BMJ Open

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053674
Article Type:	Original research
Date Submitted by the Author:	20-May-2021
Complete List of Authors:	Glaab, Enrico; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Rauschenberger, Armin; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Banzi, Rita; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Gerardi, Chiara; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Garcia, Paula; ECRIN, European Clinical Research Infrastructure Network Demotes, Jacques; ECRIN, European Clinical Research Infrastructure Network
Keywords:	BIOTECHNOLOGY & BIOINFORMATICS, NATURAL SCIENCE DISCIPLINES, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Authors

Enrico Glaab[1,*], Armin Rauschenberger[1], Rita Banzi[2], Chiara Gerardi[2], Paula Garcia[3], Jacques Demotes-Mainard[3], and the PERMIT Group

Affiliations

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg.

[2]Center for Health Regulatory Policies, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

[3]European Clinical Research Infrastructure Network (ECRIN), Paris, France

*Correspondence: enrico.glaab@uni.lu; 14, avenue du Rock'n'Roll, L-4361 Esch-sur-Alzette, Luxembourg; Phone: +352-466644 6186

Word count

Abstract: 291

Main text: 5320

Keywords

Biomarkers, Machine Learning, Review, Omics, Stratification

Abstract

Objective: To review biomarker discovery studies using omics data for patient stratification which led to clinically validated FDA-cleared tests or laboratory developed tests, in order to identify common characteristics and derive recommendations for future biomarker projects.

Design: Scoping review.

Methods: We searched PubMed, EMBASE and Web of Science to obtain a comprehensive list of biomedical literature articles describing clinically validated biomarker signatures for patient stratification, derived using statistical learning approaches. All documents were screened to retain only peer-reviewed research articles, review articles, or opinion articles, covering supervised and unsupervised machine learning applications for omics-based patient stratification. Two reviewers independently confirmed the eligibility. Disagreements were solved by consensus. We focused the final analysis on omics-based biomarkers which achieved the highest level of validation, i.e., clinical approval of the developed molecular signature as a laboratory developed test or FDA approved tests.

Results: Overall, 234 articles fulfilled the eligibility criteria. The analysis of validated biomarker signatures identified multiple common methodological and practical features that may explain the successful test development and provide guidance for future biomarker projects. These include study design choices to ensure sufficient statistical power for model building and external testing, suitable combinations of non-targeted and targeted measurement technologies, the integration of prior biological knowledge, and the adequacy of statistical and machine learning methods for discovery and validation.

Conclusions: While most clinically validated biomarker models derived from omics data have been developed for biomedical decision making in oncology, first applications for non-cancer diseases highlight the potential of multivariate omics biomarker design for other complex disorders. Distinctive characteristics of prior success stories, such as early filtering and robust discovery approaches, continuous improvements in assay design and experimental measurement technology, and rigorous multi-cohort validation approaches, enable the derivation of specific recommendations for future studies.

Article Summary

- This scoping review provides a first integrative overview of biomarker discovery studies using omics data which led to clinically validated diagnostic and prognostic tools.
- It identified shared characteristics of successful omics-based biomarker studies, which may help to guide study design, discovery and validation methods for future projects.
- Recommendations derived from the review mainly provide guidance on optimizing the design of prospective studies, but also include suggestions for retrospective studies.
- The integration of diverse, multi-omics data sources for biomarker modeling is still an exception, but has the potential to provide more robust and reliable biomedical predictions.
- Further knowledge exchange among computational, experimental and clinical experts in the field is still needed to derive comprehensive guidelines for omics-based biomarker studies.

Introduction

Personalised medicine is a rapidly developing area in health care research and practice, which aims at providing more effective and safer therapies tailored to the individual patient, by exploiting subject-specific molecular, clinical and environmental data sources (Box 1).

One of the main tools used in personalised medicine and the focus of this survey is the machine learning (ML) analysis of omics profiling data to derive molecular biomarker signatures for disease- or drug-based patient stratification. The major goals behind ML-based omics biomarker development in this domain are to develop more reliable and robust tests for drug response prediction, early diagnosis, differential diagnosis or prognosis of the future clinical disease course. Omics-derived biomarker signatures may help to guide treatment decisions, and to focus therapies on the right populations in order to prevent overtreatment, increase success rates, and reduce costs. As a research and information tool, they may also enable a better monitoring of disease progression and treatment success, and guide new drug development and discovery. In contrast to classical single-molecule biomarker approaches, omics signatures have the potential to provide more sensitive, specific and robust predictions of disease-associated outcomes.

However, while biomarker discovery projects using omics data have already led to the successful development of clinically validated diagnostic and prognostic tests (1–10), many biomarker studies are not further pursued after early development stages or fail the translation in later clinical validation stages. Dedicated statistical and ML methodologies for omics biomarker discovery and validation have been published, as well as recommendations for the study design, implementation and reporting (11,12), but it is not clear which distinctive features and approaches characterize the studies that succeed in translating omics research findings into clinically validated tests.

As part of an ongoing EU project on “Personalised Medicine Trials” (PERMIT, https://permit-eu.org), funded within the H2020 framework, we have therefore investigated the current methodological practices for personalised medicine, covering ML approaches for omics-based patient stratification as one of the major focus areas. While a broader series of questions was established for the overall scoping review (16), for this manuscript, we focused our analysis on biomarker discovery studies that have led to successful, clinically validated FDA-cleared tests or laboratory developed tests (LDTs), to determine their shared and distinctive characteristics as compared to previous biomarker studies without clinical translation. In particular, we aimed to address the following more specific research questions:

- Which omics-derived biomarker discovery studies have previously led to clinically validated tests for patient stratification (LDTs or FDA-cleared tests)?
- What are the key characteristics that are shared by successful omics biomarker studies and that distinguish them from previously published biomarker studies which have not yet led to clinically validated tests?
- Which types of model building and validation methods have been used to develop clinically validated biomarker signatures, and what are the lessons learned and recommended workflows?
- Which recommendations and guidelines have previously been proposed to address common challenges in biomarker development using omics data?

These questions lend themselves for a scoping review, because omics-derived biomarker development is still an evolving field, and a preliminary assessment of the potential scope and size of the available biomedical literature on these topics is required as a first step for further follow-up research. Therefore, the objective of this survey was to address the above questions by retrieving and examining the current literature that describe biomarker discovery and validation studies using omics data and ML approaches.

Methods

We conducted a scoping review following the methodological framework suggested by the Joanna Briggs Institute (13). This framework consists of six stages: 1) identifying the research questions, 2) identifying relevant studies, 3) study selection, 4) charting the data, 5) collating, summarising and reporting results, and 6) consultation.

The scoping review approach was considered to be the most suitable to respond to the broad scope and the evolving nature of the field. Compared to systematic reviews that aim to answer specific questions, scoping reviews are used to present a general overview of the evidence pertaining to a topic and they are useful to examine areas that are emerging, to clarify key concepts and identify gaps (14,15). Before conducting the review, a study protocol was published on the online platform Zenodo (16). Due to the iterative nature of scoping reviews, deviations from the protocol are expected and duly reported when occurred. We used the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist to report our results (17) (Online supplementary file 1).

Study identification

Relevant studies and documents were identified, balancing feasibility with breadth and comprehensiveness of searches. We searched PubMed, EMBASE and Web of Science (search date: March 13, 2020) for articles describing supervised or unsupervised ML analyses for biomarker discovery or personalised medicine, including both discovery and validation methods (see Fig. 1, illustrating the keyword-based search strategy). We included journal publications and meeting abstracts from international conferences and workshops. No other grey literature was included. Online supplementary file 2 reports the detailed search strategies applied. We restricted inclusion to reports published from January 2000 to April 2020 (covering also “online first” articles with official publication date in the near future) in English, French, Spanish, Italian and German language.

Eligibility criteria

We included peer-reviewed methodology articles, review articles, opinion articles on supervised and unsupervised ML methods for omics stratification and associated validation methods (addressing accuracy, robustness, and clinical relevance). Only approaches tested on real-world biomedical data were reviewed, while studies relying purely on simulated data were excluded. We also excluded papers on biomarker methodologies without a demonstrated biomedical application, and those with insufficient sample size (i.e., less than 50 samples per group used for the main conditions studied, unless a dedicated power calculation was presented) or statistical validation (i.e., lack of clear descriptions of cross-validation or external testing methodology, performance metrics and test statistics). These exclusion criteria were not specified in the generic review protocol, but they were agreed among the authors before starting the screening process.

To cover both data from original research papers and prior systematic reviews, we extracted information from three main article types: (1) applied research papers, (2) methodology articles with demonstrated applications, and (3) review articles on methods, applications and validation approaches.

Apart from these inclusion and exclusion criteria, for the final presentation of results, the statistical investigations covered all selected articles, whereas the detailed discussion of study characteristics in this paper focused mainly on the studies that led to clinically validated biomarker signatures tested on multiple cohorts with large sample sizes (i.e., studies using a power calculation to demonstrate the adequacy of the chosen sample sizes, or covering hundreds or thousands of samples per studied subject group).

Study selection

We exported the references retrieved from the searches into the online tool Rayyan (18). Duplicates were removed automatically using the reference manager Endnote X9 (Clarivate Analytics, Philadelphia, United States) and manually by the reviewers. One reviewer screened the titles and abstracts and retrieved full-text copies of potentially eligible reports for further assessment. Two reviewers independently confirmed the eligibility. Disagreements were solved by consensus.

Charting the data

We designed a data extraction form using Excel (Online supplementary file 3). General study characteristics extracted were for instance: authors, title, citation, type of publication (e.g., journal article, meeting abstract), study population and sample size (if applicable), methodology/study design, and outcome measures (if applicable). Specific items included key findings that relate to the review question; type of article/study (e.g., methodology, applied research, review– methods, review– applications, review–validation); generic ML domain (e.g., supervised/unsupervised); name of specific ML approach used.

To capture key findings that relate to the review question, relevant sentences were extracted from each reviewed article, and if needed, complemented by a brief explanatory remark by the reviewer.

The reviewers piloted the data extraction form using five records from the retrieved article collection. Two reviewers (EG, AR) working independently extracted the data from the included articles. In the case of disagreements, consensus was obtained by discussion.

In the final full-text review stage, the pre-selected articles were grouped by topic, categorizing articles into applied vs. methodological studies, supervised vs. unsupervised analyses, and assigning algorithm type identifiers to each article (review articles and papers on validation methodologies were considered as separate categories without a specific algorithm type assignment).

It was not within the remit of this scoping review to assess the methodological quality of individual studies included in the analysis.

Consultation exercise

The members of the PERMIT consortium, associated partners, and the PERMIT project Scientific Advisory Board discussed the preliminary findings of the scoping review in a 2-hour online workshop.

Patient and public involvement

The European Patients' Forum is a member of PERMIT project. Although not directly involved in the conduction of the scoping review, they received the draft review protocol for collecting comments and feedback.

Results

Study selection and general characteristics of reports

We retrieved 1164 abstracts from the literature search. After the removal of duplicates, we screened the remaining 1079 abstracts for eligibility. 502 records were excluded, while 577 abstracts were retained for the full-text assessment. We finally included 234 articles that passed all filtering criteria in the data extraction and analysis step (see flow chart in Fig. 2 and online supplementary file 2).

The full-text article review revealed that many studies did not meet the pre-defined inclusion criteria: 251 articles (44%) were removed because of an insufficient sample size, and 67 further articles (12%) were removed because they provided insufficient details on the validation results or methodology (see Fig. 2). This shows that the challenges of recruiting an adequate number of participants per study group or conducting sufficient omics profiling experiments for robust model building and validation are not met in a large proportion of omics biomarker studies, and that many of these studies lack adequate documentation for the study design and validation.

For the 234 selected articles, the majority (81%) rely entirely on an internal validation involving data from only a single cohort, whereas studies that use an external validation on an independent cohort are still underrepresented (only 12% of articles describe both an internal cross-validation and an external cohort validation, and an additional 7% include an external validation, but do not report internal cross-validation results). Moreover, when comparing the numbers of published studies involving different types of internal and external validations over different periods of time during the past 20 years, the relative proportion of studies including an external validation has only slightly increased in recent years (see Fig. 3).

Since a detailed discussion of all filtered articles is not within the scope of the present review, in the following, we focus on reviewing representative omics-based biomarkers studies which have achieved the highest level of validation, i.e., clinical approval of the developed molecular signature as an LDT or FDA approved test (see the overview of studies in Table 1). By investigating the shared and distinctive features of these successful studies, we also cover how they address common shortcomings and missing features of other reviewed studies, and summarize the lessons learned.

Success stories in omics-based biomarker signature development

Cancer approved omics-derived diagnostic tests (8 studies)

The first and most well-known omics-derived molecular test to receive FDA clearance was *MammaPrint*, a prognostic signature using the RNA expression activity of 70 genes to estimate the risk for distant tumor metastasis and recurrence in early-stage breast cancer patients (1,19–23). This test had been developed at the Netherlands Cancer Institute, using DNA microarray analysis to investigate primary breast tumors of 117 patients. Supervised ML was applied to the resulting data to identify a gene signature that was highly predictive of a short interval to distant metastases in lymph node negative patients (19).

A distinctive feature of the development approach behind this signature in comparison to other reviewed studies was the multi-stage filtering and cross-validation strategy used in the initial discovery study, which may explain the repeated confirmation of the signature in later validation studies (1,20–23). From 25k genes represented on the DNA microarrays, only those significantly regulated in more than 3 tumors out of the subset of 78 sporadic lymph-node negative patients were preselected, and further filtered by retaining only the genes with a minimum absolute correlation with the disease outcome of 0.3. The resulting list of 231 genes, rank-ordered by absolute correlation, was investigated by sequentially adding the next top 5 genes from the list to a candidate ML classifier and evaluating its performance by leave-one-out cross-validation (LOOCV). This procedure was repeated as long as the estimated accuracy of the classifier improved, providing a final candidate signature of 70 genes. The final signature was validated on multiple independent test sets (including a limited set of 19 external samples in the original study, but several additional validations on independent cohorts in subsequent studies (1,20–23).

The *MammaPrint* signature also provided the role model for the subsequent development of a similar prognostic test for colon cancer, *ColoPrint* (24–29). This test aims at detecting the approx. 20% of patients with stage II colon cancer expected to experience a relapse and develop distant metastases. It uses an 18-gene expression signature, developed by analyzing DNA microarray data in a similar manner to the *MammaPrint* approach. This diagnostic tool has been commercialized as an LDT to assist physicians in selecting treatment options for colon cancer patients. Similar to *MammaPrint*, the development of this signature was characterized by extensive discovery and validation studies, which involved multiple statistical reproducibility, stability and precision analyses across independent, large-scale patient cohorts (30).

Another widely used cancer-related LDT, which received FDA clearance in 2013, is the *Prosigna Breast Cancer Prognostic Gene Signature Assay*, previously called *PAM50* test (31–35). This assay assesses mRNA expression for a signature of 58 genes (50 target genes + 8 endogenous control genes) to predict the risk of distant recurrence for hormone-receptor-positive breast cancer from 5 to 10 years after diagnosis (prerequisites are that the patients have been treated with hormonal therapy and surgery, and are stage I or stage II lymph-node negative, or in stage II with one to three positive nodes). The development of the test started with a microarray discovery study and involved a multistage filtering approach, using consecutive applications of statistical tests and cross-validation to propose a subset of candidate gene markers (36). The authors compared the reproducibility of classification scores obtained with these markers for three centroid-based prediction methods to ensure the robustness of the methodology. By further developing the approach first into a more sensitive PCR-based test, and then into an assay using the NanoString nCounter Dx Analysis System, the predictive performance was improved in a step-wise fashion. The original discovery study was characterized by significantly larger sample sizes than the majority of reviewed biomarker studies, with a training set of 189 samples, test sets of 761 patients evaluated for prognosis, and 133 patients evaluated for prediction of pathologic complete response to treatment with taxane and anthracycline. These study design features in combination with multi-stage filtering and validation approaches, and the improvement of the measurement technology during the course of the study, may explain the successful progression of the PAM50 test to FDA clearance.

Among the LDTs for breast cancer prognosis, *Oncotype DX*® is a further test which is already commonly used in clinical practice (3,37–40). The underlying gene signature consists of 16 cancer-associated genes and 5 reference genes, and is therefore often also referred to as '21-gene assay'. Its main application is to predict risk of recurrence in estrogen-receptor positive tumors. The relevance of this prognostic tool for treatment selection is explained by the strong association of the provided recurrence score with the probability of positive treatment response to chemotherapy. *Oncotype DX* was developed using a consecutive refinement procedure, starting with the RT-PCR assessment of 250 candidate genes across 447 patients from three distinct studies to identify the 21-gene signature after multiple filtering steps. A recurrence score algorithm built using the signature

as input was clinically validated on 668 independent patients (41). The selection of the 16 cancer-related genes included in the assay was mainly done by scoring the performance of the candidate features in all three studies and the consistency of the primer/probe performance in the assay (42).Thus, particular strengths of the development process for this LDT include the consideration of both technical robustness and statistical robustness of the assay across distinct cohorts. However, an independent comparative clinical validation of Oncotype DX and the PAM50 signature for estimating the likelihood of distant recurrence in ER-positive, node-negative, post-menopausal breast cancer patients treated with endocrine therapy suggested that the PAM50 signature provided more predictive information than Oncotype DX (43).

While the first validated omics biomarker signatures were developed for breast cancer, similar diagnostic and prognostic tools have followed for other cancer types. One of these is the Decipher Prostate Cancer Test (4,44–48), which stands out from other omics-derived diagnostic tools in that it is provided together with an additional software platform and database, the Decipher Genomic Resource Information Database (GRID), that captures 1.4 million expression markers per patient to facilitate personalised care. The test itself uses 22 preselected RNAs to predict clinical metastasis and cancer-specific mortality for patients who have undergone radical prostatectomy. An initial discovery study by the Mayo Clinic (Rochester, MN, USA) investigated a cohort of 545 such patients, split into a training (n = 359) and a validation cohort (n = 186). Similar to other LDTs, the discovery started with a genome-wide profiling and used both statistical and ML analyses for filtering. First, *t*-tests were applied (reduction from 1.4 mil. to 18,902 differentially expressed RNAs), then regularized logistic regression (reduction to 43 candidate markers), and finally a random forest-based feature selection (reduction to final set of 22 RNAs). Apart from testing the signature in the validation cohort, further external validations were performed in subsequent studies (4,44–48). Overall, distinctive strengths of the used approach include the improved interpretability of the test results through supporting analyses on the GRID platform, and the robustness of the discovery and validation approach, involving large sample sizes and several complementary statistical and ML assessments.

While most diagnostic tests in oncology have been designed for specific cancer types, a dedicated LDT has also been developed for cancers of unknown or uncertain diagnosis. The Cancer Type ID test by bioTheranostics distinguishes between 50 different tumor types using a 92-gene RT-PCR expression measurement signature (10,49–51). This signature has been derived from analyses of a microarray data collection covering 446 frozen tumor samples and 112 formalin-fixed, paraffin-embedded (FFPE) samples of both primary and metastatic tumors. The modeling steps involved k-nearest neighbor clustering and classification, and a genetic algorithm to explore the search space of possible feature subset selections. After successful cross-validation (84% accuracy) and external validation (82% accuracy on 112 independent FFPE samples) of the microarray-based signature, it was further developed to use more sensitive RT-PCR measurements. Testing the new approach on an independent validation set provided an increased accuracy (87%). Distinctive characteristics of the overall development process that may have contributed to the positive validation include the efficient and extensive exploration of the search space of possible gene subset selections via a genetic algorithm, the large sample sizes used for discovery and validation, and the transfer of the assay from microarrays to the more sensitive RT-PCR platform.

Apart from the omics-derived biomarker signatures that address the most frequent cancer types, more recent applications in oncology focus on the diagnosis of less common malignancies, such as thyroid cancer. Typically, deciding whether a thyroid nodule is benign or cancerous is directly possible via a fine needle aspiration (FNA) biopsy, without requiring more complex measurements or analyses. However, while direct FNA-based diagnosis is feasible in most cases, indeterminate results can occur (52). To help prevent unnecessary surgeries for the corresponding patients, a molecular signature and LDT known as the Afirma™ Gene Expression Classifier (GEC) has been developed to discriminate between benign and cancerous thyroid nodules (52–57). The original discovery study behind the GEC signature used mRNA expression analysis in 315 thyroid nodules,

covering 178 retrospective surgical tissues and 137 prospectively collected FNA samples. The authors trained two ML classifiers separately on surgical tissues and FNAs, assessing the test set performance on 48 independent, prospective FNA samples (50% of which had indeterminate cytopathology). Discriminative features were selected using a linear modeling approach implemented in the software Limma, and a linear support vector machine was applied for model building and performance estimation via 30-fold cross-validation (CV). The successful cross-validation results were later confirmed on multiple distinct cohorts. While the internal validation used in the initial study cannot address cohort-specific biases, the combined use of established feature selection and modeling approaches, and the subsequent external validation across multiple cohorts with large sample sizes may account for the successful translation of this signature.

Most omics-based diagnostic tests identified in the survey rely purely on gene expression profiling data. However, more recently, first multi-omics signatures for diagnostic purposes have been developed. One of the first LDTs that integrated information from both RNA and DNA sequencing was the FoundationOne Heme assay (9,58–60). This assay aims to detect hematologic malignancies, sarcomas, pediatric malignancies, or solid tumors (including among others leukemias, myelodysplastic syndromes, myeloproliferative neoplasms, lymphomas, multiple myeloma, Ewing sarcoma, Leiomyosarcoma, and pediatric tumors). The test identifies four types of genomic alterations (base substitutions, insertions and deletions, copy number alterations, rearrangements) and reports microsatellite instability and tumor mutational burden to facilitate clinical decision making. The approach was originally developed and evaluated using reference samples of pooled cell lines in order to model the main characteristics that determine the test accuracy, including mutant allele frequency, indel length and amplitude of copy change (58). A first validation using 249 independent FFPE cancer samples, which had already been characterized by established assays, confirmed the accuracy of the test. Later external validation studies on independent cohorts also corroborated the utility of the test for further diagnostic applications (9,61). The study results highlight the potential of integrating diverse biological data sources in order to obtain more robust and reliable predictions, a strategy that may be promising in particular for complex disorders that involve very heterogeneous phenotypes.

Non cancer approved omics-derived diagnostic tests (2 Studies)

While most clinically approved omics-derived diagnostic tests have been developed in the field of oncology, one of the first LDTs that received FDA clearance for a non-cancer disease was the AlloMap Heart test (8,62–64). It uses a gene expression signature of 11 target genes and 9 control genes in peripheral blood from heart transplant recipients to estimate the risk for acute cellular cardiac allograft rejection. The development process involved statistical analyses of leukocyte microarray profiling data from 285 samples, and subsequent RT-PCR validation and bioinformatics post-processing (8). Prior knowledge from database and literature mining was integrated into the analysis by mapping the data to known alloimmune pathways. This allowed the researchers to narrow down 252 candidate marker genes. An RT-PCR validation on 145 samples confirmed 68 of these candidate genes, which distinguished rejection samples from quiescent samples according to a T-test ($p < 0.01$). Six genes were eliminated due to significant variation in gene expression with sample processing time. Next, correlated gene expression levels were averaged to create robust meta-level features, called ‘metagenes’, and 20 of these features were added as new variables. Finally, a linear discriminant analysis was applied, providing a prediction model using four individual genes and three metagenes, which aggregate information from 11 original genes. Bootstrap validation procedures and external test set validations were performed to confirm the accuracy of this signature. Overall, distinctive aspects of the development approach for the AlloMap signature include the knowledge-based gene discovery, a comprehensive RT-PCR validation of candidate genes, and the robust bootstrap and external validation analyses.

The first clinically validated LDT for a cardiovascular indication derived from omics data was the Corus CAD test, developed to identify coronary artery disease (CAD) in stable non-diabetic patients (6,65–68). In contrast to most other omics-based tests, Corus CAD is not a pure molecular signature test, but also takes the clinical covariates gender and age into account. The initial discovery study used a retrospective microarray analysis of blood samples from 195 diabetic and non-diabetic patients from the Duke University CATHGEN registry. After ranking the studied genes in terms of the statistical significance of group differences and prior biological knowledge about their disease relevance, 88 genes were selected for RT-PCR validation. Because diabetes status as a clinical covariate was significantly associated with the observed gene expression alterations, and the identified CAD-associated genes did not overlap between diabetic and non-diabetic patients, the authors decided to limit follow-up work to the non-diabetic patients. In a prospective clinical trial, the PREDICT study, microarray profiling was conducted on blood samples from 198 patients, and top-ranked genes were further validated using RT-PCR for 640 PREDICT blood samples. After multiple filtering steps, taking into account statistical significance in T-tests, biological relevance, gene correlation clustering and cell-type analyses, a final signature of 23 genes was derived, composed of 20 CAD-associated genes and 3 reference genes (69). To maximize the predictive performance, the final prediction algorithm was optimized to adjust for differences associated with age and gender. Overall, compared to other reviewed studies, the Corus CAD approach stands out by taking clinical covariates into account in the final prediction model, including a critical review and adjustment of the inclusion criteria (limiting the focus to nondiabetic patients), and integrating complementary filtering and validation analyses on large sample sizes.

Discussion

Statement of principal findings

The scoping review of articles on patient stratification using omics data revealed common limitations in the study design for many published biomarker development projects, such as insufficient and imbalanced sample sizes per study group and inadequate validation methods, but also identified multiple studies that have led to validated diagnostic and prognostic tests. These success stories were investigated in more detail to identify common characteristics in the study design, discovery and validation methods, which may have supported the clinical translation of the initial findings. Key shared aspects that are possible determinants of the study success and could help to guide future biomarker investigations are outlined in Fig. 4. These characteristics, which may serve as a guideline for future studies, cover in particular the following main features:

(1) A sample size selection, study group and replicate design that provides adequate statistical power for the ML analyses;

2) The application of robust statistical filtering and evaluation schemes (including multiple layers of statistical and ML-based feature selection, combined statistical and biological filters, robust validation schemes that involve multiple cross-validation, bootstrapping and external validation analyses, using multiple suitable and complementary performance metrics, and providing information on the statistical variation and confidence intervals for the performance estimates);

3) Clarity of the study scope and goals (involving clear inclusion and exclusion criteria, and precisely defined primary and secondary outcomes; new knowledge gained during the project may require a re-definition and adjustment of the inclusion criteria, as in the case of the Corus CAD study, or adjustments in the methodology to reach the goals, such as the progression from non-targeted microarray technology to higher-sensitivity RT-PCR in the case of the Prosigna test and the Cancer Type ID test);

4) Completeness and reproducibility of the study documentation (covering details on used instruments, parameters and settings, reproducible methods descriptions, and information on data provenance);

5) Interpretability and biological plausibility of the created predictive models (including explainable and justifiable predictions, human-interpretable model descriptions, and biologically plausible models that agree with the current mechanistic understanding of the studied disorder);

6) Integration of prior biological knowledge into the predictive feature selection, model building and validation procedures (e.g., using public data on disease-associated molecular pathways and networks; complementary clinical and real-world data, and relevant multi-omics data).

Strengths and Limitations

The majority of methodological recommendations derived from the literature survey relate to the early planning and study design for biomarker discovery projects, involving considerations associated with the choice of the study group, sampling and blocking design, the measurement technology, and the input and output variables (11,12). These recommendations are therefore mainly applicable to prospective studies. For retrospective biomarker investigations of already collected data, the suggestions derived from the review are limited to guidance on filtering and evaluation analyses, the integration of prior knowledge from multi-omics data and public annotation databases; and the choice of robust and interpretable modelling approaches for the generation of biologically plausible and reproducible prediction models. While the main focus of the review on studies that have already led to validated biomarker models helps to ensure the quality of the surveyed articles, more recent methodological developments in the ML and cross-validation analysis of omics data, such as meta-learning (70) and bolstered cross-validation (71), will require further dedicated surveys in the future.

Discussing important differences in results

Previous reviews of ML approaches using omics data for patient stratification have mostly focused on domain-specific analyses for specific types of diseases, or specific types of ML methodologies (72–80). By contrast, this scoping review focused on disease-agnostic workflows with generic applicability across human disorders that involve multifactorial molecular alterations in the affected tissues and body fluids. The coverage of statistical and ML approaches for stratification did not focus on the detailed discussion of specific algorithms or statistics, but rather aimed at identifying key determinants of success for generic analysis workflows that have been applied successfully in practice for biomedical stratification studies.

Meaning of the study: implications for clinicians and policymakers

The previous clinical translation successes in omics-based biomarker development reviewed in this survey, which have mostly been achieved for oncology applications, highlight the potential for similar biomarker discovery and validation projects in other fields of biomedicine. In contrast to conventional statistical biomarker discovery approaches, which focus on identifying single-molecule markers, systems-level analysis of omics data using multivariate ML approaches can identify biomarker signatures which are not only more sensitive, specific and robust, but also more biologically insightful in terms of reflecting disease-associated cellular process alterations in a more detailed and comprehensive fashion.

This scoping review has identified common characteristics of omics studies which have led to clinically validated diagnostic and prognostic tests. Thus, it may help to guide clinical researchers on study design choices, the selection of methodologies for statistical and biological data filtering and ML, and the implementation of adequate validation schemes. By creating awareness of potential pitfalls, such as issues associated with batch effects, biases, confounding factors, lack of statistical power, and multiple hypothesis testing, common reasons for failures in biomarker development can be avoided. This information may also be relevant for policymakers and funding bodies, by facilitating the design of public and private funding schemes for biomedical research in a manner that addresses risks in the funded projects upfront through appropriate guidelines and regulations. Finally, it may help clinicians involved in biomarker discovery to make better use of already available public knowledge and data sources, e.g. cellular pathway and molecular interaction databases, that may allow them to exploit prior knowledge more effectively in biomarker modelling, and create more robust and interpretable prediction models.

Unanswered questions & future research

Since the recommendations and guidelines identified from the surveyed articles are mostly derived from already established biomarker discovery and validation approaches, new upcoming methodologies could only be covered to a limited extent and may lead to changed recommendations in the future. In particular, in the reviewed patient stratification studies, some of the more recently developed ML and validation workflows, e.g. involving meta-learning, bootstrapping or bolstered cross-validation, are still underrepresented among the reviewed studies, and may play a more important role in the future.

Overall, while the currently available literature on validated stratification biomarkers already provides ample information on common pitfalls and best practices, the development of widely accepted standard guidelines on methodologies for omics biomarker discovery will require further knowledge exchange and deliberation among the stakeholders in the field. In particular, integration of domain-specific expertise in discussions involving clinicians, experimental and data scientists, and regulatory and legal experts is required as a follow-up effort in order to derive comprehensive methodological guidelines for future biomarker development.

Acknowledgments

The authors thank Vanna Pistotti for her assistance with search strategy development and conduction.

List of Figures

Figure 1: Keyword based search strategy for the scoping review

Figure 2: Study selection flow diagram

Figure 3: Validation methods used in omics biomarker studies

Figure 4: Characteristics of successful omics-based studies

Definitions (In boxes)

Box 1: *What is Personalised Medicine?*

According to the European Council Conclusion on personalised medicine for patients, personalised medicine is 'a medical model using characterisation of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention (81).

In the context of the Permit project, we applied the following common operational definition of personalised medicine research: a set of comprehensive methods (methodology, statistics, validation, technology) to be applied in the different phases of the development of a personalised approach to treatment, diagnosis, prognosis, or risk prediction. Ideally, robust and reproducible methods should cover all the steps between the generation of the hypothesis (e.g., a given stratum of patients could better respond to a treatment), its validation and pre-clinical development, and up to the definition of its value in a clinical setting (16).

References

1. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717–29.
2. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;
3. Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genomics. 2013;14(1).
4. Torres A, Alshalalfa M, Tomlins SA, Erho N, Gibb EA, Chelliserry J, et al. Comprehensive Determination of Prostate Tumor ETS Gene Status in Clinical Samples Using the CLIA Decipher Assay. J Mol Diagnostics. 2017;19(3):475–84.
5. Angell TE, Babiarz J, Barth N, Blevins T, Duh Q, Ghossein RA, et al. Clinical validation of the AFIRMA genomic sequencing braf V600E classifier. Thyroid [Internet]. 2017;27:A50. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L624116485
6. Ladapo JA, Budoff MJ, Sharp D, Kuo JZ, Huang L, Maniet B, et al. Utility of a Precision Medicine Test in Elderly Adults with Symptoms Suggestive of Coronary Artery Disease. J Am Geriatr Soc. 2018;66(2):309–15.
7. Tabari E, Lovejoy AF, Lin H, Bolen CR, Saelee SL, Lefkowitz JP, et al. Molecular characteristics and disease burden metrics determined by next-generation sequencing on circulating tumor DNA correlate with progression free survival in previously untreated diffuse large B-cell lymphoma. Blood [Internet]. 2019;134. Available from: http://dx.doi.org/10.1182/blood-2019-123633
8. Deng MC. The AlloMap™ genomic biomarker story: 10 years after. Clin Transplant. 2017;31(3).
9. He J, Abdel-Wahab O, Nahas MK, Wang K, Rampal RK, Intlekofer AM, et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood. 2016;127(24):3004–14.
10. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med. 2006;130(4):465–73.
11. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013.
12. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.
13. Vinet L, Zhedanov A. A "missing" family of classical orthogonal polynomials. J Phys A Math Theor. 2011;44(8).
14. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, Mcewen SA. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Res Synth Methods. 2014;5(4):371–85.
15. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping

review approach. BMC Med Res Methodol. 2018;18(1):143.

16. Banzi R, Gerardi C, Fratelli M, Garcia P, Torres T, Abad JMH, et al. Methodological approaches for personalised medicine: protocol for a series of scoping reviews [Internet]. 10.5281/zenodo.3770937. Available from: https://zenodo.org/record/3770937

17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018 Oct;169(7):467–73.

18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.

19. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.

20. Winner BS, Sgroi DC, Ryan PD, Bruinsma TJ, Glas AM, Male A, et al. Analysis of the mamma print breast cancer assay in a predominantly postmenopausal cohort. Clin Cancer Res. 2008;14(10):2988–93.

21. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: Another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn. 2009;9(5):417–22.

22. Sapino A, Roepman P, Linn SC, Snel MHJ, Delahaye LJMJ, Van Den Akker J, et al. MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue. J Mol Diagnostics. 2014;16(2):190–7.

23. Mook S, Knauer M, Bueno-De-Mesquita JM, Retel VP, Wesseling J, Linn SC, et al. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. Ann Surg Oncol. 2010;17(5):1406–13.

24. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, et al. Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. Ann Surg. 2013;257(6):1053–8.

25. Kopetz S, Tabernero J, Rosenberg R, Jiang Z, Moreno V, Bachleitner-Hofmann T, et al. Genomic Classifier ColoPrint Predicts Recurrence in Stage II Colorectal Cancer Patients More Accurately Than Clinical Factors. Oncologist. 2015;20(2):127–33.

26. Tan IB, Tan P. Genetics: An 18-gene signature (ColoPrint®) for colon cancer prognosis. Nat Rev Clin Oncol. 2011;8(3):131–3.

27. Rosenberg R, Maak M, Simon I, Nitsche U, Schuster T, Kuenzli B, et al. Independent validation of a prognostic genomic profile (ColoPrint) for stage II colon cancer (CC) patients. J Clin Oncol. 2011;29(4_suppl):358–358.

28. Salazar R, de Waard JW, Glimelius B, Marshall J, Klaase J, Van Der Hoeven J, et al. The PARSC trial, a prospective study for the assessment of recurrence risk in stage II colon cancer (CC) patients using ColoPrint. J Clin Oncol. 2012;30(4_suppl):678–678.

29. Tabernero J, Moreno V, Rosenberg R, Nitsche U, Bachleitner-Hofmann T, Lanza G, et al. Clinical and technical validation of a genomic classifier (ColoPrint) for predicting outcome of patients with stage II colon cancer. J Clin Oncol. 2012;30(4_suppl):384–384.

30. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;23:ix179.

31. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC

Cancer. 2014;14(1).

32. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics. 2015;8(1).

33. Alvarado MD, Prasad C, Rothney M, Cherbavaz DB, Sing AP, Baehner FL, et al. A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer. Adv Ther. 2015;32(12):1237–47.

34. Jensen MB, Lænkholm AV, Nielsen TO, Eriksen JO, Wehn P, Hood T, et al. The Prosigna gene expression assay and responsiveness to adjuvant cyclophosphamide-based chemotherapy in premenopausal high-risk patients with breast cancer. Breast Cancer Res. 2018;20(1).

35. Hequet D, Callens C, Gentien D, Albaud B, Mouret-Reynier MA, Dubot C, et al. Prospective, multicenter French study evaluating the clinical impact of the Breast Cancer Intrinsic Subtype-Prosigna® Test in the management of early-stage breast cancers. PLoS One. 2017;12(10).

36. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.

37. Kelly CM, Krishnamurthy S, Bianchini G, Litton JK, Gonzalez-Angulo AM, Hortobagyi GN, et al. Utility of oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. Cancer. 2010;116(22):5161–7.

38. Lo SS, Mumby PB, Norton J, Rychlik K, Smerage J, Kash J, et al. Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection. J Clin Oncol. 2010;28(10):1671–6.

39. Carlson JJ, Roth JA. The impact of the Oncotype Dx breast cancer assay in clinical practice: A systematic review and meta-analysis. Vol. 141, Breast Cancer Research and Treatment. 2013. p. 13–22.

40. Thakur SS, Li H, Chan AMY, Tudor R, Bigras G, Morris D, et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One. 2018/01/06. 2018;13(1):e0188983.

41. Gianni L, Zambetti M, Clark K, Baker J, Cronin M, Wu J, et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J Clin Oncol. 2005;23(29):7265–77.

42. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817–26.

43. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol. 2013;31(22):2783–90.

44. Marrone M, Potosky AL, Penson D, Freedman AN. A 22 gene-expression assay, decipher® (GenomeDx biosciences) to predict five-year risk of metastatic prostate cancer in men treated with radical prostatectomy. PLoS Curr. 2015;7(EVIDENCEONGENOMICTESTS).

45. Nguyen PL, Haddad Z, Lam LLC, Ong K, Buerki C, Deheshi S, et al. Evaluation of the Decipher prostate cancer classifier to predict metastasis and disease-specific mortality from genomic analysis of diagnostic prostate needle biopsy specimens. J Clin Oncol. 2017;35(6_suppl):4–4.

46. Magi-Galluzzi C, Yousefi K, Haddad Z, Palmer-Aronsten B, Lam LLC, Buerki C, et al. Validation of the Decipher prostate cancer classifier for predicting 10-year postoperative metastasis from analysis of diagnostic needle biopsy specimens. J Clin Oncol. 2016;34(2_suppl):59–59.

47. Dalela D, Löppenberg B, Sood A, Sammon J, Abdollah F. Contemporary Role of the Decipher® Test in Prostate Cancer Management: Current Practice and Future Perspectives. Rev Urol. 2016;18(1):1–9.

48. Klein EA, Haddad Z, Yousefi K, Lam LLC, Wang Q, Choeurng V, et al. Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology. 2016;90:148–52.

49. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, et al. Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagnostics. 2013;15(2):263–9.

50. Greco FA, Spigel DR, Yardley DA, Erlander MG, Ma X, Hainsworth JD. Molecular Profiling in Unknown Primary Cancer: Accuracy of Tissue of Origin Prediction. Oncologist. 2010;15(5):500–6.

51. Hainsworth JD, Rubin MS, Spigel DR, Boccia R V., Raby S, Quinn R, et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: A prospective trial of the Sarah cannon research institute. J Clin Oncol. 2013;31(2):217–23.

52. Harrison G, Sosa JA, Jiang X. Evaluation of the afirma gene expression classifier in repeat indeterminate thyroid nodules. Arch Pathol Lab Med. 2017;141(7):985–9.

53. Chudova D, Wilde JI, Wang ET, Wang H, Rabbee N, Egidio CM, et al. Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol Metab. 2010;95(12):5296–304.

54. Kim MI, Alexander EK. Diagnostic use of molecular markers in the evaluation of thyroid nodules. Vol. 18, Endocrine Practice. 2012. p. 796–802.

55. Ali SZ, Fish SA, Lanman R, Randolph GW, Sosa JA. Use of the Afirma® gene expression classifier for preoperative identification of benign thyroid nodules with indeterminate fine needle aspiration cytopathology. PLoS Currents. 2013. p. 1–7.

56. McIver B, Castro MR, Morris JC, Bernet V, Smallridge R, Henry M, et al. An independent study of a gene expression classifier (Afirma) in the evaluation of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab. 2014;99(11):4069–77.

57. Lastra RR, Pramick MR, Crammer CJ, LiVolsi VA, Baloch ZW. Implications of a suspicious afirma test result in thyroid fine-needle aspiration cytology: An institutional experience. Cancer Cytopathol. 2014;122(10):737–44.

58. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31(11):1023–31.

59. Wang K, Sanchez-Martin M, Wang X, Knapp KM, Koche R, Vu L, et al. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. Leukemia. 2017;31(1):151–8.

60. Tarlock K, He J, Zhong S, Ries RE, Bailey M, Morley S, et al. Distinct age-associated genomic profiles in acute myeloid leukemia (AML) using FoundationOne heme. J Clin Oncol. 2016;34(15_suppl):7041–7041.

61. Lieber DS, Kennedy MR, Johnson DB, Greenbowe JR, Frampton GM, Schrock AB, et al. Abstract B16: Validation and clinical feasibility of a Foundation Medicine assay to identify

immunotherapy response potential through tumor mutational burden (TMB). In 2017.

62. Yamani MH, Taylor DO, Rodriguez ER, Cook DJ, Zhou L, Smedira N, et al. Transplant Vasculopathy Is Associated With Increased AlloMap Gene Expression Score. J Hear Lung Transplant. 2007;26(4):403–6.

63. Yamani MH, Taylor DO, Haire C, Smedira N, Starling RC. Post-transplant ischemic injury is associated with up-regulated AlloMap gene expression. Clin Transplant. 2007;21(4):523–5.

64. Kobashigawa J, Patel J, Azarbal B, Kittleson M, Chang D, Czer L, et al. Randomized Pilot Trial of Gene Expression Profiling Versus Heart Biopsy in the First Year after Heart Transplant: Early Invasive Monitoring Attenuation Through Gene Expression Trial. Circ Hear Fail. 2015;8(3):557–64.

65. Wingrove JA, Daniels SE, Sehnert AJ, Tingley W, Elashoff MR, Rosenberg S, et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. Circ Cardiovasc Genet. 2008;1(1):31–8.

66. Rosenberg S, Dehais C, Ducray F, Alentron A, Tanguy M, De Reyneis A, et al. OS11.3 Machine learning for better prognostic stratification and driver genes identification in 1p/19q-codeleted grade III gliomas. Neuro Oncol [Internet]. 2017;19(suppl_3):iii22–iii22. Available from: http://dx.doi.org/10.1093/neuonc/nox036

67. Vargas J, Lima JAC, Kraus WE, Douglas PS, Rosenberg S. Use of the Corus® CAD Gene Expression Test for Assessment of Obstructive Coronary Artery Disease Likelihood in Symptomatic Non-Diabetic Patients. PLoS Currents. 2013.

68. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC Med Genomics. 2011;4.

69. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010;153(7):425–34.

70. Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies. Artif Intell Rev. 2015;

71. Sima C, Braga-Neto UM, Dougherty ER. High-dimensional bolstered error estimation. Bioinformatics. 2011;

72. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews. 2019.

73. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Br Bioinform. 2008/03/04. 2008;9(2):119–28.

74. Grollemund V, Pradat PF, Querin G, Delbot F, Le Chat G, Pradat-Peyre JF, et al. Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions. Front Neurosci [Internet]. 2019;13. Available from: http://dx.doi.org/10.3389/fnins.2019.00135

75. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet [Internet]. 2019;10(MAR). Available from: http://dx.doi.org/10.3389/fgene.2019.00267

76. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. Curr Hypertens Rep. 2018/07/08. 2018;20(9):75.

77. Long NP, Yoon SJ, Anh NH, Nghi TD, Lim DK, Hong YJ, et al. A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer.

Metabolomics. 2019/03/05. 2018;14(8):109.

78. Martinez BI, Stabenfeldt SE. Current trends in biomarker discovery and analysis tools for traumatic brain injury. J Biol Eng [Internet]. 2019;13(1). Available from: http://dx.doi.org/10.1186/s13036-019-0145-8

79. Patil S, Awan KH, Arakeri G, Seneviratne CJ, Muddur N, Malik S, et al. Machine learning and its potential applications to the genomic study of head and neck cancer-A systematic review. J Oral Pathol Med. 2019;48(9):773–9.

80. Saini G, Mittal K, Rida P, Janssen EAM, Gogineni K, Aneja R. Panoptic view of prognostic models for personalized breast cancer management. Cancers (Basel) [Internet]. 2019;11(9). Available from: http://dx.doi.org/10.3390/cancers11091325

81. Europen Council. Council conclusions on personalised medicine for patients. Off J Eur Union [Internet]. 2015;431(2):1–4. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015XG1217(01)&from=EN

Author Contributions

Study conception and design: EG, AR.

Methodology: CG, RB.

Data collection and analysis: EG, AR

Original draft preparation: EG

Review and editing: AR, PG, CG, JDM, RB.

Project supervision: PG

Funding acquisition: JDM.

All authors have read and revised the manuscript and approved the final version.

The members of the PERMIT group were involved in the preparation or revision of the joint protocol of the four scoping reviews of the PERMIT series, attended the joint workshop (consultation exercise) and are co-authors of the other scoping reviews of the PERMIT series.

Collaborators

PERMIT group:

1. Antonio L. Andreu
2. Florence Bietrix,
3. Florie Brion Bouvier
4. Montserrat Carmona Rodriguez
5. Maria del Mar Polo-de Santos,
6. Maddalena Fratelli,
7. Rainer Girgenrath,
8. Alexander Grundmann,
9. Josep Maria Haro,
10. Frank Hulstaert,
11. Iñaki Imaz-Iglesia,
12. Setefilla Luengo Matos
13. Emmet McCormack,
14. Albert Sanchez Niubo,
15. Emanuela Oldoni,
16. Raphael Porcher,
17. Vibeke Fosse,
18. Luis M. Sánchez-Gómez,
19. Lorena San Miguel,
20. Cecilia Superchi,
21. Teresa Torres,
22. Anna Monistrol Mula

Funding statement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825.

Competing interests statement

None declared

Ethics approval

This study was based entirely on a systematic review of relevant published literature and did not require an ethics approval.

Patient consent

This study did not require consent from patients as it uses no individual data.

Permission to reproduce material from other sources

This study has cited all references which are published and publicly available.

Data sharing statement

The study protocol was published on the online platform Zenodo (https://zenodo.org/record/3770937). Copies of searches and data extraction sheets will be made publicly available on Zenodo as part of the database collection for all scoping reviews conducted in the PERMIT project.

Figure legends

Fig. 1. Keyword based search strategy for the scoping review. Four categories of keywords were defined to retrieve relevant articles from the biomedical literature on machine learning analyses of omics data for personalised medicine, which include a validation study (highlighted by the colored boxes in the center). For each category relevant keywords were determined, including controlled vocabulary terms from the Medical Subject Headings (MeSH) thesaurus by the US National Library of Medicine (upper and lower boxes with frames colored according to the corresponding category). As indicated by the keyword "AND" in the center, a conjunctive search was conducted, i.e., every retrieved article had to contain at least one keyword from each category. This strategy was adapted for searching the other databases.

Fig. 2. Study selection flow diagram. Flow diagram of the procedure for the scoping review article identification, screening, eligibility assessment, and final inclusion, according to the PRISMA scheme (17). Reasons for excluding full-text were not mutually exclusive.

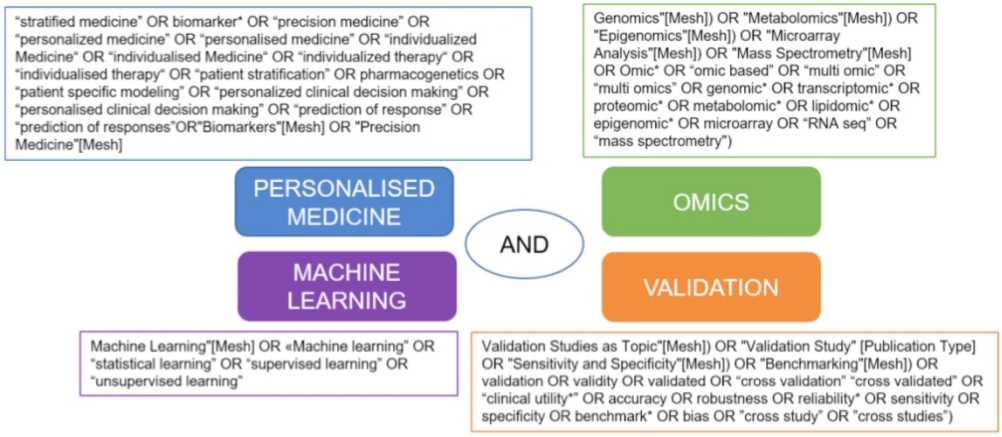
Fig. 3. Validation methods used in omics biomarker studies. Stacked bar chart of the number of articles retrieved in the scoping review for different categories of validation methods used in the underlying biomarker studies (covering time periods from 2000 to 2020). The majority of studies use only internal cohort validation approaches, such as cross-validation (CV), training/test set split validation, out-of-bag validation (for tree-based classifiers), and combinations of CV and test set validation within the same cohort. Studies with an external validation on an independent patient cohort (with or without an additional internal cross-validation) are still underrepresented, even in more recent time periods. All filtered full-text articles derived from the scoping review except for review articles were included in the analysis.

Fig. 4. Characteristics of successful omics-based studies. Six main categories of design and implementation aspects that characterize successful omics-based biomarker development studies were identified (starting from the centre left in the figure and proceeding clockwise): 1) Adequacy of the study design & sample size selection; 2) Rigor and robustness of the statistical evaluation; 3) Clarity of scope and goals; 4) Completeness and reproducibility of the study documentation; 5) Interpretability and biological plausibility of the created predictive models; 6) Integration of prior biological knowledge into the model building and validation procedures.

Tables

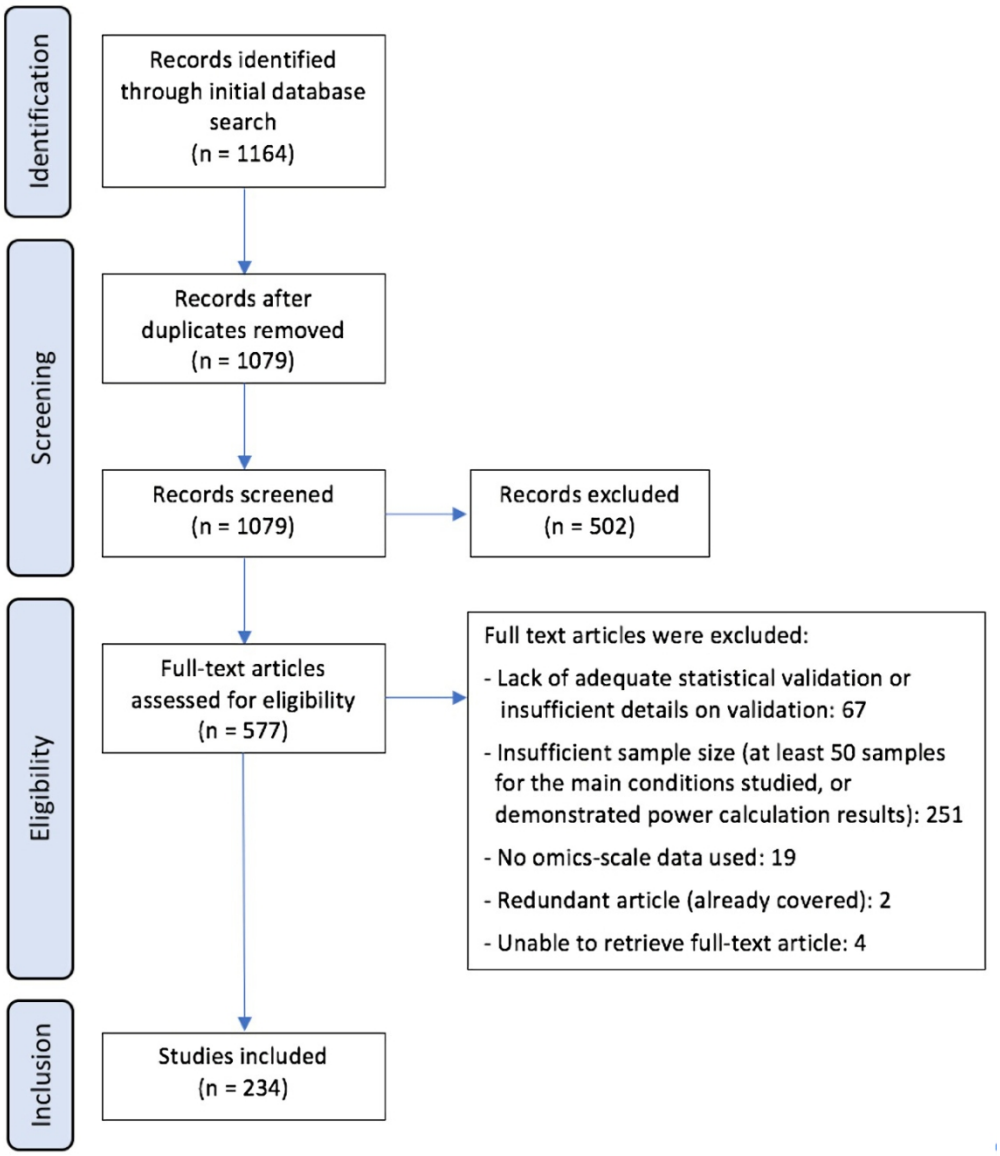
Name	Test approval (FDA-cleared and/or LDT)	Purpose	References
MammaPrint	FDA-cleared, LDT	breast cancer risk-of-recurrence assessment	(1,20–23)
ColoPrint	LDT	colon cancer development of distant metastasis prediction	(24–29)
Prosigna Assay / PAM50	FDA-cleared, LDT	breast cancer risk of distant recurrence prediction	(31–35)
Oncotype DX	LDT	breast cancer risk-of-recurrence assessment	(3,37–40)
Decipher	LDT	prostate cancer metastatic risk prediction	(4,44–48)
Cancer Type ID	LDT	predict tumor type for cancers of unknown / uncertain diagnosis	(10,49–51)
Afirma™ Gene Expression Classifier	LDT	discriminate between benign and cancerous thyroid nodules	(52–57)
Foundation One Heme	LDT	test for hematologic malignancies, sarcomas, or solid tumors	(9,58–60)
AlloMap Heart	FDA-cleared, LDT	identifying heart transplant recipients with risk of cellular rejection	(8,62–64)
Corus CAD	LDT	identify obstructive coronary artery disease	(6,65–68)

Tab. 1. Examples of clinically approved omics-derived diagnostic or prognostic tests designs applied to personalised medicine (synonyms for the same test are separated by the "/"-symbol). FDA-approval status was checked on the web-site https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests and reflects the status as of Oct. 22nd 2020.



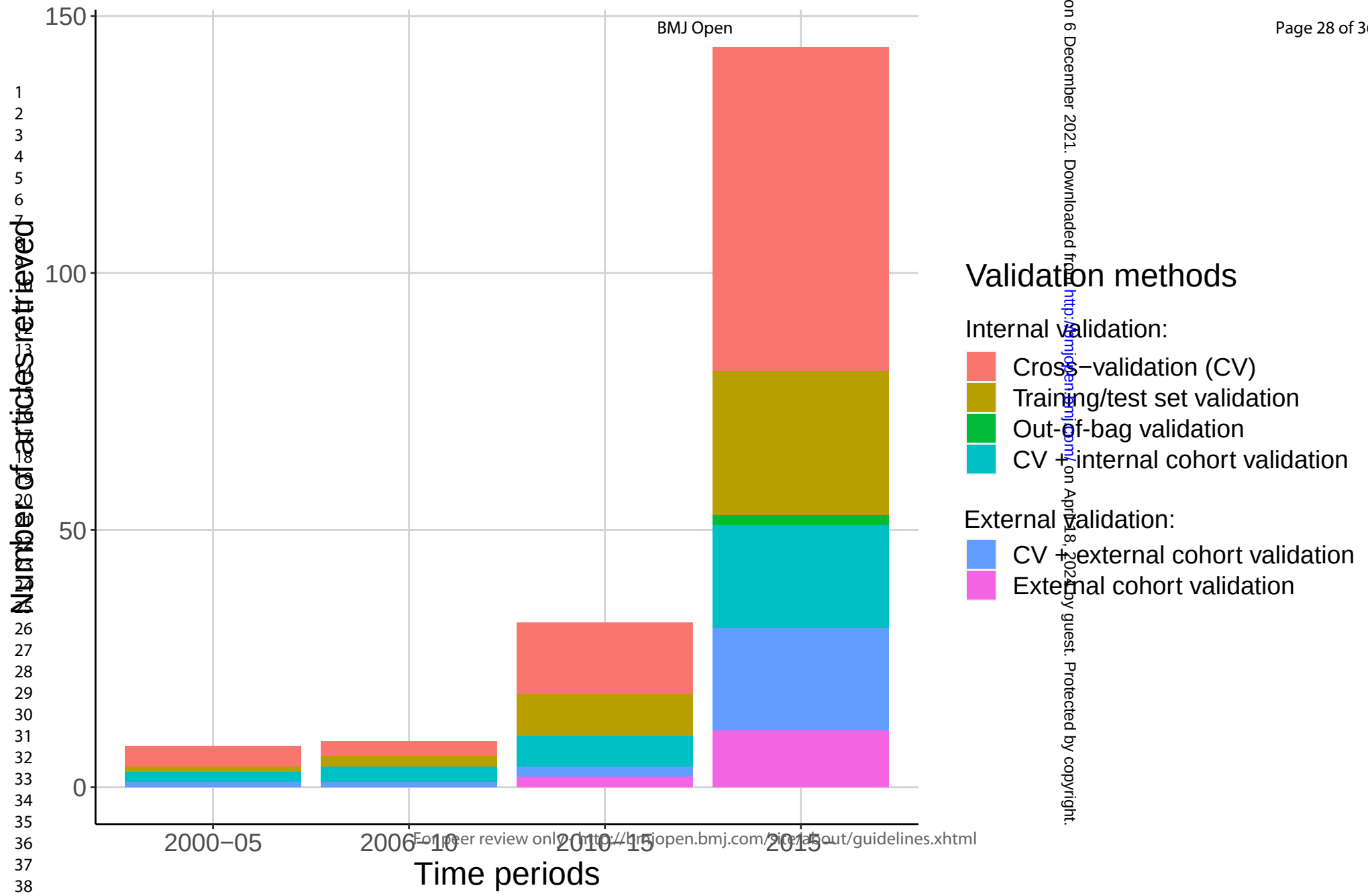
Keyword based search strategy for the scoping review. Four categories of keywords were defined to retrieve relevant articles from the biomedical literature on machine learning analyses of omics data for personalised medicine, which include a validation study (highlighted by the colored boxes in the center). For each category relevant keywords were determined, including controlled vocabulary terms from the Medical Subject Headings (MeSH) thesaurus by the US National Library of Medicine (upper and lower boxes with frames colored according to the corresponding category). As indicated by the keyword "AND" in the center, a conjunctive search was conducted, i.e., every retrieved article had to contain at least one keyword from each category. This strategy was adapted for searching the other databases.

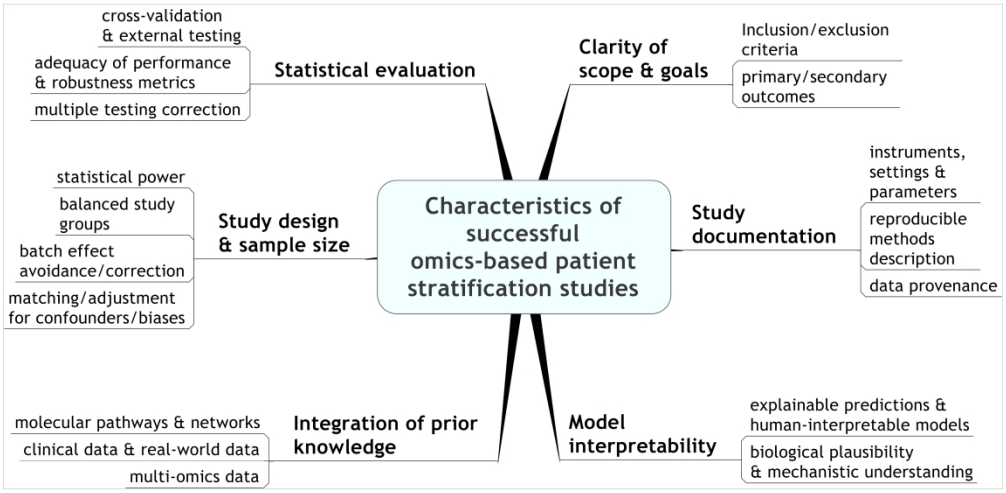
164x71mm (300 x 300 DPI)



Study selection flow diagram. Flow diagram of the procedure for the scoping review article identification, screening, eligibility assessment, and final inclusion, according to the PRISMA scheme. Reasons for excluding full-text were not mutually exclusive.

109x127mm (300 x 300 DPI)





Characteristics of successful omics-based studies. Six main categories of design and implementation aspects that characterize successful omics-based biomarker development studies were identified (starting from the centre left in the figure and proceeding clockwise): 1) Adequacy of the study design & sample size selection; 2) Rigor and robustness of the statistical evaluation; 3) Clarity of scope and goals; 4) Completeness and reproducibility of the study documentation; 5) Interpretability and biological plausibility of the created predictive models; 6) Integration of prior biological knowledge into the model building and validation procedures.

338x165mm (300 x 300 DPI)

Online Supplementary file 2 – Search strategy

Precise queries and number of items retrieved for each query for the keyword searches conducted in the databases *PubMed*, *EMBASE* and *Web of Science* as part of the scoping review.

1) PubMed Query

No.	Query	Items found
#14	Search **#4 AND #7 AND #10 AND #13** Sort by: **PublicationDate**	365
#13	Search **#11 OR #12** Sort by: **PublicationDate**	2480122
#12	Search **(Validation Studies as Topic"[Mesh]) OR "Validation Study" [Publication Type] OR "Sensitivity and Specificity"[Mesh]) OR "Benchmarking"[Mesh])**	0
#11	Search **(validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR ”cross study” OR ”cross studies”)**	2480122
#10	Search **#8 OR #9** Sort by: **PublicationDate**	982942
#9	Search **("Genomics"[Mesh]) OR "Metabolomics"[Mesh]) OR "Epigenomics"[Mesh]) OR "Microarray Analysis"[Mesh]) OR "Mass Spectrometry"[Mesh])**	422277
#8	Search **(Omic* OR “omic based” OR “multi omic” OR “multi omics” OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR “RNA seq” OR “mass spectrometry")**	944832
#7	Search **#5 OR #6** Sort by: **PublicationDate**	743829
#6	Search **("Biomarkers"[Mesh]) OR "Precision Medicine"[Mesh])**	743829
#5	Search **(stratified medicine” OR cluster* OR “sub group*” OR Subgroup* OR biomarker* OR diagnos* OR prognos* OR “precision medicine” OR “personalized medicine”OR “personalised medicine” OR “individualized Medicine“ OR “individualised Medicine“ OR “individualized therapy“ OR “individualised therapy“)**	0
#4	Search **#2 OR #3** Sort by: **PublicationDate**	40640
#3	Search **"Machine Learning"[Mesh]** Sort by: **PublicationDate**	16481
#2	Search **(“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)** Sort by: **PublicationDate**	34840

2) Embase Query

No.	Query	Items found
#25	#23 AND #24	688
#24	[embase]/lim NOT [medline]/lim	9568801
#23	#20 AND #21 AND ([english]/lim OR [french]/lim OR [italian]/lim OR [spanish]/lim)	1423
#22	#20 AND #21	1433
#21	omic*:ti,ab OR 'machine learning':ti,ab OR 'personalized medicine':ti,ab OR 'personalised medicine':ti,ab	59092
#20	#4 AND #10 AND #16 AND #19	4830
#19	#17 OR #18	6287177
#18	validation:ti,ab OR validity:ti,ab OR validated:ti,ab OR 'cross validation':ti,ab OR 'cross validated':ti,ab OR test*:ti,ab OR 'clinical utility*':ti,ab OR accuracy:ti,ab OR robustness:ti,ab OR reliability*:ti,ab OR sensitivity:ti,ab OR specificity:ti,ab OR benchmark*:ti,ab OR bias:ti,ab OR 'cross study:ti,ab' OR 'cross studies':ti,ab	6150811
#17	'validation study'/exp OR 'reliability'/exp OR 'sensitivity and specificity'/exp OR 'benchmarking'/exp	580344
#16	#14 OR #15	1174400
#15	omic*:ti,ab OR 'omic based':ti,ab OR 'multi omic*':ti,ab OR genomic*:ti,ab OR transcriptomic*:ti,ab OR proteomic*:ti,ab OR metabolomic*:ti,ab OR lipidomic*:ti,ab OR epigenomic*:ti,ab OR microarray:ti,ab OR 'rna seq':ti,ab OR 'mass spectrometr*':ti,ab	852339
#14	#11 OR #12 OR #13	758269
#13	'mass spectrometry'/exp	455591
#12	'microarray analysis'/exp	68369
#11	'omics'/exp OR 'genomics'/exp OR 'epigenetics'/exp	299009
#10	#5 OR #6 OR #7 OR #8 OR #9	5000844
#9	'individualized medicine':ti,ab OR 'individualised medicine':ti,ab OR 'individualized therapy':ti,ab OR 'individualised therapy':ti,ab	3459
#8	'personalised medicine':ti,ab	1713
#7	'personalized medicine':ti,ab	13669
#6	'stratified medicine':ti,ab OR cluster*:ti,ab OR 'sub group*':ti,ab OR subgroup*:ti,ab OR biomarker*:ti,ab OR diagnos*:ti,ab OR prognos*:ti,ab OR 'precision medicine':ti,ab	4904827
#5	'biological marker'/exp OR 'personalized medicine'/exp	330768
#4	#1 OR #2 OR #3	200079
#3	'machine learning'/exp	193633
#2	'statistical learning'/exp	46
#1	'machine learning':ti,ab OR 'statistical learning':ti,ab OR 'supervised learning':ti,ab OR 'unsupervised learning':ti,ab	34557

3) Web of Science Query

No.	Query	Items found
# 7	#6 AND #5 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	193
# 6	**TITLE:** ((omic* OR "machine learning" OR "personalized medicine" OR "personalised Medicine")) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	33,111
# 5	#4 AND #3 AND #2 AND #1 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	1,075
# 4	**TOPIC:** ((validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR ”cross study” OR ”cross studies”)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	5,015,557
# 3	**TOPIC:** ((Omic* OR “omic based” OR “multi omic*” OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR “RNA seq” OR “mass spectrometr*”)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	1,024,118
# 2	**TOPIC:** ((“ stratified medicine” OR cluster* OR “sub group*” OR Subgroup* OR biomarker* OR diagnos* OR prognos* OR “precision medicine” OR “personalized medicine”OR “personalised medicine” OR “individualized Medicine“ OR “individualised Medicine“ OR “individualized therapy“ OR “individualised therapy“)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	4,137,042
# 1	**TOPIC:** ((“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	134,956

Online Supplementary file 3 – Data extraction form

Data items extracted from each processed article during the full-text scoping review, and associated qualifications for each item.

Item	Qualifications
Authors	
Title	
Journal	
Volume	
Issue	(if applicable)
Pages	(if applicable)
Year	
Location	
URL / DOI	
Type of publication	• Research article • Meeting abstract • Review
Study population and sample size	(if applicable)
Methodology / Study Design	• Case-control study • Cases only stratification study (+ further qualification, e.g. treatment response prediction, tumor subtype categorization, recurrence/relapse prediction, survival prediction, tissue-of-origin prediction)
Outcome assessment	• Performance measures (e.g. accuracy, sensitivity, specificity, Kohen's Kappa, F-score, AUC) • Validation scheme (cross-validation approach, external validation approach, single cohort or multiple cohorts)
Generic machine learning category	• Supervised learning • Unsupervised learning • Other / mixed approaches
Name of specific machine learning approach	(if applicable)

Main results / key findings that relate to the research question	short description

Online Supplementary file 1 – PRISMA-ScR Checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (17).

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	2
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	(16)
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5-6
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	5
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	6

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	6 (Online Suppl. File 2)
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	Click here to enter text.
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	6
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	7 (Fig. 2)
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	7 (Table 1)
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	Click here to enter text.
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	7-11
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	7-11
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	11-12
Limitations	20	Discuss the limitations of the scoping review process.	11-12
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and	12-13

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
		objectives, as well as potential implications and/or next steps.	
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	20

BMJ Open

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053674.R1
Article Type:	Original research
Date Submitted by the Author:	12-Aug-2021
Complete List of Authors:	Glaab, Enrico; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Rauschenberger, Armin; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Banzi, Rita; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Gerardi, Chiara; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Garcia, Paula; ECRIN, European Clinical Research Infrastructure Network Demotes, Jacques; ECRIN, European Clinical Research Infrastructure Network
Primary Subject Heading:	Patient-centred medicine
Secondary Subject Heading:	Diagnostics, Research methods
Keywords:	BIOTECHNOLOGY & BIOINFORMATICS, NATURAL SCIENCE DISCIPLINES, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Authors

Enrico Glaab[1,*], Armin Rauschenberger[1], Rita Banzi[2], Chiara Gerardi[2], Paula Garcia[3], Jacques Demotes-Mainard[3], and the PERMIT Group

Affiliations

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg.

[2]Center for Health Regulatory Policies, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

[3]European Clinical Research Infrastructure Network (ECRIN), Paris, France

*Correspondence: enrico.glaab@uni.lu; 14, avenue du Rock'n'Roll, L-4361 Esch-sur-Alzette, Luxembourg; Phone: +352-466644 6186

Word count

Abstract: 300

Main text: 6641

Keywords

Biomarkers, Scoping Review, Omics, Machine Learning, Stratification

Abstract

Objective: To review biomarker discovery studies using omics data for patient stratification which led to clinically validated FDA-cleared tests or laboratory developed tests, in order to identify common characteristics and derive recommendations for future biomarker projects.

Design: Scoping review.

Methods: We searched PubMed, EMBASE and Web of Science to obtain a comprehensive list of articles from the biomedical literature published between January 2000 to July 2021, describing clinically validated biomarker signatures for patient stratification, derived using statistical learning approaches. All documents were screened to retain only peer-reviewed research articles, review articles, or opinion articles, covering supervised and unsupervised machine learning applications for omics-based patient stratification. Two reviewers independently confirmed the eligibility. Disagreements were solved by consensus. We focused the final analysis on omics-based biomarkers which achieved the highest level of validation, i.e., clinical approval of the developed molecular signature as a laboratory developed test or FDA approved tests.

Results: Overall, 352 articles fulfilled the eligibility criteria. The analysis of validated biomarker signatures identified multiple common methodological and practical features that may explain the successful test development and guide future biomarker projects. These include study design choices to ensure sufficient statistical power for model building and external testing, suitable combinations of non-targeted and targeted measurement technologies, the integration of prior biological knowledge, strict filtering and inclusion/exclusion criteria, and the adequacy of statistical and machine learning methods for discovery and validation.

Conclusions: While most clinically validated biomarker models derived from omics data have been developed for personalised oncology, first applications for non-cancer diseases show the potential of multivariate omics biomarker design for other complex disorders. Distinctive characteristics of prior success stories, such as early filtering and robust discovery approaches, continuous improvements in assay design and experimental measurement technology, and rigorous multi-cohort validation approaches, enable the derivation of specific recommendations for future studies.

Strengths and limitations of this study

- This scoping review provides an overview of biomarker discovery studies using machine learning analysis of omics data which have led to clinically validated diagnostic and prognostic tools.
- The review discusses shared characteristics of successful biomarker studies as a guidance for study design, discovery and validation method choices for future projects.
- Data extraction and analysis methods focus on deriving recommendations to optimize the design of prospective studies and improve analysis workflows for retrospective studies.
- The review applied minimum eligibility criteria for sample size and statistical validation, but did not assess the quality of the included studies.

Introduction

Personalised medicine is a rapidly developing area in health care research and practice, which aims at providing more effective and safer therapies tailored to the individual patient, by exploiting subject-specific molecular, clinical and environmental data sources (Box 1).

A central tool in personalised medicine and the focus of this study is the machine learning (ML) analysis of omics profiling data to derive molecular biomarker signatures for disease- or drug-based patient stratification (1). The major goals for ML-based omics biomarker development are to develop more reliable and robust tests for drug response prediction, early diagnosis, differential diagnosis or prognosis of the future clinical disease course (2). Omics-derived biomarker signatures may help to guide treatment decisions, and to focus therapies on the right populations to prevent overtreatment, increase success rates, and reduce costs (3). As a research and information tool, they may enable a better monitoring of disease progression and treatment success, and guide new drug development and discovery (4). In contrast to classical single-molecule biomarker approaches, omics signatures have the potential to provide more sensitive, specific and robust predictions of disease-associated outcomes (5).

However, while biomarker discovery projects using omics data have already led to the successful development of clinically validated diagnostic and prognostic tests (6–15), many biomarker studies are discontinued after early development stages or fail in later clinical validation stages. Dedicated statistical and ML methodologies for omics biomarker discovery and validation have been published, as well as recommendations for study design, implementation and reporting (16,17). The distinctive features and approaches which characterize prior successes in translating omics research findings into clinically validated tests have however not yet been investigated in detail. In order to guide future projects on suitable method choices, there is a need for dedicated studies on the key determinants of previous translational successes in ML-based omics biomarker development.

As part of an EU project on “Personalised Medicine Trials” (PERMIT (18)), funded within the H2020 framework, we have therefore investigated the current methodological practices for personalised medicine, covering ML approaches for omics-based patient stratification as a major focus area. While a broader series of questions was established and examined for the overall scoping review (19), for this manuscript, we focused our analysis on biomarker discovery studies that have led to successful, clinically validated FDA-cleared tests or laboratory developed tests (LDTs), to determine their shared and distinctive characteristics compared to studies with no clinical translation. In particular, we aimed to address the following specific research questions:

- Which omics-derived biomarker discovery studies have led to clinically validated tests for patient stratification (LDTs or FDA-cleared tests)?
- What are the key characteristics shared by successful omics biomarker studies and distinguishing them from previously published biomarker studies which have not yet led to clinically validated tests?
- Which types of model building and validation methods have been used to develop clinically validated biomarker signatures, and what are the lessons learned and recommended workflows?
- Which recommendations and guidelines have been proposed to address common challenges in biomarker development using omics data?

These questions lend themselves to a scoping review, because omics-derived biomarker development is still an evolving field, and a preliminary assessment of the potential scope and size of the available biomedical literature on these topics is required as a first step for further follow-up research. Therefore, the objective of this study was to address the above questions by retrieving and examining the current literature on biomarker discovery and validation studies using omics data and ML approaches.

Methods

We conducted a scoping review following the methodological framework suggested by the Joanna Briggs Institute (20). This framework consists of six stages: 1) identifying the research questions, 2) identifying relevant studies, 3) study selection, 4) charting the data, 5) collating, summarising and reporting results, and 6) consultation.

The scoping review approach was considered most suitable to respond to the broad scope and the evolving nature of the field. Compared to systematic reviews that aim to answer specific questions, scoping reviews present a general overview of the evidence pertaining to a topic and are useful to examine emerging trends, to clarify key concepts and identify gaps (21,22). Before conducting the review, a study protocol was published on the online platform Zenodo (19). Due to the iterative nature of scoping reviews, deviations from the protocol are expected and duly reported when occurred. We used the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist to report our results (23) (Online supplementary file 1).

Study identification

Relevant studies and documents were identified, balancing feasibility with breadth and comprehensiveness of searches. We searched PubMed, EMBASE and Web of Science (last search date: July 27, 2021) for articles describing supervised or unsupervised ML analyses for biomarker discovery or personalised medicine, including both discovery and validation methods. The relevance of the search methodology was ensured by using a strict multi-stage filtering, considering only articles including at least one relevant search term per category from four categories of keywords ("Personalized medicine / Biomarkers", "Omics", "Machine Learning" and "Validation", covering both synonyms for these terms and closely related keywords, see Fig. 1, illustrating the keyword-based search strategy, and Online Supplementary file 2 for the detailed search queries), and subsequently post-filtering the retrieved articles manually to exclude studies not involving omics-based biomarker research or lacking a description of machine learning and validation analyses (see sections on *Eligibility criteria* and *Study selection*). To cover only relevant scientific content, the scope was limited to journal publications and meeting abstracts from international conferences and workshops, and no other grey literature was included. We restricted inclusion to reports published from January 2000 to July 2021 (covering also "online first" articles with official publication date in the future) in English, French, Spanish, Italian and German language. Since to the best of our knowledge, the first clinically validated FDA-cleared omics-derived biomarker signature was published in 2002 (24), only few preliminary discovery studies were expected to have taken place significantly earlier than 2002, and we therefore did not extent the search further backwards in time than January 2000.

Eligibility criteria

We included peer-reviewed methodology articles, review articles, opinion articles on supervised and unsupervised ML methods for omics disease prediction and stratification and associated statistical cross-validation and multi-cohort validation methods (addressing accuracy, robustness, and clinical

relevance). Only approaches tested on real-world biomedical omics data were reviewed, while studies relying purely on simulated data or were excluded. We also excluded papers on biomarker methods without a demonstrated biomedical application, and those with insufficient sample size (i.e., removing studies covering less than 50 samples per group for the main conditions studied, unless a dedicated power calculation was presented) or statistical validation (i.e., lack of clear descriptions of cross-validation or external testing methodology, performance metrics and test statistics). These exclusion criteria were not specified in the generic review protocol, but they were agreed among the authors prior to the screening process.

To cover both data from original research papers and prior systematic reviews, we extracted information from three main article types: (1) applied research papers, (2) methodology articles with demonstrated applications, and (3) review articles on methods, applications and validation approaches.

Apart from these inclusion and exclusion criteria, for the final result presentation, the statistical investigations covered all selected articles, whereas the detailed discussion of study characteristics focused on the studies that led to clinically validated biomarker signatures tested on multiple cohorts with large sample sizes (i.e., studies using a power calculation to demonstrate the adequacy of the chosen sample sizes, or covering hundreds or thousands samples per studied subject group).

Study selection

We exported the references retrieved from the searches into the online tool Rayyan (25). Duplicates were removed automatically using the reference manager Endnote X9 (Clarivate Analytics, Philadelphia, United States) and manually by the reviewers. One reviewer loaded the retrieved records into the online screening tool Rayyan (25), and two reviewers confirmed the eligibility independently by covering both the screening for all records and the full-text review for the articles pre-selected by the screening. Disagreements were solved by consensus.

Charting the data and synthesis of results

We designed a data extraction form using Excel (Online supplementary file 3). General study characteristics extracted covered author names, title, citation, type of publication (e.g., journal article, meeting abstract), study population and sample size (if applicable), methodology/study design, and outcome measures (if applicable). Specific items associated with the topic of the scoping review included the study type (e.g., Case-control study, differential diagnosis study, prognostic study, review – methods, review – applications, review – validation); the article type (journal or conference article), the generic ML domain (e.g., supervised/unsupervised); and the name of specific approaches for outcome prediction and for validation. Moreover, to capture key findings related to the review questions, relevant sentences were extracted from each reviewed article, and if needed, complemented by a brief explanatory remark, and by writing out abbreviations used in the original text.

The reviewers piloted the data extraction form using five records from the retrieved article collection. Two reviewers (EG, AR) working independently extracted the data from the included articles. In the case of disagreements, consensus was obtained by discussion.

In the final full-text review stage, the pre-selected articles were grouped by topic, categorizing articles into applied vs. methodological studies, supervised vs. unsupervised analyses, and assigning algorithm type identifiers to each article (review articles and papers on validation methodologies were considered as separate categories without a specific algorithm type assignment). The full-text

review and categorization of articles into different publication types was done through independent manual inspection by the two reviewers.

While the information on sample sizes and validation methods was documented as part of the data extraction, it was not within the remit of this scoping review to assess the methodological quality of individual studies included in the analysis.

Consultation exercise

The members of the PERMIT consortium, associated partners, and the PERMIT project Scientific Advisory Board discussed the preliminary findings of the scoping review in a 2-hour online workshop.

Patient and public involvement

The European Patients' Forum is a member of PERMIT project. Although not directly involved in the conduction of the scoping review, they received the draft review protocol for collecting comments and feedback.

Results

Study selection and general characteristics of reports

We retrieved 1563 abstracts from the literature search. After the removal of duplicates, we screened the remaining 1475 abstracts for eligibility. 619 records were excluded, while 856 abstracts were retained for the full-text assessment. Finally, we included 352 articles that passed all filtering criteria in the data extraction and analysis (see flow chart in Fig. 2).

The full-text article review revealed that many studies did not meet the pre-defined inclusion criteria: 371 articles (43%) were removed because of an insufficient sample size, and 105 further articles (12%) were excluded because they provided insufficient details on the validation results or methodology (see Fig. 2). This shows that the challenges of recruiting an adequate number of participants per study group or conducting sufficient omics profiling experiments for robust model building and validation are not met in a large proportion of omics biomarker studies. Moreover, many studies lack adequate documentation for the study design and validation.

For the selected articles that cover primary research on omics biomarker studies, the majority (77%) rely entirely on an internal validation involving data from only a single cohort, whereas studies that use an external validation on an independent cohort are still underrepresented (only 12% of articles describe both an internal cross-validation and an external cohort validation, and an additional 11% include an external validation, but do not report internal cross-validation results). However, when comparing the numbers of published studies over different periods of time during the past 20 years, the relative proportion of studies including an external validation has increased in recent years (see Fig. 3), suggesting a growing recognition of the importance of independent, multi-cohort validation.

Next, we investigated the countries of origin for the selected articles, showing that the United States of America (USA) are contributing the largest proportion of validated biomarker studies (28%), followed by China (18%), Canada (5%), Germany (4%), and the United Kingdom and India (both 3%; see also Fig. 4, providing a map visualization of the country statistics). These country representations show limited correlation with population sizes and may largely reflect worldwide variation in relative biomedical research productivity reviewed in previous study (26). Since the most

prolific countries in the development of molecular diagnostics have already set up policies and regulations for omics- and ML-based *in vitro* diagnostics and medical devices (e.g., see the life cycle regulation of AI- and ML-based software devices in the USA (27)), they may also provide a role model for countries still in the process of establishing similar regulatory frameworks.

When inspecting the representation of study design types in the filtered article collection, the great majority of documents described diagnostic studies (67%), prognostic and survival prediction studies were covered in 8% of articles, and studies examining therapy or drug response in 7% (see Fig. 5). Apart from this, 13% of articles were reviews on methodologies and applications in the field, and 5% of articles described other rare study types (e.g. tissue-of-origin prediction studies or combinations of different study types).

Since a detailed discussion of all filtered articles is not within the scope of the present review, in the following, we focus on reviewing representative omics biomarker studies which achieved the highest validation level, i.e., clinical approval of the developed molecular signature as an LDT or FDA approved test (see the overview of studies in Table 1 and the FDA web-site (28)). We investigate the shared features of these successful studies, examine how they address common shortcomings and missing features of other reviewed studies, and summarize the lessons learned.

Success stories in omics-based biomarker signature development

Cancer approved omics-derived diagnostic tests (9 studies)

The first and most well-known omics-derived molecular test to receive FDA clearance was *MammaPrint*®, a prognostic signature using the RNA expression activity of 70 genes to estimate the risk for distant tumours metastasis and recurrence in early-stage breast cancer patients (6,24,29–32). This test was developed at the Netherlands Cancer Institute, using DNA microarray analysis to investigate primary breast tumours of 117 patients. Supervised ML was applied to the resulting data to identify a highly predictive gene signature for a short interval to distant metastases in lymph node negative patients (24).

A distinctive feature of the development approach behind this signature in comparison to other reviewed studies was the multi-stage filtering and cross-validation strategy used in the initial discovery study, which may explain the repeated confirmation of the signature in later validation studies (6,29–32). From 25k genes represented on the DNA microarrays, only those significantly regulated in more than 3 tumours out 78 sporadic lymph-node negative patients were preselected, and further filtered by retaining only the genes with a minimum absolute correlation with the disease outcome of 0.3. The resulting list of 231 genes, rank-ordered by absolute correlation, was investigated by sequentially adding the next top 5 genes from the list to a candidate ML classifier and evaluating its performance by leave-one-out cross-validation (LOOCV). This procedure was repeated as long as the estimated accuracy of the classifier improved, providing a final candidate signature of 70 genes. The final signature was validated on multiple independent test sets, including a set of 19 external samples in the original study and several additional validations on independent cohorts in follow-up studies (6,29–32).

The *MammaPrint* signature provided the role model for the subsequent development of a similar prognostic test for colon cancer, *ColoPrint*® (33–38). This test aims at detecting the approx. 20% of patients with stage II colon cancer expected to experience a relapse and develop distant metastases. It uses an 18-gene expression signature, developed by analysing DNA microarray data in a similar manner to the *MammaPrint* approach. The diagnostic approach has been commercialized as an

LDT to assist physicians in selecting treatment options for colon cancer patients. Similar to *MammaPrint*, the signature development was characterized by extensive discovery and validation studies, which involved multiple statistical reproducibility, stability and precision analyses for independent, large-scale patient cohorts (39).

Another widely used cancer-related LDT, which received FDA clearance in 2013, is the *Prosigna*® *Breast Cancer Prognostic Gene Signature Assay*, previously called *PAM50* test (40–44). This assay assesses mRNA expression for a signature of 58 genes (50 target genes + 8 endogenous control genes) to predict the risk of distant recurrence for hormone-receptor-positive breast cancer between 5 to 10 years after diagnosis (prerequisites are that the patients have been treated with hormonal therapy and surgery, and are stage I or stage II lymph-node negative, or in stage II with one to three positive nodes). The test development started with a microarray discovery study and involved a multistage filtering, using consecutive applications of statistical tests and cross-validation to propose a subset of candidate gene markers (45). The authors compared the reproducibility of classification scores obtained with these markers for three centroid-based prediction methods to ensure the robustness of the methodology. By further developing the approach into a more sensitive PCR-based test, and later into an assay using the NanoString nCounter Dx Analysis System, the predictive performance was improved in a step-wise fashion. The original discovery study was characterized by significantly larger sample sizes than the majority of reviewed biomarker studies, with a training set of 189 samples, test sets of 761 patients evaluated for prognosis, and 133 patients evaluated for prediction of pathologic complete response to treatment with taxane and anthracycline. These study design features in combination with multi-stage filtering and validation approaches, and improved measurement technology during the course of the study, may explain the successful progression of the PAM50 test to FDA clearance.

Among the LDTs for breast cancer prognosis, *Oncotype DX*® is a further test commonly used in clinical practice (8,46–49). The underlying gene signature consists of 16 cancer-associated genes and 5 reference genes, and is therefore often also referred to as '21-gene assay'. Its main application is to predict risk of recurrence in oestrogen-receptor positive tumours. The relevance of this prognostic tool for treatment selection may be explained by the strong association of the provided recurrence score with the probability of positive treatment response to chemotherapy (50). *Oncotype DX* was developed using a consecutive refinement procedure, starting with the RT-PCR assessment of 250 candidate genes across 447 patients from three distinct studies to identify the 21-gene signature after multiple filtering steps. A recurrence score algorithm built using the signature as input was clinically validated on 668 independent patients (51). The selection of the 16 cancer-related genes included in the assay involved scoring the performance of the candidate features in all three studies and the consistency of the primer/probe performance in the assay (52).Thus, particular strengths of the development process for this LDT include the consideration of both technical robustness and statistical robustness of the assay across distinct cohorts. However, an independent comparative clinical validation of Oncotype DX and the PAM50 signature for estimating the likelihood of distant recurrence in ER-positive, node-negative, post-menopausal breast cancer patients treated with endocrine therapy suggested that the PAM50 signature provided more prognostic information than Oncotype DX (53).

While the first validated omics biomarker signatures were developed for breast cancer, similar diagnostic and prognostic tools have followed for other cancer types. One of these is the Decipher® Prostate Cancer Test (9,54–58), which differs from other omics-derived diagnostic tools by being provided together with a software platform and database, the Decipher Genomic Resource Information Database (GRID), that captures 1.4 million expression markers per patient to facilitate personalised care. The test itself uses 22 preselected RNAs to predict clinical metastasis and cancer-specific mortality for patients who have undergone radical prostatectomy. An initial discovery study by the Mayo Clinic (Rochester, MN, USA) investigated a cohort of 545 such patients, split into a training (n = 359) and a validation cohort (n = 186). Similar to other LDTs, the discovery started

with a genome-wide profiling and used both statistical and ML analyses for filtering. First, *t*-tests were applied (reduction from 1.4 mil. to 18,902 differentially expressed RNAs), then regularized logistic regression (reduction to 43 candidate markers), and finally a random forest-based feature selection (reduction to final set of 22 RNAs). Apart from testing the signature in the validation cohort, further external validations were performed in subsequent studies (9,54–58). Overall, distinctive strengths of the used approach include the improved interpretability of the test results through supporting analyses on the GRID platform, and the robustness of the discovery and validation approach, involving large sample sizes and several complementary statistical and ML assessments.

While most diagnostic tests in oncology have been designed for specific cancer types, a dedicated LDT has also been developed for cancers of unknown or uncertain diagnosis. The Cancer Type ID® test by bioTheranostics distinguishes between 50 different tumour types using a 92-gene RT-PCR expression measurement signature (15,59–61). This signature was derived from analyses of a microarray data collection covering 446 frozen tumour samples and 112 formalin-fixed, paraffin-embedded (FFPE) samples of both primary and metastatic tumours. Modelling steps involved k-nearest neighbour clustering and classification, and a genetic algorithm to explore the search space of possible feature subset selections. After successful cross-validation (84% accuracy) and external validation (82% accuracy on 112 independent FFPE samples), the microarray-based signature was further developed to use more sensitive RT-PCR measurements. Testing the new approach on an independent validation set provided an increased accuracy (87%). Distinctive characteristics of the development process that may have contributed to the positive validation include the efficient and extensive exploration of the search space of possible gene subset selections via a genetic algorithm, the large sample sizes used for discovery and validation, and the transfer of the assay from microarrays to the more sensitive RT-PCR platform.

The first omics-derived biomarker signatures addressed only the most frequent cancer types, but more recent applications in oncology focus on the diagnosis of less common malignancies, such as thyroid cancer. Typically, deciding whether a thyroid nodule is benign or cancerous is possible via a fine needle aspiration (FNA) biopsy, without requiring more complex measurements or analyses. However, while direct FNA-based diagnosis is feasible in most cases, indeterminate results can occur (62). To help prevent unnecessary surgeries for the corresponding patients, a molecular signature and LDT known as the Afirma™ Gene Expression Classifier (GEC) has been developed to discriminate benign from cancerous thyroid nodules (62–67). The original discovery study behind the GEC signature used mRNA expression analysis in 315 thyroid nodules, covering 178 retrospective surgical tissues and 137 prospectively collected FNA samples. Two ML classifiers were trained separately on surgical tissues and FNAs, assessing the test set performance on 48 independent, prospective FNA samples (50% of which had indeterminate cytopathology). Discriminative features were selected using a linear modelling approach implemented in the software Limma, and a linear support vector machine was applied for model building and performance estimation via 30-fold cross-validation (CV). The successful cross-validation results were confirmed on multiple distinct cohorts (62,65–68). While the internal validation used in the initial study cannot address cohort-specific biases, the combined use of established feature selection and modelling approaches, and the subsequent external validation across multiple cohorts with large sample sizes may account for the successful translation of this signature.

Most omics-based diagnostic tests identified in our study rely purely on gene expression profiling data. However, more recently, first multi-omics signatures for diagnostic purposes have been developed. One of the first LDTs that integrated information from both RNA and DNA sequencing was the FoundationOne® Heme assay (14,69–71). This assay aims to detect hematologic malignancies, sarcomas, pediatric malignancies, or solid tumours (including among others leukaemias, myelodysplastic syndromes, myeloproliferative neoplasms, lymphomas, multiple myeloma, Ewing sarcoma, Leiomyosarcoma, and paediatric tumours). The test identifies four types of genomic alterations (base substitutions, insertions and deletions, copy number alterations,

rearrangements) and reports microsatellite instability and tumour mutational burden to facilitate clinical decision making. This approach was originally developed and evaluated using reference samples of pooled cell lines in order to model the main characteristics that determine the test accuracy, including mutant allele frequency, indel length and amplitude of copy change (69). A first validation using 249 independent FFPE cancer samples, which had already been characterized by established assays, confirmed the accuracy of the test. External validation studies on independent cohorts corroborated the utility of the test for further diagnostic applications (14,72). The study results highlight the potential of integrating diverse biological data sources in order to obtain more robust and reliable predictions, a strategy that may be promising in particular for complex disorders that involve very heterogeneous phenotypes.

A common limitation of genomic profiling approaches for diagnostic testing is that most analyses have to be performed in centralized specialty laboratories, which limits a wider use and results in long waiting times. To address this shortcoming, the Elio™ Tissue Complete assay, an *in vitro* diagnostic test cleared in 2020 by the FDA for assessing somatic mutations and tumour mutation burden (TMB) in solid tumours, has been developed as an integrated DNA-to-report approach to enable a decentralized evaluation in all diagnostic labs with next generation sequencing (NGS) technology (73). The analytical performance of the test was assessed by comparing it with the FoundationOne test (see above) using a concordance analysis on 147 tumour specimens. It provided a positive percent agreement (PPA) above 95% for single nucleotide variants (SNVs) and insertions/deletions, and 80-83% PPA for copy number alterations and gene translocations (73). The test has recently also been applied to investigate the response to immune checkpoint inhibitors (ICI) in metastatic renal cell carcinoma (mRCC), using a retrospective evaluation of SNVs, TMB, microsatellite status and genomic status of antigen presentation genes (74). While no correlation between treatment response and TMB was observed, one third of patients with progressive disease following ICI therapy displayed loss of heterozygosity of major histocompatibility complex class I genes (LOH-MHC) vs. 6% of disease control patients, suggesting that loss of antigen presentation may restrict ICI response (74). In summary, the Elio Tissue Complete assay provides an example of how integrating NGS analyses with bioinformatics in a combined DNA-to-report approach could help to broaden the access to genomic diagnostics for both clinical and research applications.

Non-cancer approved omics-derived diagnostic tests (4 Studies)

While most clinically approved omics-derived diagnostic tests have been developed in the field of oncology, one of the first LDTs that received FDA clearance for a non-cancer disease was the AlloMap® Heart test (13,75–77). It uses a gene expression signature of 11 target genes and 9 control genes in peripheral blood from heart transplant recipients to estimate the risk for acute cellular cardiac allograft rejection. The development process involved statistical analyses of leukocyte microarray profiling data from 285 samples, and subsequent RT-PCR validation and bioinformatics post-processing (13). Prior knowledge from database and literature mining was included in the analysis by mapping the data to known alloimmune pathways. This allowed the researchers to narrow down 252 candidate marker genes. An RT-PCR validation on 145 samples confirmed 68 of these candidate genes, which distinguished rejection samples from quiescent samples according to a T-test ($p < 0.01$). Six genes were eliminated due to significant variation in gene expression with sample processing time. Next, the investigators averaged correlated gene expression levels to create robust meta-level features, called ‘metagenes’, and added 20 of these features as new variables. A linear discriminant analysis was applied, providing a prediction model using four individual genes and three metagenes, which aggregate information from 11 original genes. Finally, bootstrap validation procedures and external test set validations were performed to confirm the accuracy of this signature. Overall, distinctive aspects of the development approach for the AlloMap signature include the knowledge-based gene discovery, a comprehensive RT-PCR validation of candidate genes, and the robust bootstrap and external validation analyses.

The first clinically validated LDT for a cardiovascular indication derived from omics data was the Corus® CAD test, developed to identify coronary artery disease (CAD) in stable non-diabetic patients (11,78–81). In contrast to most other omics-based tests, Corus CAD is not a pure molecular signature test, but takes the clinical covariates gender and age into account. The initial discovery study used a retrospective microarray analysis of blood samples from 195 diabetic and non-diabetic patients from the Duke University CATHGEN registry. After ranking the studied genes by the statistical significance of group differences and prior biological knowledge on their disease relevance, 88 genes were selected for RT-PCR validation. Because diabetes status as a clinical covariate was significantly associated with the observed gene expression alterations, and the identified CAD-associated genes did not overlap between diabetic and non-diabetic patients, the authors decided to limit follow-up work to non-diabetic patients. In a prospective clinical trial, microarray profiling was conducted on blood samples from 198 patients, and top-ranked genes were further validated using RT-PCR for 640 blood samples. After multiple filtering steps, taking into account statistical significance in T-tests, biological relevance, gene correlation clustering and cell-type analyses, a final signature of 23 genes was derived, composed of 20 CAD-associated genes and 3 reference genes (82). To maximize the predictive performance, the final prediction algorithm was optimized to adjust for differences associated with age and gender. Compared to most other reviewed studies, the Corus CAD approach stands out by taking clinical covariates into account in the final prediction model, including an intermediate critical review and adjustment of the inclusion criteria (limiting the focus to nondiabetic patients), and integrating complementary filtering and validation analyses on large sample sizes.

For inflammatory diseases, a first omics-derived signature recently received approval for measuring rheumatoid arthritis (RA) inflammatory disease activity, the Vectra® DA multi-biomarker test (83–87). It uses blood serum samples and multi-spot 96-well immunoassay plates to assess serum concentrations of 12 protein biomarkers associated with the pathobiology of RA. The original Vectra DA score, which combines these measurements into a composite score between 1 and 100, was assessed via multivariate regression and displayed a high predictive power in estimating a standard RA score, the Disease Activity Score in 28 joints using the C-reactive protein level (DAS28-CRP), in both seropositive (AUC 0.77, $P < 0.001$) and seronegative (AUC 0.70, $P < 0.001$) patients (87). This score was later adjusted for age, gender and adiposity (based on leptin concentration), and validated in two cohorts against DAS28-CRP as a prognostic test for radiographic progression during the next year. The results showed that the new adjusted score was the most accurate, independent predicator of progression, with the rate of progression increasing from < 2% in the low (1-29) adjusted score category to 16% in the high (45-100) category (85). Overall, the Vectra DA approach illustrates the utility of omics-based biomarker signatures for prognostic applications in inflammatory disorders, and further highlights the benefit of integrating omics signatures with information from clinical covariates.

For neurodegenerative disorders, clinically approved diagnostic and prognostic omics-derived tests are still lacking. However, recently the Helix® Genetic Health Risk App for Late-onset Alzheimer's Disease (AD) was cleared by the FDA for over-the-counter use. It detects clinically relevant variants in genomic DNA isolated from human saliva of individuals ≥18 years in order to report and interpret genetic health risks, and evaluates the information of variants with established genome-wide significant associations to AD. When tested on 99 human saliva samples, the accuracy was 100% with a lower 95% CI bound of 96.3% (88). The approach uses a whole exome sequencing (WES) constituent device, the Helix® Laboratory Platform (89–91), as a qualitative *in vitro* diagnostics approach covering measurements for approximately 20k genes. The Helix Laboratory Platform has received FDA clearance through a new regulatory approval pathway established by the FDA for WES devices (Regulation 21 CFR 866.6000). Due to the generic applicability of the WES profiling assay used by this platform, called Exome+, the assay has also been applied to find statistically significant gene-based associations for several other phenotypes in large-scale cohort studies (89) and to identify carriers of autosomal dominant diseases by population-based genetic screening (91).

Thus, the Helix Laboratory Platform provides a first example for a new approval pathway for omics-based diagnostic tests, in which a clinically approved genomic testing device is not anymore linked to a single diagnostic application or a specific disease type. Instead, the market authorization for diagnostic tests is obtained separately from the device and facilitated and accelerated by the prior approval of the constituent measurement device. For the future development of omics-derived biomarker signatures, this may allow researcher to focus on demonstrating the clinical utility of a new signature, while the analytical validity of the underlying testing device has already been established previously.

Discussion

Statement of principal findings

The scoping review of articles on patient stratification using omics data revealed common limitations in the study design for many published biomarker development projects, such as insufficient and imbalanced sample sizes per study group and inadequate validation methods, but also identified multiple studies that have led to validated diagnostic and prognostic tests. These success stories were investigated in more detail to identify common characteristics in the study design, discovery and validation methods, which may have supported the clinical translation of the initial findings. Fig. 6 outlines key shared aspects that are possible determinants of the study success and could help to guide future biomarker investigations. In particular, they cover the following main features:

(1) A sample size selection, study group and replicate design that provides adequate statistical power for the ML analyses;

2) The application of robust statistical filtering and evaluation schemes (including multiple layers of statistical and ML-based feature selection, combined statistical and biological filters, robust validation schemes that involve multiple cross-validation, bootstrapping and external validation analyses, using multiple suitable and complementary performance metrics, and providing information on the statistical variation and confidence intervals for the performance estimates);

3) Clarity of the study scope and goals (involving clear inclusion and exclusion criteria, primary and secondary outcomes, and decision processes to make necessary adjustments due to new knowledge gained during the project, such as the adjusted inclusion criteria in the Corus CAD study and the progression from non-targeted microarray technology to higher-sensitivity RT-PCR in the case of the Prosigna test and the Cancer Type ID test);

4) Completeness and reproducibility of the study documentation (covering details on used instruments, parameters and settings, reproducible methods descriptions, and information on data provenance);

5) Interpretability and biological plausibility of the created predictive models (including explainable and justifiable predictions, human-interpretable model descriptions, and biologically plausible models that agree with the current mechanistic understanding of the studied disorder);

6) Integration of prior biological knowledge into the predictive feature selection, model building and validation procedures (e.g., using public data on disease-associated molecular pathways and networks; complementary clinical and real-world data, and relevant multi-omics data).

Strengths and Limitations

The majority of methodological recommendations derived from the study relate to the early planning and study design for biomarker discovery projects, involving considerations associated with the

choice of the study group, sampling and blocking design, the measurement technology, and the input and output variables (16,17). These recommendations are therefore mainly applicable to prospective studies. For retrospective biomarker investigations of already collected data, the suggestions derived from the review are limited to guidance on improving analysis workflows, e.g. for filtering and evaluation analyses, the integration of prior knowledge from multi-omics data and public annotation databases, and the choice of robust and interpretable modelling approaches for the generation of biologically plausible and reproducible prediction models. While the focus of the review on studies that have already led to validated biomarker models and that fulfil minimum requirements for sample size and statistical model assessment helps to ensure the quality of the selected articles, no further quality evaluation was performed. Finally, more recent methodological developments in the machine learning and cross-validation analysis of omics data, such as meta-learning (92) and bolstered cross-validation (93), have only limited coverage among the articles that passed the eligibility criteria, and will therefore require further dedicated study in the future.

Discussing important differences in results

Previous reviews of ML approaches using omics data for patient stratification have focused on domain-specific analyses for specific types of diseases, or specific types of ML methodologies (94–102). By contrast, this scoping review focuses on disease-agnostic workflows with generic applicability across complex human disorders involving multifactorial molecular alterations. The coverage of statistical and ML approaches for stratification does not aim to provide a detailed discussion of specific algorithms, statistical methods or scoring metrics, but rather at identifying key determinants of success for generic analysis and validation workflows in biomedical stratification studies. Therefore, the results describe general workflow characteristics that distinguish omics biomarker studies with clinical translation from other studies, and cover associated disease-agnostic recommendations for future studies, whereas method recommendations specific to particular disease types or ML analysis types are covered elsewhere in domain-specific reviews (94–102).

Meaning of the study: implications for clinicians and policymakers

The previous clinical translation successes in omics-based biomarker development reviewed in this study, which have mostly been achieved in the field of oncology, highlight the potential for developing similar biomarker signatures for further disease indications. In contrast to conventional statistical biomarker discovery approaches, which focus on identifying single-molecule markers, systems-level analysis of omics data using multivariate ML approaches can identify multifactorial signatures which are robust against noise in individual gene or protein measurements, and more biologically insightful by reflecting disease-associated cellular process alterations in a more comprehensive fashion.

This scoping review has identified common characteristics of omics studies which have led to clinically validated diagnostic and prognostic tests. Thus, the conclusions drawn on recommended practices for sample size selection, biological data filtering and ML, and the implementation of adequate validation schemes may help to guide clinical researchers on study design choices and the selection of analysis methodologies. Additionally, the scoping review results can help to raise awareness of common pitfalls, such as issues associated with batch effects, biases, confounding factors, lack of statistical power, and multiple hypothesis testing, and thus contribute to preventing these failure causes in biomarker development. For policymakers and funding bodies, findings on the distinctive characteristics of studies with successful clinical biomarker translation, e.g. concerning the specific requirements for robust cross-validation and external result validation methods, may provide relevant information for the design of public and private funding schemes for biomedical research. Risks in funded research projects may be addressed upfront through appropriate guidelines and regulations for the study design and validation (e.g. recommendations

on power calculations and specific validation and documentation requirements). Finally, the scoping review results can guide clinicians involved in biomarker discovery on how to make better use of available public knowledge and data sources, e.g. cellular pathway and molecular interaction databases, that may allow them to exploit prior knowledge effectively, and create more robust and interpretable biomarker models.

Unanswered questions & future research

Since the recommendations and guidelines identified from the reviewed articles are mostly derived from established biomarker discovery and validation approaches, new methodologies and upcoming trends could only be covered to a limited extent and may lead to changed recommendations in the future. In particular, in the reviewed patient stratification studies, some of more recently introduced ML concepts (e.g. transfer learning, distance metric learning, semi-supervised learning, structured machine learning, meta learning, multi-view learning, and generative models), data processing techniques (e.g. new dimension reduction approaches, outlier removal methods, data augmentation techniques), and model validation methods (e.g. bootstrapping or bolstered cross-validation, uncertainty quantification), are still underrepresented among the eligible studies reviewed, and may provide suitable topics for follow-up research.

Overall, while the currently available literature on validated stratification biomarkers already provides ample information on common pitfalls and established practices, the development of widely accepted standard guidelines on methodologies for omics biomarker discovery will require further knowledge exchange and deliberation among stakeholders in the field. In particular, integration of domain-specific expertise in discussions involving clinicians, experimental and data scientists, and regulatory and legal experts is required as a follow-up effort to derive comprehensive methodological guidelines for future biomarker development.

Acknowledgments

The authors thank Vanna Pistotti for her assistance with search strategy development and conduction.

List of Figures

Figure 1: Keyword based search strategy for the scoping review

Figure 2: Study selection flow diagram

Figure 3: Validation methods used in omics biomarker studies

Figure 4: Map representation of country statistics for the selected articles

Figure 5: Representation of study types among the selected articles

Figure 6: Characteristics of successful omics-based studies

Definitions (In boxes)

Box 1: *What is Personalised Medicine?*

According to the European Council Conclusion on personalised medicine for patients, personalised medicine is 'a medical model using characterisation of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention (103).

In the context of the Permit project, we applied the following common operational definition of personalised medicine research: a set of comprehensive methods (methodology, statistics, validation, technology) to be applied in the different phases of the development of a personalised approach to treatment, diagnosis, prognosis, or risk prediction. Ideally, robust and reproducible methods should cover all the steps between the generation of the hypothesis (e.g., a given stratum of patients could better respond to a treatment), its validation and pre-clinical development, and up to the definition of its value in a clinical setting (19).

References

1. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49(107739).
2. Goecks J, Jalili V, Heiser LM, Gray JW. How Machine Learning Will Transform Biomedicine. Vol. 181, Cell. 2020. p. 92–101.
3. Jiang Y, Wang M. Personalized medicine in oncology: Tailoring the right drug to the right patient. Biomarkers in Medicine. 2010.
4. Hopp WJ, Li J, Wang G. Big Data and the Precision Medicine Revolution. Prod Oper Manag. 2018;27(9):1647–64.
5. Glaab E. Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification. Brief Bioinform. 2016;
6. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717–29.
7. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;
8. Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genomics. 2013;14(1).
9. Torres A, Alshalalfa M, Tomlins SA, Erho N, Gibb EA, Chelliserry J, et al. Comprehensive Determination of Prostate Tumor ETS Gene Status in Clinical Samples Using the CLIA Decipher Assay. J Mol Diagnostics. 2017;19(3):475–84.
10. Angell TE, Babiarz J, Barth N, Blevins T, Duh Q, Ghossein RA, et al. Clinical validation of the AFIRMA genomic sequencing braf V600E classifier. Thyroid [Internet]. 2017;27:A50. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L624116485
11. Ladapo JA, Budoff MJ, Sharp D, Kuo JZ, Huang L, Maniet B, et al. Utility of a Precision Medicine Test in Elderly Adults with Symptoms Suggestive of Coronary Artery Disease. J Am Geriatr Soc. 2018;66(2):309–15.
12. Tabari E, Lovejoy AF, Lin H, Bolen CR, Saelee SL, Lefkowitz JP, et al. Molecular characteristics and disease burden metrics determined by next-generation sequencing on circulating tumor DNA correlate with progression free survival in previously untreated diffuse large B-cell lymphoma. Blood [Internet]. 2019;134. Available from: http://dx.doi.org/10.1182/blood-2019-123633
13. Deng MC. The AlloMap™ genomic biomarker story: 10 years after. Clin Transplant. 2017;31(3).
14. He J, Abdel-Wahab O, Nahas MK, Wang K, Rampal RK, Intlekofer AM, et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood. 2016;127(24):3004–14.
15. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med. 2006;130(4):465–73.
16. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013.

17. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.

18. Banzi R, Gerardi C, Fratelli M, Garcia P, Torres T, Abad JMH, et al. Web-page for the Personalized Medicine Trials (PERMIT) project [Internet]. 2020 [cited 2021 Aug 2]. Available from: https://permit-eu.org

19. Banzi R, Gerardi C, Fratelli M, Garcia P, Torres T, Abad JMH, et al. Methodological approaches for personalised medicine: protocol for a series of scoping reviews [Internet]. 10.5281/zenodo.3770937. Available from: https://zenodo.org/record/3770937

20. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13(3):141–6.

21. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, Mcewen SA. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Res Synth Methods. 2014;5(4):371–85.

22. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143.

23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018 Oct;169(7):467–73.

24. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.

25. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.

26. Rahman M, Fukui T. Biomedical research productivity: Factors across the countries. Int J Technol Assess Health Care. 2003;19(1):249–52.

27. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle Regulation of Artificial Intelligence- And Machine Learning-Based Software Devices in Medicine. Vol. 322, JAMA - Journal of the American Medical Association. 2019. p. 2285–6.

28. FDA Center for Devices and Radiological Health. Web-page on Nucleic Acid Based Tests by the Food and Drug Administration (FDA) [Internet]. 2021 [cited 2021 Aug 2]. Available from: https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests

29. Winner BS, Sgroi DC, Ryan PD, Bruinsma TJ, Glas AM, Male A, et al. Analysis of the mamma print breast cancer assay in a predominantly postmenopausal cohort. Clin Cancer Res. 2008;14(10):2988–93.

30. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: Another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn. 2009;9(5):417–22.

31. Sapino A, Roepman P, Linn SC, Snel MHJ, Delahaye LJMJ, Van Den Akker J, et al. MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue. J Mol Diagnostics. 2014;16(2):190–7.

32. Mook S, Knauer M, Bueno-De-Mesquita JM, Retel VP, Wesseling J, Linn SC, et al. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. Ann Surg Oncol. 2010;17(5):1406–13.

33. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, et al. Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. Ann Surg.

2013;257(6):1053–8.

34. Kopetz S, Tabernero J, Rosenberg R, Jiang Z, Moreno V, Bachleitner-Hofmann T, et al. Genomic Classifier ColoPrint Predicts Recurrence in Stage II Colorectal Cancer Patients More Accurately Than Clinical Factors. Oncologist. 2015;20(2):127–33.

35. Tan IB, Tan P. Genetics: An 18-gene signature (ColoPrint®) for colon cancer prognosis. Nat Rev Clin Oncol. 2011;8(3):131–3.

36. Rosenberg R, Maak M, Simon I, Nitsche U, Schuster T, Kuenzli B, et al. Independent validation of a prognostic genomic profile (ColoPrint) for stage II colon cancer (CC) patients. J Clin Oncol. 2011;29(4_suppl):358–358.

37. Salazar R, de Waard JW, Glimelius B, Marshall J, Klaase J, Van Der Hoeven J, et al. The PARSC trial, a prospective study for the assessment of recurrence risk in stage II colon cancer (CC) patients using ColoPrint. J Clin Oncol. 2012;30(4_suppl):678–678.

38. Tabernero J, Moreno V, Rosenberg R, Nitsche U, Bachleitner-Hofmann T, Lanza G, et al. Clinical and technical validation of a genomic classifier (ColoPrint) for predicting outcome of patients with stage II colon cancer. J Clin Oncol. 2012;30(4_suppl):384–384.

39. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;23:ix179.

40. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer. 2014;14(1).

41. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics. 2015;8(1).

42. Alvarado MD, Prasad C, Rothney M, Cherbavaz DB, Sing AP, Baehner FL, et al. A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer. Adv Ther. 2015;32(12):1237–47.

43. Jensen MB, Lænkholm AV, Nielsen TO, Eriksen JO, Wehn P, Hood T, et al. The Prosigna gene expression assay and responsiveness to adjuvant cyclophosphamide-based chemotherapy in premenopausal high-risk patients with breast cancer. Breast Cancer Res. 2018;20(1).

44. Hequet D, Callens C, Gentien D, Albaud B, Mouret-Reynier MA, Dubot C, et al. Prospective, multicenter French study evaluating the clinical impact of the Breast Cancer Intrinsic Subtype-Prosigna® Test in the management of early-stage breast cancers. PLoS One. 2017;12(10).

45. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.

46. Kelly CM, Krishnamurthy S, Bianchini G, Litton JK, Gonzalez-Angulo AM, Hortobagyi GN, et al. Utility of oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. Cancer. 2010;116(22):5161–7.

47. Lo SS, Mumby PB, Norton J, Rychlik K, Smerage J, Kash J, et al. Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection. J Clin Oncol. 2010;28(10):1671–6.

48. Carlson JJ, Roth JA. The impact of the Oncotype Dx breast cancer assay in clinical practice:

A systematic review and meta-analysis. Vol. 141, Breast Cancer Research and Treatment. 2013. p. 13–22.

49. Thakur SS, Li H, Chan AMY, Tudor R, Bigras G, Morris D, et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One. 2018/01/06. 2018;13(1):e0188983.

50. Pease AM, Riba LA, Gruner RA, Tung NM, James TA. Oncotype DX ® Recurrence Score as a Predictor of Response to Neoadjuvant Chemotherapy. Ann Surg Oncol. 2019;

51. Gianni L, Zambetti M, Clark K, Baker J, Cronin M, Wu J, et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J Clin Oncol. 2005;23(29):7265–77.

52. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817–26.

53. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol. 2013;31(22):2783–90.

54. Marrone M, Potosky AL, Penson D, Freedman AN. A 22 gene-expression assay, decipher® (GenomeDx biosciences) to predict five-year risk of metastatic prostate cancer in men treated with radical prostatectomy. PLoS Curr. 2015;7(EVIDENCEONGENOMICTESTS).

55. Nguyen PL, Haddad Z, Lam LLC, Ong K, Buerki C, Deheshi S, et al. Evaluation of the Decipher prostate cancer classifier to predict metastasis and disease-specific mortality from genomic analysis of diagnostic prostate needle biopsy specimens. J Clin Oncol. 2017;35(6_suppl):4–4.

56. Magi-Galluzzi C, Yousefi K, Haddad Z, Palmer-Aronsten B, Lam LLC, Buerki C, et al. Validation of the Decipher prostate cancer classifier for predicting 10-year postoperative metastasis from analysis of diagnostic needle biopsy specimens. J Clin Oncol. 2016;34(2_suppl):59–59.

57. Dalela D, Löppenberg B, Sood A, Sammon J, Abdollah F. Contemporary Role of the Decipher® Test in Prostate Cancer Management: Current Practice and Future Perspectives. Rev Urol. 2016;18(1):1–9.

58. Klein EA, Haddad Z, Yousefi K, Lam LLC, Wang Q, Choeurng V, et al. Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology. 2016;90:148–52.

59. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, et al. Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagnostics. 2013;15(2):263–9.

60. Greco FA, Spigel DR, Yardley DA, Erlander MG, Ma X, Hainsworth JD. Molecular Profiling in Unknown Primary Cancer: Accuracy of Tissue of Origin Prediction. Oncologist. 2010;15(5):500–6.

61. Hainsworth JD, Rubin MS, Spigel DR, Boccia R V., Raby S, Quinn R, et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: A prospective trial of the Sarah cannon research institute. J Clin Oncol. 2013;31(2):217–23.

62. Harrison G, Sosa JA, Jiang X. Evaluation of the Afirma gene expression classifier in repeat indeterminate thyroid nodules. Arch Pathol Lab Med. 2017;141(7):985–9.

63. Chudova D, Wilde JI, Wang ET, Wang H, Rabbee N, Egidio CM, et al. Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol

Metab. 2010;95(12):5296–304.

64. Kim MI, Alexander EK. Diagnostic use of molecular markers in the evaluation of thyroid nodules. Vol. 18, Endocrine Practice. 2012. p. 796–802.

65. Ali SZ, Fish SA, Lanman R, Randolph GW, Sosa JA. Use of the Afirma® gene expression classifier for preoperative identification of benign thyroid nodules with indeterminate fine needle aspiration cytopathology. PLoS Currents. 2013. p. 1–7.

66. McIver B, Castro MR, Morris JC, Bernet V, Smallridge R, Henry M, et al. An independent study of a gene expression classifier (Afirma) in the evaluation of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab. 2014;99(11):4069–77.

67. Lastra RR, Pramick MR, Crammer CJ, LiVolsi VA, Baloch ZW. Implications of a suspicious Afirma test result in thyroid fine-needle aspiration cytology: An institutional experience. Cancer Cytopathol. 2014;122(10):737–44.

68. Kim JY, Park SC, Lee JK, Choi SJ, Hahm KS, Park Y. Novel Antibacterial Activity of β2-Microglobulin in Human Amniotic Fluid. PLoS One. 2012;

69. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31(11):1023–31.

70. Wang K, Sanchez-Martin M, Wang X, Knapp KM, Koche R, Vu L, et al. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. Leukemia. 2017;31(1):151–8.

71. Tarlock K, He J, Zhong S, Ries RE, Bailey M, Morley S, et al. Distinct age-associated genomic profiles in acute myeloid leukemia (AML) using FoundationOne heme. J Clin Oncol. 2016;34(15_suppl):7041–7041.

72. Lieber DS, Kennedy MR, Johnson DB, Greenbowe JR, Frampton GM, Schrock AB, et al. Abstract B16: Validation and clinical feasibility of a Foundation Medicine assay to identify immunotherapy response potential through tumor mutational burden (TMB). In 2017.

73. Lee Deak K, Jackson JB, Valkenburg KC, Keefer LA, Robinson Gerding KM, Angiuoli SV, et al. Next-Generation Sequencing Concordance Analysis of Comprehensive Solid Tumor Profiling between a Centralized Specialty Laboratory and the Decentralized PGDx Elio Tissue Complete Kitted solution. J Mol Diagnostics [Internet]. 2021 Jul;in press. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1525157821002105

74. Labriola MK, Zhu J, Gupta R, McCall S, Jackson J, Kong EF, et al. Characterization of tumor mutation burden, PD-L1 and DNA repair genes to assess relationship to immune checkpoint inhibitors response in metastatic renal cell carcinoma. J Immunother Cancer [Internet]. 2020 Mar;8(1):e000319. Available from: https://jitc.bmj.com/lookup/doi/10.1136/jitc-2019-000319

75. Yamani MH, Taylor DO, Rodriguez ER, Cook DJ, Zhou L, Smedira N, et al. Transplant Vasculopathy Is Associated With Increased AlloMap Gene Expression Score. J Hear Lung Transplant. 2007;26(4):403–6.

76. Yamani MH, Taylor DO, Haire C, Smedira N, Starling RC. Post-transplant ischemic injury is associated with up-regulated AlloMap gene expression. Clin Transplant. 2007;21(4):523–5.

77. Kobashigawa J, Patel J, Azarbal B, Kittleson M, Chang D, Czer L, et al. Randomized Pilot Trial of Gene Expression Profiling Versus Heart Biopsy in the First Year after Heart Transplant: Early Invasive Monitoring Attenuation Through Gene Expression Trial. Circ Hear Fail. 2015;8(3):557–64.

78. Wingrove JA, Daniels SE, Sehnert AJ, Tingley W, Elashoff MR, Rosenberg S, et al.

Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. Circ Cardiovasc Genet. 2008;1(1):31–8.

79. Rosenberg S, Dehais C, Ducray F, Alentron A, Tanguy M, De Reyneis A, et al. OS11.3 Machine learning for better prognostic stratification and driver genes identification in 1p/19q-codeleted grade III gliomas. Neuro Oncol [Internet]. 2017;19(suppl_3):iii22–iii22. Available from: http://dx.doi.org/10.1093/neuonc/nox036

80. Vargas J, Lima JAC, Kraus WE, Douglas PS, Rosenberg S. Use of the Corus® CAD Gene Expression Test for Assessment of Obstructive Coronary Artery Disease Likelihood in Symptomatic Non-Diabetic Patients. PLoS Currents. 2013.

81. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC Med Genomics. 2011;4.

82. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010;153(7):425–34.

83. Brahe CH, Østergaard M, Johansen JS, Defranoux N, Wang X, Bolce R, et al. Predictive value of a multi-biomarker disease activity score for clinical remission and radiographic progression in patients with early rheumatoid arthritis: a post-hoc study of the OPERA trial. Scand J Rheumatol. 2019;

84. Chernoff D, Scott Eastman P, Hwang CC, Flake DD, Wang X, Kivitz A, et al. Determination of the minimally important difference (MID) in multi-biomarker disease activity (MBDA) test scores: impact of diurnal and daily biomarker variation patterns on MBDA scores. Clin Rheumatol. 2019;38(2):437–45.

85. Curtis JR, Weinblatt ME, Shadick NA, Brahe CH, Østergaard M, Hetland ML, et al. Validation of the adjusted multi-biomarker disease activity score as a prognostic test for radiographic progression in rheumatoid arthritis: a combined analysis of multiple studies. Arthritis Res Ther. 2021;

86. Curtis JR, Xie F, Yang S, Danila MI, Owensby JK, Chen L. Uptake and Clinical Utility of Multibiomarker Disease Activity Testing in the United States. J Rheumatol [Internet]. 2019;46(3):237–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30442830

87. Curtis JR, Van Der Helm-Van Mil AH, Knevel R, Huizinga TW, Haney DJ, Shen Y, et al. Validation of a novel multibiomarker test to assess rheumatoid arthritis disease activity. Arthritis Care Res. 2012;64(12):1794–803.

88. Food and Drug Administration. Helix Genetic Health Risk App For Late-Onset Alzheimer's Disease - FDA Review Decision Summary. 2020; Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K192073

89. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun. 2020;

90. Lu JT, Ferber M, Hagenkord J, Levin E, South S, Kang HP, et al. Evaluation for Genetic Disorders in the Absence of a Clinical Indication for Testing: Elective Genomic Testing. Journal of Molecular Diagnostics. 2019.

91. Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. Nat Med. 2020;

92. Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies. Artif Intell

Rev. 2015;

93. Sima C, Braga-Neto UM, Dougherty ER. High-dimensional bolstered error estimation. Bioinformatics. 2011;

94. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews. 2019.

95. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Br Bioinform. 2008/03/04. 2008;9(2):119–28.

96. Grollemund V, Pradat PF, Querin G, Delbot F, Le Chat G, Pradat-Peyre JF, et al. Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions. Front Neurosci [Internet]. 2019;13. Available from: http://dx.doi.org/10.3389/fnins.2019.00135

97. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet [Internet]. 2019;10(MAR). Available from: http://dx.doi.org/10.3389/fgene.2019.00267

98. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. Curr Hypertens Rep. 2018/07/08. 2018;20(9):75.

99. Long NP, Yoon SJ, Anh NH, Nghi TD, Lim DK, Hong YJ, et al. A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer. Metabolomics. 2019/03/05. 2018;14(8):109.

100. Martinez BI, Stabenfeldt SE. Current trends in biomarker discovery and analysis tools for traumatic brain injury. J Biol Eng [Internet]. 2019;13(1). Available from: http://dx.doi.org/10.1186/s13036-019-0145-8

101. Patil S, Awan KH, Arakeri G, Seneviratne CJ, Muddur N, Malik S, et al. Machine learning and its potential applications to the genomic study of head and neck cancer-A systematic review. J Oral Pathol Med. 2019;48(9):773–9.

102. Saini G, Mittal K, Rida P, Janssen EAM, Gogineni K, Aneja R. Panoptic view of prognostic models for personalized breast cancer management. Cancers (Basel) [Internet]. 2019;11(9). Available from: http://dx.doi.org/10.3390/cancers11091325

103. Europen Council. Council conclusions on personalised medicine for patients. Off J Eur Union [Internet]. 2015;431(2):1–4. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015XG1217(01)&from=EN

104. Labriola MK, Zhu J, Gupta R, McCall S, Jackson J, Kong EF, et al. Characterization of tumor mutation burden, PD-L1 and DNA repair genes to assess relationship to immune checkpoint inhibitors response in metastatic renal cell carcinoma. J Immunother Cancer. 2020;8(1).

Author Contributions

Study conception and design: EG, AR.

Methodology: CG, RB.

Data collection and analysis: EG, AR

Original draft preparation: EG

Review and editing: AR, EG, PG, CG, JDM, RB.

Project supervision: PG

Funding acquisition: JDM.

All authors have read and revised the manuscript and approved the final version.

The members of the PERMIT group were involved in the preparation or revision of the joint protocol of the four scoping reviews of the PERMIT series, attended the joint workshop (consultation exercise) and are co-authors of the other scoping reviews of the PERMIT series.

Collaborators

PERMIT group:

1. Antonio L. Andreu
2. Florence Bietrix,
3. Florie Brion Bouvier
4. Montserrat Carmona Rodriguez
5. Maria del Mar Polo-de Santos,
6. Maddalena Fratelli,
7. Rainer Girgenrath,
8. Alexander Grundmann,
9. Josep Maria Haro,
10. Frank Hulstaert,
11. Iñaki Imaz-Iglesia,
12. Setefilla Luengo Matos
13. Emmet McCormack,
14. Albert Sanchez Niubo,
15. Emanuela Oldoni,
16. Raphael Porcher,
17. Vibeke Fosse,
18. Luis M. Sánchez-Gómez,
19. Lorena San Miguel,
20. Cecilia Superchi,
21. Teresa Torres,
22. Anna Monistrol Mula

Funding statement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825.

Competing interests statement

None declared

Ethics approval

This study was based entirely on a scoping review of relevant published literature and did not require an ethics approval.

Patient consent

This study did not require consent from patients, because it does not use individual data.

Permission to reproduce material from other sources

This study has cited all references which are published and publicly available.

Data sharing statement

The study protocol was published on the online platform Zenodo (19). Copies of searches and data extraction sheets will be made publicly available on Zenodo as part of the database collection for all scoping reviews conducted in the PERMIT project.

Figure legends

Fig. 1. Keyword based search strategy for the scoping review. Four categories of keywords were defined to retrieve relevant articles from the biomedical literature on machine learning analyses of omics data for personalised medicine, which include a validation study (highlighted by the coloured boxes in the centre). For each category relevant keywords were determined, including controlled vocabulary terms from the Medical Subject Headings (MeSH) thesaurus by the US National Library of Medicine (upper and lower boxes with frames coloured according to the corresponding category). As indicated by the keyword "AND" in the centre, a conjunctive search was conducted, i.e., every retrieved article had to contain at least one keyword from each category. This strategy was adapted for searching the other databases.

Fig. 2. Study selection flow diagram. Flow diagram of the procedure for the scoping review article identification, screening, eligibility assessment, and final inclusion, according to the PRISMA scheme (23). Reasons for excluding full-text were not mutually exclusive.

Fig. 3. Validation methods used in omics biomarker studies. Stacked bar chart of the number of articles retrieved in the scoping review for different categories of validation methods used in the underlying biomarker studies (covering time periods from 2000 to 2021). The majority of studies use only internal cohort validation approaches, such as cross-validation (CV), training/test set split validation, resampling/bootstrapping-based validation, out-of-bag validation (for tree-based classifiers), and combinations of CV and test set validation within the same cohort. Studies with an external validation on an independent patient cohort (with or without an additional internal cross-validation) are still underrepresented, even in more recent time periods. All filtered full-text articles derived from the scoping review except for review articles were included in the analysis.

Fig. 4. Map representation of country statistics for the selected articles. The number of articles originating from different countries among the studies selected in the full-text review are visualized on a world map representation using a colour gradient from blue (1 article) to red (98 articles = maximum contribution by a single country; using a logarithmic colour gradient scale to highlight differences over a broad value range).

Fig. 5. Representation of study types among the selected articles. The percentage of articles describing case-control studies, therapy/drug response studies, differential diagnosis studies, prognostic and survival prediction studies, as well as review studies and other study types is represented as a pie chart.

Fig. 6. Characteristics of successful omics-based studies. Six main categories of design and implementation aspects that characterize successful omics-based biomarker development studies were identified (starting from the centre left in the figure and proceeding clockwise): 1) Adequacy of the study design & sample size selection; 2) Rigor and robustness of the statistical evaluation; 3) Clarity of scope and goals; 4) Completeness and reproducibility of the study documentation; 5) Interpretability and biological plausibility of the created predictive models; 6) Integration of prior biological knowledge into the model building and validation procedures.

Tables

Name	Test approval type	Purpose	References
MammaPrint®	FDA-cleared Assay	breast cancer risk-of-recurrence assessment	(6,29–32)
ColoPrint®	LDT	colon cancer development of distant metastasis prediction	(33–38)
Prosigna® Assay / PAM50	FDA-cleared Assay	breast cancer risk of distant recurrence prediction	(40–44)
Oncotype DX®	LDT	breast cancer risk-of-recurrence assessment	(8,46–49)
Decipher®	LDT	prostate cancer metastatic risk prediction	(9,54–58)
Cancer Type ID®	LDT	predict tumour type for cancers of unknown / uncertain diagnosis	(15,59–61)
Afirma™ Gene Expression Classifier	LDT	discriminate between benign and cancerous thyroid nodules	(62–67)
Foundation One® Heme	LDT	test for hematologic malignancies, sarcomas, or solid tumours	(14,69–71)
PGDx Elio™ Tissue Complete	FDA-cleared Assay	test to assess somatic mutations and tumour mutation burden for solid tumours	(73,104)
AlloMap® Heart	FDA-cleared Assay	identifying heart transplant recipients with risk of cellular rejection	(13,75–77)
Corus® CAD	LDT	identify obstructive coronary artery disease	(11,78–81)
Vectra® DA	LDT	multi-biomarker blood test for rheumatoid arthritis	(83–86)
Helix® Laboratory Platform & Health Risk App for Late-onset Alzheimer's	FDA-cleared medical device	whole exome sequencing constituent device based for reporting and interpreting general genetic health risks	(89–91)

Tab. 1. Examples of clinically approved omics-derived diagnostic or prognostic tests designs applied to personalised medicine (synonyms for the same test are separated by the "/"-symbol). FDA-approval status was checked on the web-site by the FDA (28) and reflects the status as of July 2021.

"stratified medicine" OR biomarker* OR "precision medicine" OR "personalized medicine" OR "personalised medicine" OR "individualized Medicine" OR "individualised Medicine" OR "individualized therapy" OR "individualised therapy" OR "patient stratification" OR pharmacogenetics OR "patient specific modeling" OR "personalized clinical decision making" OR "personalised clinical decision making" OR "prediction of response" OR "prediction of responses" OR "Biomarkers"[Mesh] OR "Precision Medicine"[Mesh]

Genomics"[Mesh]) OR "Metabolomics"[Mesh]) OR "Epigenomics"[Mesh]) OR "Microarray Analysis"[Mesh]) OR "Mass Spectrometry"[Mesh] OR Omic* OR "omic based" OR "multi omic" OR "multi omics" OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR "RNA seq" OR "mass spectrometry")

PERSONALISED MEDICINE

MACHINE LEARNING

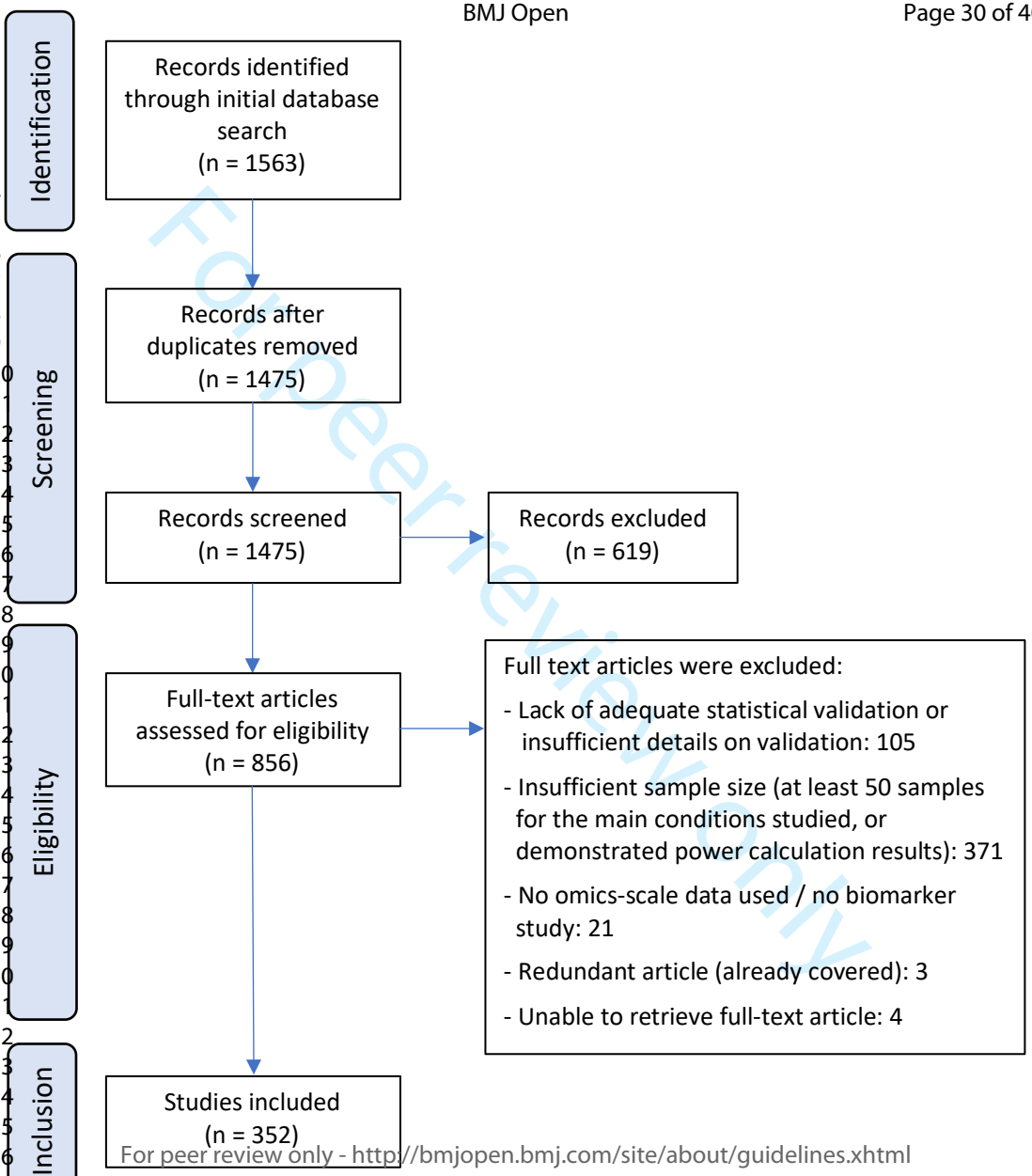
AND

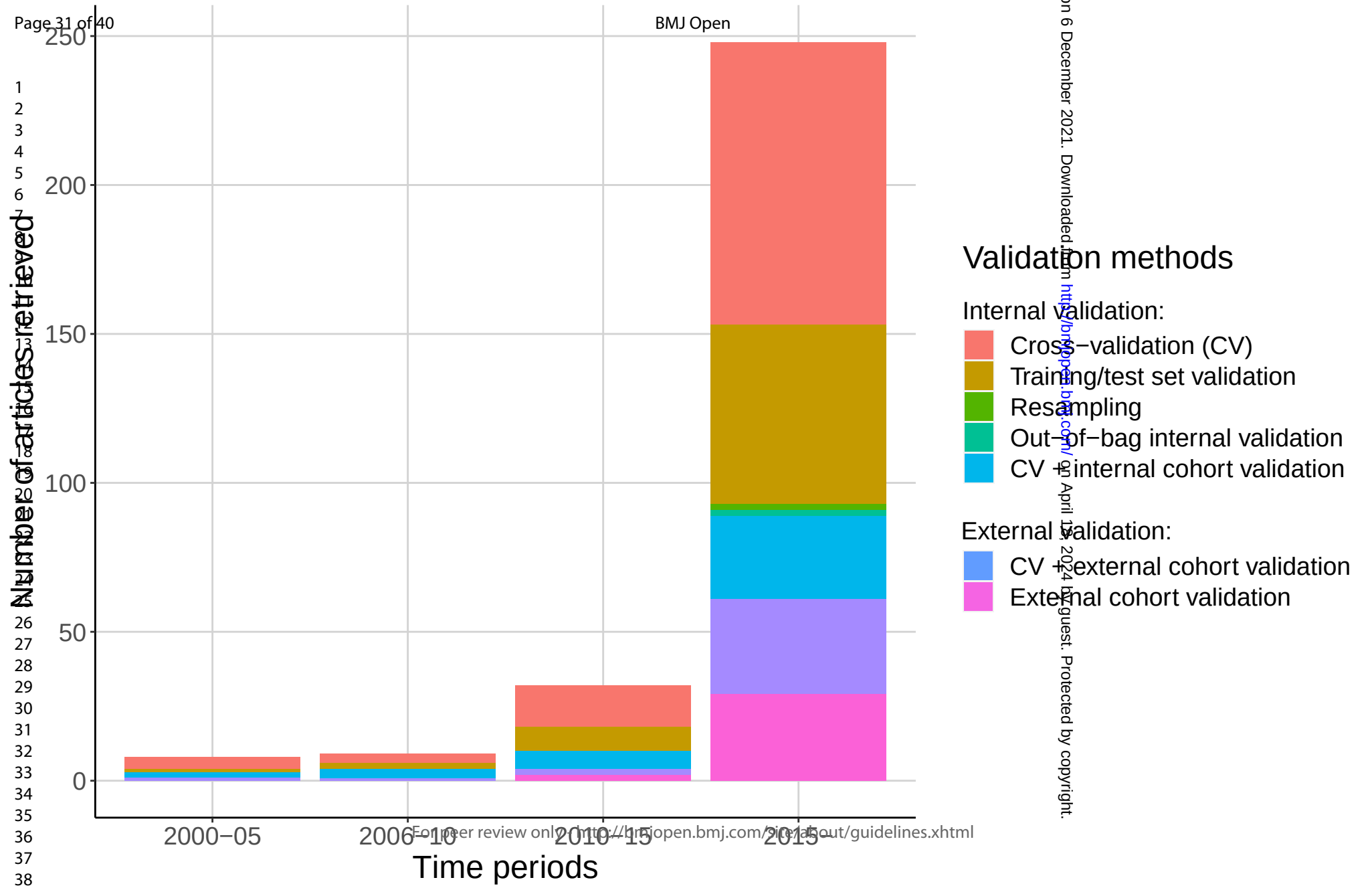
OMICS

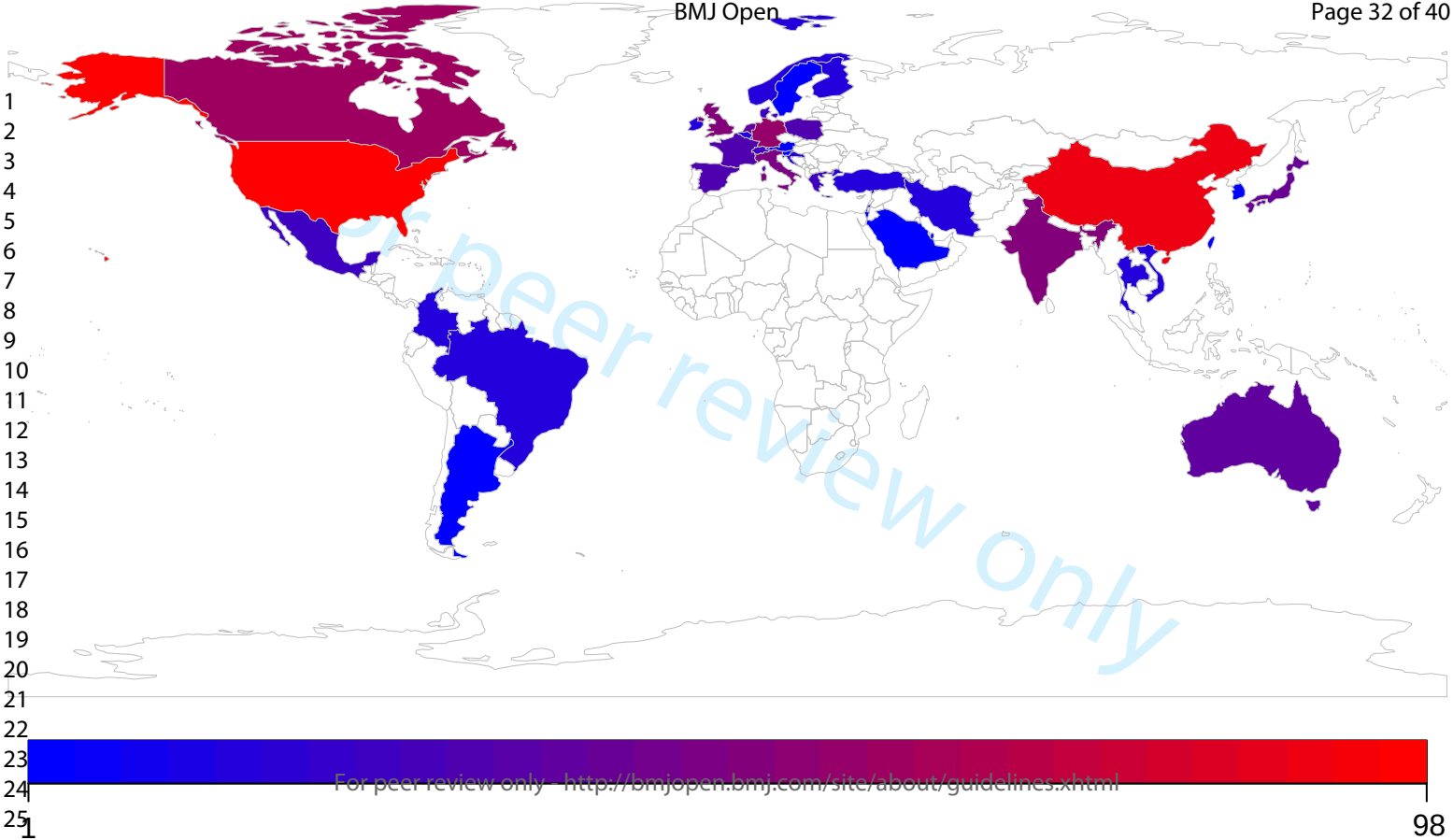
VALIDATION

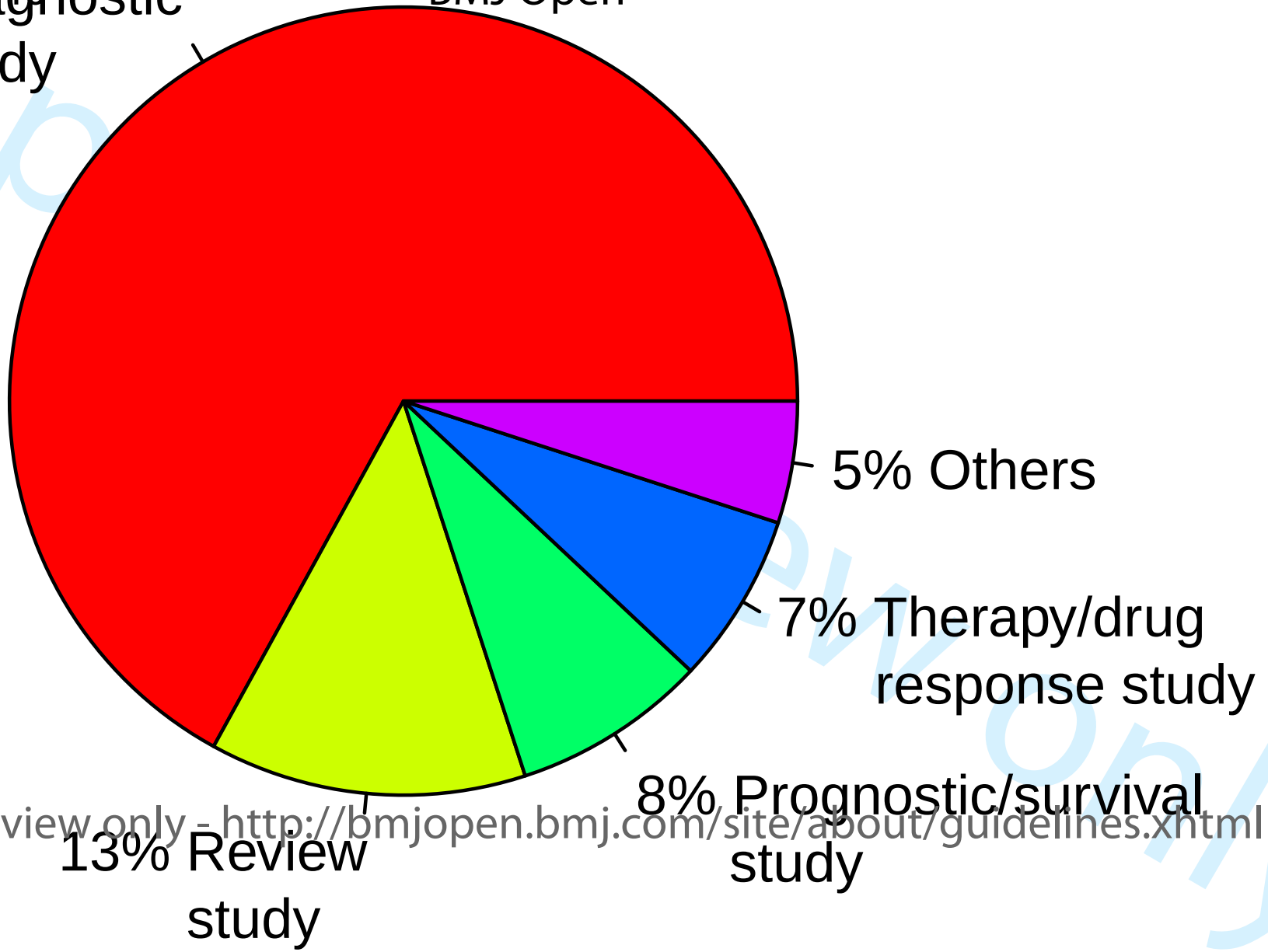
"Machine Learning"[Mesh] OR «Machine learning" OR "statistical learning" OR "supervised learning" OR "unsupervised learning"

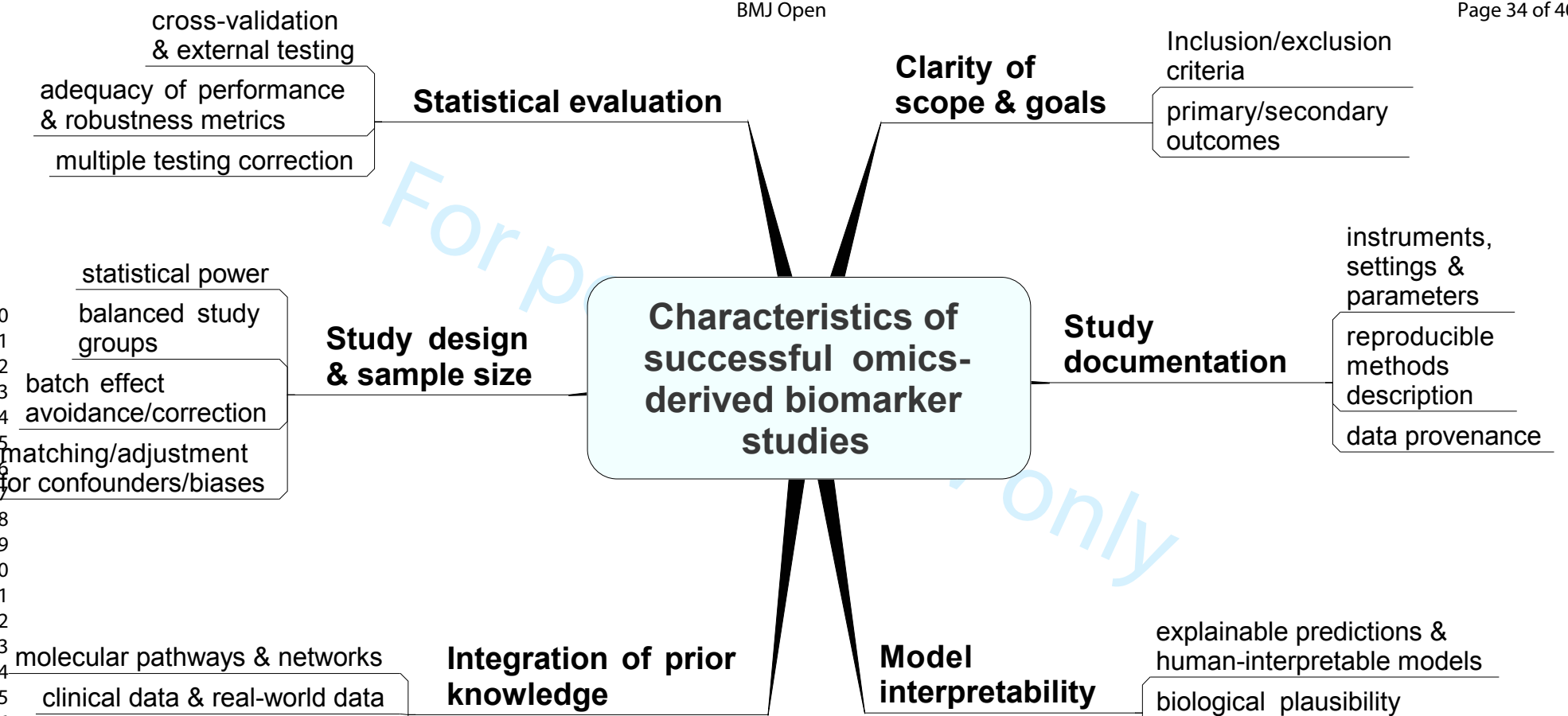
Validation Studies as Topic"[Mesh]) OR "Validation Study" [Publication Type] OR "Sensitivity and Specificity"[Mesh]) OR "Benchmarking"[Mesh]) OR validation OR validity OR validated OR "cross validation" "cross validated" OR "clinical utility*" OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR "cross study" OR "cross studies")











Online Supplementary file 1 – PRISMA-ScR Checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (17).

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	5
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5-6
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	5
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5 (Online Suppl. File 2)
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	6

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	6 (Online Suppl. File 3)
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	Click here to enter text.
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	6-7
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	7 (Fig. 2)
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	7 (Table 1)
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	Click here to enter text.
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	7-13
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	7-13
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	13
Limitations	20	Discuss the limitations of the scoping review process.	13-14
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and	14-15

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
		objectives, as well as potential implications and/or next steps.	
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	24

Online Supplementary file 2 – Search strategy

Keyword searches conducted in the databases *PubMed*, *EMBASE* and *Web of Science* as part of the scoping review.

1) PubMed Query

Search: (((((("Machine learning" OR "statistical learning" OR "supervised learning" OR "unsupervised learning")) OR ("Machine Learning"[Mesh])) AND (("Biomarkers"[Mesh]) OR "Precision Medicine"[Mesh])) AND (((Omic* OR "omic based" OR "multi omic" OR "multi omics" OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR "RNA seq" OR "mass spectrometry")) OR ("Genomics"[Mesh] OR "Metabolomics"[Mesh] OR "Epigenomics"[Mesh] OR "Microarray Analysis"[Mesh] OR "Mass Spectrometry"[Mesh]))) AND ((validation OR validity OR validated OR "cross validation" "cross validated" OR "clinical utility*" OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR "cross study" OR "cross studies"))
AND
(("2000/01/01"[Date - Entry]: "2021/07/20"[Date - Entry])) Filters: English, French, Italian, Spanish

2) Embase Query

#25: #24 AND [embase]/lim NOT [medline]/lim
#24: #23 AND [2000-2021]/py
#23: #20 AND #21 AND ([english]/lim OR [french]/lim OR [italian]/lim OR [spanish]/lim)
#22: #20 AND #21
#21: omic*:ti,ab OR 'machine learning':ti,ab OR 'personalized medicine':ti,ab OR 'personalised medicine':ti,ab
#20: #4 AND #10 AND #16 AND #19
#19: #17 OR #18
#18: validation:ti,ab OR validity:ti,ab OR validated:ti,ab OR 'cross validation':ti,ab OR 'cross validated':ti,ab OR test*:ti,ab OR 'clinical utility*':ti,ab OR accuracy:ti,ab OR robustness:ti,ab OR reliability*:ti,ab OR sensitivity:ti,ab OR specificity:ti,ab OR benchmark*:ti,ab OR bias:ti,ab OR 'cross study:ti,ab' OR 'cross studies':ti,ab
#17: 'validation study'/exp OR 'reliability'/exp OR 'sensitivity and specificity'/exp OR 'benchmarking'/exp
#16: #14 OR #15
#15: omic*:ti,ab OR 'omic based':ti,ab OR 'multi omic*':ti,ab OR genomic*:ti,ab OR transcriptomic*:ti,ab OR proteomic*:ti,ab OR metabolomic*:ti,ab OR lipidomic*:ti,ab OR epigenomic*:ti,ab OR microarray:ti,ab OR 'rna seq':ti,ab OR 'mass spectrometr*':ti,ab
#14: #11 OR #12 OR #13
#13: 'mass spectrometry'/exp
#12: 'microarray analysis'/exp
#11: 'omics'/exp OR 'genomics'/exp OR 'epigenetics'/exp
#10: #5 OR #6 OR #7 OR #8 OR #9
#9: 'individualized medicine':ti,ab OR 'individualised medicine':ti,ab OR 'individualized therapy':ti,ab OR 'individualised therapy':ti,ab
#8: 'personalised medicine':ti,ab
#7: 'personalized medicine':ti,ab
#6: 'stratified medicine':ti,ab OR cluster*:ti,ab OR 'sub group*':ti,ab OR subgroup*:ti,ab OR biomarker*:ti,ab OR diagnos*:ti,ab OR prognos*:ti,ab OR 'precision medicine':ti,ab
#5: 'biological marker'/exp OR 'personalized medicine'/exp
#4: #1 OR #2 OR #3
#3: 'machine learning'/exp
#2: 'statistical learning'/exp

#1: 'machine learning':ti,ab OR 'statistical learning':ti,ab OR 'supervised learning':ti,ab OR 'unsupervised learning':ti,ab

3) Web of Science Query

(((((#1) AND #1) AND #2) AND #3) AND #4) AND #5
5: (ALL=(((validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark*
OR bias OR ”cross study” OR ”cross studies”)))) AND ALL=(((omic* OR "machine learning" OR
"personalized medicine" OR "personalised Medicine")))
4: ALL=(((validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark*
OR bias OR ”cross study” OR ”cross studies”)))
3: ALL=(TOPIC: ((Omic* OR “omic based” OR “multi omic*” OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR “RNA seq”
OR “mass spectrometr*”)))
2: (ALL=(TOPIC: ((“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)))) AND ALL=(TOPIC: ((“stratified medicine” OR cluster* OR “sub group*” OR Subgroup* OR biomarker* OR diagnos* OR prognos* OR “precision medicine” OR
“personalized medicine”OR “personalised medicine” OR “individualized Medicine“ OR “individualised Medicine“ OR “individualized therapy“ OR “individualised therapy“)))
1: ALL=(TOPIC: ((“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)))

Online Supplementary file 3 – Data extraction form

Data items extracted from each processed article during the full-text scoping review, and associated qualifications for each item.

Item	Qualifications
Authors	
Title	
Journal	
Volume	
Issue	(if applicable)
Pages	(if applicable)
Year	
Location	
URL / DOI	
Type of publication	• Research article • Meeting abstract • Review
Study population and sample size	(if applicable)
Methodology / Study Design	• Case-control study • Cases only stratification study (+ further qualification, e.g. treatment response prediction, tumor subtype categorization, recurrence/relapse prediction, survival prediction, tissue-of-origin prediction)
Outcome assessment	• Performance measures (e.g. accuracy, sensitivity, specificity, Kohen's Kappa, F-score, AUC) • Validation scheme (cross-validation approach, external validation approach, single cohort or multiple cohorts)
Generic machine learning category	• Supervised learning • Unsupervised learning • Other / mixed approaches
Name of specific machine learning approach	(if applicable)

Main results / key findings that relate to the research question	short description

BMJ Open

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053674.R2
Article Type:	Original research
Date Submitted by the Author:	06-Nov-2021
Complete List of Authors:	Glaab, Enrico; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Rauschenberger, Armin; University of Luxembourg, Luxembourg Centre for Systems Biomedicine Banzi, Rita; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Gerardi, Chiara; Mario Negri Institute for Pharmacological Research, Center for Health Regulatory Policies Garcia, Paula; ECRIN, European Clinical Research Infrastructure Network Demotes, Jacques; ECRIN, European Clinical Research Infrastructure Network
Primary Subject Heading:	Patient-centred medicine
Secondary Subject Heading:	Diagnostics, Research methods
Keywords:	BIOTECHNOLOGY & BIOINFORMATICS, NATURAL SCIENCE DISCIPLINES, STATISTICS & RESEARCH METHODS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review

Authors

Enrico Glaab[1,*], Armin Rauschenberger[1], Rita Banzi[2], Chiara Gerardi[2], Paula Garcia[3], Jacques Demotes-Mainard[3], and the PERMIT Group

Affiliations

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, Esch-sur-Alzette, Luxembourg.

[2]Center for Health Regulatory Policies, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

[3]European Clinical Research Infrastructure Network (ECRIN), Paris, France

*Correspondence: enrico.glaab@uni.lu; 14, avenue du Rock'n'Roll, L-4361 Esch-sur-Alzette, Luxembourg; Phone: +352-466644 6186

Word count

Abstract: 300

Main text: 7153

Keywords

Biomarkers, Scoping Review, Omics, Machine Learning, Stratification

Abstract

Objective: To review biomarker discovery studies using omics data for patient stratification which led to clinically validated FDA-cleared tests or laboratory developed tests, in order to identify common characteristics and derive recommendations for future biomarker projects.

Design: Scoping review.

Methods: We searched PubMed, EMBASE and Web of Science to obtain a comprehensive list of articles from the biomedical literature published between January 2000 to July 2021, describing clinically validated biomarker signatures for patient stratification, derived using statistical learning approaches. All documents were screened to retain only peer-reviewed research articles, review articles, or opinion articles, covering supervised and unsupervised machine learning applications for omics-based patient stratification. Two reviewers independently confirmed the eligibility. Disagreements were solved by consensus. We focused the final analysis on omics-based biomarkers which achieved the highest level of validation, i.e., clinical approval of the developed molecular signature as a laboratory developed test or FDA approved tests.

Results: Overall, 352 articles fulfilled the eligibility criteria. The analysis of validated biomarker signatures identified multiple common methodological and practical features that may explain the successful test development and guide future biomarker projects. These include study design choices to ensure sufficient statistical power for model building and external testing, suitable combinations of non-targeted and targeted measurement technologies, the integration of prior biological knowledge, strict filtering and inclusion/exclusion criteria, and the adequacy of statistical and machine learning methods for discovery and validation.

Conclusions: While most clinically validated biomarker models derived from omics data have been developed for personalised oncology, first applications for non-cancer diseases show the potential of multivariate omics biomarker design for other complex disorders. Distinctive characteristics of prior success stories, such as early filtering and robust discovery approaches, continuous improvements in assay design and experimental measurement technology, and rigorous multi-cohort validation approaches, enable the derivation of specific recommendations for future studies.

Strengths and limitations of this study

- This scoping review provides an overview of biomarker discovery studies using machine learning analysis of omics data which have led to clinically validated diagnostic and prognostic tools.
- The review discusses shared characteristics of successful biomarker studies as a guidance for study design, discovery and validation method choices for future projects.
- Data extraction and analysis methods focus on deriving recommendations to optimize the design of prospective studies and improve analysis workflows for retrospective studies.
- The review applied minimum eligibility criteria for sample size and statistical validation, but did not assess the quality of the included studies.

Introduction

Personalised medicine is a rapidly developing area in health care research and practice, which aims at providing more effective and safer therapies tailored to the individual patient, by exploiting subject-specific molecular, clinical and environmental data sources (Box 1).

A central tool in personalised medicine and the focus of this study is the machine learning (ML) analysis of omics profiling data to derive molecular biomarker signatures for disease- or drug-based patient stratification (1). The major goals for ML-based omics biomarker development are to develop more reliable and robust tests for drug response prediction, early diagnosis, differential diagnosis or prognosis of the future clinical disease course (2). Omics-derived biomarker signatures may help to guide treatment decisions, and to focus therapies on the right populations to prevent overtreatment, increase success rates, and reduce costs (3). As a research and information tool, they may enable a better monitoring of disease progression and treatment success, and guide new drug development and discovery (4). In contrast to classical single-molecule biomarker approaches, omics signatures have the potential to provide more sensitive, specific and robust predictions of disease-associated outcomes (5).

However, while biomarker discovery projects using omics data have already led to the successful development of clinically validated diagnostic and prognostic tests (6–15), many biomarker studies are discontinued after early development stages or fail in later clinical validation stages. Dedicated statistical and ML methodologies for omics biomarker discovery and validation have been published, as well as recommendations for study design, implementation and reporting (16,17). The distinctive features and approaches which characterize prior successes in translating omics research findings into clinically validated tests have however not yet been investigated in detail. In order to guide future projects on suitable method choices, there is a need for dedicated studies on the key determinants of previous translational successes in ML-based omics biomarker development.

As part of an EU project on “Personalised Medicine Trials” (PERMIT (18)), funded within the H2020 framework, we have therefore investigated the current methodological practices for personalised medicine, covering ML approaches for omics-based patient stratification as a major focus area. While a broader series of questions was established and examined for the overall scoping review (19), for this manuscript, we focused our analysis on biomarker discovery studies that have led to successful, clinically validated FDA-cleared tests or laboratory developed tests (LDTs), to determine their shared and distinctive characteristics compared to studies with no clinical translation. In particular, we aimed to address the following specific research questions:

- Which omics-derived biomarker discovery studies have led to clinically validated tests for patient stratification (LDTs or FDA-cleared tests)?
- What are the key characteristics shared by successful omics biomarker studies and distinguishing them from previously published biomarker studies which have not yet led to clinically validated tests?
- Which types of model building and validation methods have been used to develop clinically validated biomarker signatures, and what are the lessons learned and recommended workflows?
- Which recommendations and guidelines have been proposed to address common challenges in biomarker development using omics data?

These questions lend themselves to a scoping review, because omics-derived biomarker development is still an evolving field, and a preliminary assessment of the potential scope and size of the available biomedical literature on these topics is required as a first step for further follow-up research. Therefore, the objective of this study was to address the above questions by retrieving and examining the current literature on biomarker discovery and validation studies using omics data and ML approaches. While the focus on articles describing discovery and validation approaches covers relevant aspects for clinical translation, we point out that other translational and regulatory aspects, such as the assessment of the clinical efficacy of biomarker-associated treatment decisions, the assessment of cost-effectiveness and research ethics, are not addressed in the present review, but have been discussed in previous dedicated articles (20–24). Our scoping review also does not aim at providing a quantitate benchmark evaluation of different ML approaches, but relevant studies have previously been presented for supervised machine learning (25), and unsupervised clustering (26) and survival prediction (27) on multiple omics data types.

Methods

We conducted a scoping review following the methodological framework suggested by the Joanna Briggs Institute (28). This framework consists of six stages: 1) identifying the research questions, 2) identifying relevant studies, 3) study selection, 4) charting the data, 5) collating, summarising and reporting results, and 6) consultation.

The scoping review approach was considered most suitable to respond to the broad scope and the evolving nature of the field. Compared to systematic reviews that aim to answer specific questions, scoping reviews present a general overview of the evidence pertaining to a topic and are useful to examine emerging trends, to clarify key concepts and identify gaps (29,30). Before conducting the review, a study protocol was published on the online platform Zenodo (19). Due to the iterative nature of scoping reviews, deviations from the protocol are expected and duly reported when occurred. We used the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist to report our results (31) (Online supplementary file 1).

Study identification

Relevant studies and documents were identified, balancing feasibility with breadth and comprehensiveness of searches. We searched PubMed, EMBASE and Web of Science (last search date: July 27, 2021) for articles describing supervised or unsupervised ML analyses for biomarker discovery or personalised medicine, including both discovery and validation methods. The relevance of the search methodology was ensured by using a strict multi-stage filtering, considering only articles including at least one relevant search term per category from four categories of keywords (“Personalized medicine / Biomarkers”, “Omics”, “Machine Learning” and “Validation”, covering both synonyms for these terms and closely related keywords, see Fig. 1, illustrating the keyword-based search strategy, and Online Supplementary file 2 for the detailed search queries), and subsequently post-filtering the retrieved articles manually to exclude studies not involving omics-based biomarker research or lacking a description of machine learning and validation analyses (see sections on *Eligibility criteria* and *Study selection*). To cover only relevant scientific content, the scope was limited to journal publications and meeting abstracts from international conferences and workshops, and no other grey literature was included. We restricted inclusion to reports published from January 2000 to July 2021 (covering also “online first” articles with official publication date in the future) in English, French, Spanish, Italian and German language. Since to the best of our knowledge, the first clinically validated FDA-cleared omics-derived biomarker signature was published in 2002 (32), only few preliminary discovery studies were expected to have taken place significantly earlier than 2002, and we therefore did not extent the search further backwards in time than January 2000.

Eligibility criteria

We included peer-reviewed methodology articles, review articles, opinion articles on supervised and unsupervised ML methods for omics disease prediction and stratification and associated statistical cross-validation and multi-cohort validation methods (addressing accuracy, robustness, and clinical relevance). Only approaches tested on real-world biomedical omics data were reviewed, while studies relying purely on simulated data or were excluded. We also excluded papers on biomarker methods without a demonstrated biomedical application, and those with insufficient sample size (i.e., removing studies covering less than 50 samples per group for the main conditions studied, unless a dedicated power calculation was presented) or statistical validation (i.e., lack of clear descriptions of cross-validation or external testing methodology, performance metrics and test statistics). These exclusion criteria were not specified in the generic review protocol, but they were agreed among the authors prior to the screening process.

To cover both data from original research papers and prior systematic reviews, we extracted information from three main article types: (1) applied research papers, (2) methodology articles with demonstrated applications, and (3) review articles on methods, applications and validation approaches.

Apart from these inclusion and exclusion criteria, for the final result presentation, the statistical investigations covered all selected articles, whereas the detailed discussion of study characteristics focused on the studies that led to clinically validated biomarker signatures tested on multiple cohorts with large sample sizes (i.e., studies using a power calculation to demonstrate the adequacy of the chosen sample sizes, or covering hundreds or thousands samples per studied subject group).

Study selection

We exported the references retrieved from the searches into the online tool Rayyan (33). Duplicates were removed automatically using the reference manager Endnote X9 (Clarivate Analytics, Philadelphia, United States) and manually by the reviewers. One reviewer loaded the retrieved records into the online screening tool Rayyan (33), and two reviewers confirmed the eligibility independently by covering both the screening for all records and the full-text review for the articles pre-selected by the screening. Disagreements were solved by consensus.

Charting the data and synthesis of results

We designed a data extraction form using Excel (Online supplementary file 3). General study characteristics extracted covered author names, title, citation, type of publication (e.g., journal article, meeting abstract), study population and sample size (if applicable), methodology/study design, and outcome measures (if applicable). Specific items associated with the topic of the scoping review included the study type (e.g., Case-control study, differential diagnosis study, prognostic study, review – methods, review – applications, review – validation); the article type (journal or conference article), the generic ML domain (e.g., supervised/unsupervised); and the name of specific approaches for outcome prediction and for validation. Moreover, to capture key findings related to the review questions, relevant sentences were extracted from each reviewed article, and if needed, complemented by a brief explanatory remark, and by writing out abbreviations used in the original text.

The reviewers piloted the data extraction form using five records from the retrieved article collection. Two reviewers (EG, AR) working independently extracted the data from the included articles. In the case of disagreements, consensus was obtained by discussion.

In the final full-text review stage, the pre-selected articles were grouped by topic, categorizing articles into applied vs. methodological studies, supervised vs. unsupervised analyses, and assigning algorithm type identifiers to each article (review articles and papers on validation methodologies were considered as separate categories without a specific algorithm type assignment). The full-text review and categorization of articles into different publication types was done through independent manual inspection by the two reviewers.

While the information on sample sizes and validation methods was documented as part of the data extraction (Online supplementary file 4, a spreadsheet version has been made available on the online platform Zenodo (34)), it was not within the remit of this scoping review to assess the methodological quality of individual studies included in the analysis.

Consultation exercise

The members of the PERMIT consortium, associated partners, and the PERMIT project Scientific Advisory Board discussed the preliminary findings of the scoping review in a 2-hour online workshop.

Patient and public involvement

The European Patients' Forum is a member of PERMIT project. Although not directly involved in the conduction of the scoping review, they received the draft review protocol for collecting comments and feedback.

Results

Study selection and general characteristics of reports

We retrieved 1563 abstracts from the literature search. After the removal of duplicates, we screened the remaining 1475 abstracts for eligibility. 619 records were excluded, while 856 abstracts were retained for the full-text assessment. Finally, we included 352 articles that passed all filtering criteria in the data extraction and analysis (see flow chart in Fig. 2, and Online supplementary file 4, providing the reference for each selected article, as well as information on the study type and methodology, the outcome measures, the validation type, and representative sentences from each article on the main study results and key findings; a spreadsheet version of this table has been available on the online platform Zenodo (34)).

The full-text article review revealed that many studies did not meet the pre-defined inclusion criteria: 371 articles (43%) were removed because of an insufficient sample size, and 105 further articles (12%) were excluded because they provided insufficient details on the validation results or methodology (see Fig. 2). This shows that the challenges of recruiting an adequate number of participants per study group or conducting sufficient omics profiling experiments for robust model building and validation are not met in a large proportion of omics biomarker studies. Moreover, many studies lack adequate documentation for the study design and validation.

For the selected articles that cover primary research on omics biomarker studies, the majority (78%) rely entirely on an internal validation involving data from only a single cohort, whereas studies that

use an external validation on an independent cohort are still underrepresented (only 12% of articles describe both an internal cross-validation and an external cohort validation, and an additional 10% include an external validation, but do not report internal cross-validation results). However, when comparing the numbers of published studies over different periods of time during the past 20 years, the relative proportion of studies including an external validation has increased in recent years (see Fig. 3), suggesting a growing recognition of the importance of independent, multi-cohort validation.

Next, we investigated the countries of origin for the selected articles, showing that the United States of America (USA) are contributing the largest proportion of validated biomarker studies (28%), followed by China (18%), Canada (5%), Germany (4%), and the United Kingdom and India (both 3%; see also Fig. 4, providing a map visualization of the country statistics). These country representations show limited correlation with population sizes and may largely reflect worldwide variation in relative biomedical research productivity reviewed in previous study (35). Since the most prolific countries in the development of molecular diagnostics have already set up policies and regulations for omics- and ML-based *in vitro* diagnostics and medical devices (e.g., see the life cycle regulation of AI- and ML-based software devices in the USA (36)), they may also provide a role model for countries still in the process of establishing similar regulatory frameworks.

When inspecting the representation of study design types in the filtered article collection, the great majority of documents described diagnostic studies (67%), prognostic and survival prediction studies were covered in 8% of articles, and studies examining therapy or drug response in 7% (see Fig. 5). Apart from this, 13% of articles were reviews on methodologies and applications in the field, and 5% of articles described other rare study types (e.g. tissue-of-origin prediction studies or combinations of different study types).

Since a detailed discussion of all filtered articles is not within the scope of the present review, in the following, we focus on reviewing representative omics biomarker studies which achieved the highest validation level, i.e., clinical approval of the developed molecular signature as an LDT or FDA approved test (see the overview of studies in Table 1 and the FDA web-site (37)). We investigate the shared features of these successful studies, examine how they address common shortcomings and missing features of other reviewed studies, and summarize the lessons learned.

Success stories in omics-based biomarker signature development

Cancer approved omics-derived diagnostic tests (9 studies)

The first and most well-known omics-derived molecular test to receive FDA clearance was *MammaPrint*®, a prognostic signature using the RNA expression activity of 70 genes to estimate the risk for distant tumours metastasis and recurrence in early-stage breast cancer patients (6,32,38–41). This test was developed at the Netherlands Cancer Institute, using DNA microarray analysis to investigate primary breast tumours of 117 patients. Supervised ML was applied to the resulting data to identify a highly predictive gene signature for a short interval to distant metastases in lymph node negative patients (32).

A distinctive feature of the development approach behind this signature in comparison to other reviewed studies was the multi-stage filtering and cross-validation strategy used in the initial discovery study, which may explain the repeated confirmation of the signature in later validation studies (6,38–41). From 25k genes represented on the DNA microarrays, only those significantly regulated in more than 3 tumours out 78 sporadic lymph-node negative patients were preselected, and further filtered by retaining only the genes with a minimum absolute correlation with the disease

outcome of 0.3. The resulting list of 231 genes, rank-ordered by absolute correlation, was investigated by sequentially adding the next top 5 genes from the list to a candidate ML classifier and evaluating its performance by leave-one-out cross-validation (LOOCV). This procedure was repeated as long as the estimated accuracy of the classifier improved, providing a final candidate signature of 70 genes. The final signature was validated on multiple independent test sets, including a set of 19 external samples in the original study and several additional validations on independent cohorts in follow-up studies (6,38–41).

The *MammaPrint* signature provided the role model for the subsequent development of a similar prognostic test for colon cancer, *ColoPrint*® (42–47). This test aims at detecting the approx. 20% of patients with stage II colon cancer expected to experience a relapse and develop distant metastases. It uses an 18-gene expression signature, developed by analysing DNA microarray data in a similar manner to the *MammaPrint* approach. The diagnostic approach has been commercialized as an LDT to assist physicians in selecting treatment options for colon cancer patients. Similar to *MammaPrint*, the signature development was characterized by extensive discovery and validation studies, which involved multiple statistical reproducibility, stability and precision analyses for independent, large-scale patient cohorts (48).

Another widely used cancer-related LDT, which received FDA clearance in 2013, is the *Prosigna*® *Breast Cancer Prognostic Gene Signature Assay*, previously called *PAM50* test (49–53). This assay assesses mRNA expression for a signature of 58 genes (50 target genes + 8 endogenous control genes) to predict the risk of distant recurrence for hormone-receptor-positive breast cancer between 5 to 10 years after diagnosis (prerequisites are that the patients have been treated with hormonal therapy and surgery, and are stage I or stage II lymph-node negative, or in stage II with one to three positive nodes). The test development started with a microarray discovery study and involved a multistage filtering, using consecutive applications of statistical tests and cross-validation to propose a subset of candidate gene markers (54). The authors compared the reproducibility of classification scores obtained with these markers for three centroid-based prediction methods to ensure the robustness of the methodology. By further developing the approach into a more sensitive PCR-based test, and later into an assay using the NanoString nCounter Dx Analysis System, the predictive performance was improved in a step-wise fashion. The original discovery study was characterized by significantly larger sample sizes than the majority of reviewed biomarker studies, with a training set of 189 samples, test sets of 761 patients evaluated for prognosis, and 133 patients evaluated for prediction of pathologic complete response to treatment with taxane and anthracycline. These study design features in combination with multi-stage filtering and validation approaches, and improved measurement technology during the course of the study, may explain the successful progression of the PAM50 test to FDA clearance. The test has only three genes in common with the MammaPrint approach (*KNTC2*, *MELK*, *ORC6L*), which may be explained by the different technical and analytical approaches used, but a previous comparative evaluation concluded that the tests provide broadly equivalent risk information for females with oestrogen receptor (ER)–positive breast cancers (55).

Among the LDTs for breast cancer prognosis, *Oncotype DX*® is a further test commonly used in clinical practice (8,56–59). The underlying gene signature consists of 16 cancer-associated genes and 5 reference genes, and is therefore often also referred to as ‘21-gene assay’. Its main application is to predict risk of recurrence in oestrogen-receptor positive tumours. The relevance of this prognostic tool for treatment selection may be explained by the strong association of the provided recurrence score with the probability of positive treatment response to chemotherapy (60). *Oncotype DX* was developed using a consecutive refinement procedure, starting with the RT-PCR assessment of 250 candidate genes across 447 patients from three distinct studies to identify the 21-gene signature after multiple filtering steps. A recurrence score algorithm built using the signature as input was clinically validated on 668 independent patients (61). The selection of the 16 cancer-related genes included in the assay involved scoring the performance of the candidate features in all three studies and the consistency of the primer/probe performance in the assay (62).Thus, particular

strengths of the development process for this LDT include the consideration of both technical robustness and statistical robustness of the assay across distinct cohorts. The *Oncotype DX* signature shares one gene with *MammaPrint* (*SCUBE2*), and 9 genes with the Prosigna / PAM50 test (*BIRC5, CCNB1, MYBL2, MMP11, GRB7, ESR1, PGR, BCL, BAG1*). However, an independent clinical validation of Oncotype DX and the PAM50 signature for estimating the likelihood of distant recurrence in ER-positive, node-negative, post-menopausal breast cancer patients treated with endocrine therapy suggested that the PAM50 signature provided more prognostic information than Oncotype DX (63).

While the first validated omics biomarker signatures were developed for breast cancer, similar diagnostic and prognostic tools have followed for other cancer types. One of these is the Decipher® Prostate Cancer Test (9,64–68), which differs from other omics-derived diagnostic tools by being provided together with a software platform and database, the Decipher Genomic Resource Information Database (GRID), that captures 1.4 million expression markers per patient to facilitate personalised care. The test itself uses 22 preselected RNAs to predict clinical metastasis and cancer-specific mortality for patients who have undergone radical prostatectomy. An initial discovery study by the Mayo Clinic (Rochester, MN, USA) investigated a cohort of 545 such patients, split into a training (n = 359) and a validation cohort (n = 186). Similar to other LDTs, the discovery started with a genome-wide profiling and used both statistical and ML analyses for filtering. First, *t*-tests were applied (reduction from 1.4 mil. to 18,902 differentially expressed RNAs), then regularized logistic regression (reduction to 43 candidate markers), and finally a random forest-based feature selection (reduction to final set of 22 RNAs). Apart from testing the signature in the validation cohort, further external validations were performed in subsequent studies (9,64–68). Overall, distinctive strengths of the used approach include the improved interpretability of the test results through supporting analyses on the GRID platform, and the robustness of the discovery and validation approach, involving large sample sizes and several complementary statistical and ML assessments.

While most diagnostic tests in oncology have been designed for specific cancer types, a dedicated LDT has also been developed for cancers of unknown or uncertain diagnosis. The Cancer Type ID® test by bioTheranostics distinguishes between 50 different tumour types using a 92-gene RT-PCR expression measurement signature (15,69–71). This signature was derived from analyses of a microarray data collection covering 446 frozen tumour samples and 112 formalin-fixed, paraffin-embedded (FFPE) samples of both primary and metastatic tumours. Modelling steps involved k-nearest neighbour clustering and classification, and a genetic algorithm to explore the search space of possible feature subset selections. After successful cross-validation (84% accuracy) and external validation (82% accuracy on 112 independent FFPE samples), the microarray-based signature was further developed to use more sensitive RT-PCR measurements. Testing the new approach on an independent validation set provided an increased accuracy (87%). Distinctive characteristics of the development process that may have contributed to the positive validation include the efficient and extensive exploration of the search space of possible gene subset selections via a genetic algorithm, the large sample sizes used for discovery and validation, and the transfer of the assay from microarrays to the more sensitive RT-PCR platform.

The first omics-derived biomarker signatures addressed only the most frequent cancer types, but more recent applications in oncology focus on the diagnosis of less common malignancies, such as thyroid cancer. Typically, deciding whether a thyroid nodule is benign or cancerous is possible via a fine needle aspiration (FNA) biopsy, without requiring more complex measurements or analyses. However, while direct FNA-based diagnosis is feasible in most cases, indeterminate results can occur (72). To help prevent unnecessary surgeries for the corresponding patients, a molecular signature and LDT known as the Afirma™ Gene Expression Classifier (GEC) has been developed to discriminate benign from cancerous thyroid nodules (72–77). The original discovery study behind the GEC signature used mRNA expression analysis in 315 thyroid nodules, covering 178 retrospective surgical tissues and 137 prospectively collected FNA samples. Two ML classifiers were

trained separately on surgical tissues and FNAs, assessing the test set performance on 48 independent, prospective FNA samples (50% of which had indeterminate cytopathology). Discriminative features were selected using a linear modelling approach implemented in the software Limma, and a linear support vector machine was applied for model building and performance estimation via 30-fold cross-validation (CV). The successful cross-validation results were confirmed on multiple distinct cohorts (72,75–78). While the internal validation used in the initial study cannot address cohort-specific biases, the combined use of established feature selection and modelling approaches, and the subsequent external validation across multiple cohorts with large sample sizes may account for the successful translation of this signature.

Most omics-based diagnostic tests identified in our study rely purely on gene expression profiling data. However, more recently, first multi-omics signatures for diagnostic purposes have been developed. One of the first LDTs that integrated information from both RNA and DNA sequencing was the FoundationOne® Heme assay (14,79–81). This assay aims to detect hematologic malignancies, sarcomas, pediatric malignancies, or solid tumours (including among others leukaemias, myelodysplastic syndromes, myeloproliferative neoplasms, lymphomas, multiple myeloma, Ewing sarcoma, Leiomyosarcoma, and paediatric tumours). The test identifies four types of genomic alterations (base substitutions, insertions and deletions, copy number alterations, rearrangements) and reports microsatellite instability and tumour mutational burden to facilitate clinical decision making. This approach was originally developed and evaluated using reference samples of pooled cell lines in order to model the main characteristics that determine the test accuracy, including mutant allele frequency, indel length and amplitude of copy change (79). A first validation using 249 independent FFPE cancer samples, which had already been characterized by established assays, confirmed the accuracy of the test. External validation studies on independent cohorts corroborated the utility of the test for further diagnostic applications (14,82). The study results highlight the potential of integrating diverse biological data sources in order to obtain more robust and reliable predictions, a strategy that may be promising in particular for complex disorders that involve very heterogeneous phenotypes.

A common limitation of genomic profiling approaches for diagnostic testing is that most analyses have to be performed in centralized specialty laboratories, which limits a wider use and results in long waiting times. To address this shortcoming, the Elio™ Tissue Complete assay, an *in vitro* diagnostic test cleared in 2020 by the FDA for assessing somatic mutations and tumour mutation burden (TMB) in solid tumours, has been developed as an integrated DNA-to-report approach to enable a decentralized evaluation in all diagnostic labs with next generation sequencing (NGS) technology (83). The analytical performance of the test was assessed by comparing it with the FoundationOne test (see above) using a concordance analysis on 147 tumour specimens. It provided a positive percent agreement (PPA) above 95% for single nucleotide variants (SNVs) and insertions/deletions, and 80-83% PPA for copy number alterations and gene translocations (83). The test has recently also been applied to investigate the response to immune checkpoint inhibitors (ICI) in metastatic renal cell carcinoma (mRCC), using a retrospective evaluation of SNVs, TMB, microsatellite status and genomic status of antigen presentation genes (84). While no correlation between treatment response and TMB was observed, one third of patients with progressive disease following ICI therapy displayed loss of heterozygosity of major histocompatibility complex class I genes (LOH-MHC) vs. 6% of disease control patients, suggesting that loss of antigen presentation may restrict ICI response (84). In summary, the Elio Tissue Complete assay provides an example of how integrating NGS analyses with bioinformatics in a combined DNA-to-report approach could help to broaden the access to genomic diagnostics for both clinical and research applications.

Non-cancer approved omics-derived diagnostic tests (4 Studies)

While most clinically approved omics-derived diagnostic tests have been developed in the field of oncology, one of the first LDTs that received FDA clearance for a non-cancer disease was the AlloMap® Heart test (13,85–87). It uses a gene expression signature of 11 target genes and 9 control genes in peripheral blood from heart transplant recipients to estimate the risk for acute cellular cardiac allograft rejection. The development process involved statistical analyses of leukocyte microarray profiling data from 285 samples, and subsequent RT-PCR validation and bioinformatics post-processing (13). Prior knowledge from database and literature mining was included in the analysis by mapping the data to known alloimmune pathways. This allowed the researchers to narrow down 252 candidate marker genes. An RT-PCR validation on 145 samples confirmed 68 of these candidate genes, which distinguished rejection samples from quiescent samples according to a T-test ($p < 0.01$). Six genes were eliminated due to significant variation in gene expression with sample processing time. Next, the investigators averaged correlated gene expression levels to create robust meta-level features, called 'metagenes', and added 20 of these features as new variables. A linear discriminant analysis was applied, providing a prediction model using four individual genes and three metagenes, which aggregate information from 11 original genes. Finally, bootstrap validation procedures and external test set validations were performed to confirm the accuracy of this signature. Overall, distinctive aspects of the development approach for the AlloMap signature include the knowledge-based gene discovery, a comprehensive RT-PCR validation of candidate genes, and the robust bootstrap and external validation analyses.

The first clinically validated LDT for a cardiovascular indication derived from omics data was the Corus® CAD test, developed to identify coronary artery disease (CAD) in stable non-diabetic patients (11,88–91). In contrast to most other omics-based tests, Corus CAD is not a pure molecular signature test, but takes the clinical covariates gender and age into account. The initial discovery study used a retrospective microarray analysis of blood samples from 195 diabetic and non-diabetic patients from the Duke University CATHGEN registry. After ranking the studied genes by the statistical significance of group differences and prior biological knowledge on their disease relevance, 88 genes were selected for RT-PCR validation. Because diabetes status as a clinical covariate was significantly associated with the observed gene expression alterations, and the identified CAD-associated genes did not overlap between diabetic and non-diabetic patients, the authors decided to limit follow-up work to non-diabetic patients. In a prospective clinical trial, microarray profiling was conducted on blood samples from 198 patients, and top-ranked genes were further validated using RT-PCR for 640 blood samples. After multiple filtering steps, taking into account statistical significance in T-tests, biological relevance, gene correlation clustering and cell-type analyses, a final signature of 23 genes was derived, composed of 20 CAD-associated genes and 3 reference genes (92). To maximize the predictive performance, the final prediction algorithm was optimized to adjust for differences associated with age and gender. Compared to most other reviewed studies, the Corus CAD approach stands out by taking clinical covariates into account in the final prediction model, including an intermediate critical review and adjustment of the inclusion criteria (limiting the focus to nondiabetic patients), and integrating complementary filtering and validation analyses on large sample sizes.

For inflammatory diseases, a first omics-derived signature recently received approval for measuring rheumatoid arthritis (RA) inflammatory disease activity, the Vectra® DA multi-biomarker test (93–97). It uses blood serum samples and multi-spot 96-well immunoassay plates to assess serum concentrations of 12 protein biomarkers associated with the pathobiology of RA. The original Vectra DA score, which combines these measurements into a composite score between 1 and 100, was assessed via multivariate regression and displayed a high predictive power in estimating a standard RA score, the Disease Activity Score in 28 joints using the C-reactive protein level (DAS28-CRP), in both seropositive (AUC 0.77, $P < 0.001$) and seronegative (AUC 0.70, $P < 0.001$) patients (97). This score was later adjusted for age, gender and adiposity (based on leptin concentration), and validated in two cohorts against DAS28-CRP as a prognostic test for radiographic progression during the next year. The results showed that the new adjusted score was the most accurate, independent predicator

of progression, with the rate of progression increasing from < 2% in the low (1-29) adjusted score category to 16% in the high (45-100) category (95). Overall, the Vectra DA approach illustrates the utility of omics-based biomarker signatures for prognostic applications in inflammatory disorders, and further highlights the benefit of integrating omics signatures with information from clinical covariates.

For neurodegenerative disorders, clinically approved diagnostic and prognostic omics-derived tests are still lacking. However, recently the Helix® Genetic Health Risk App for Late-onset Alzheimer's Disease (AD) was cleared by the FDA for over-the-counter use. It detects clinically relevant variants in genomic DNA isolated from human saliva of individuals ≥18 years in order to report and interpret genetic health risks, and evaluates the information of variants with established genome-wide significant associations to AD. When tested on 99 human saliva samples, the accuracy was 100% with a lower 95% CI bound of 96.3% (98). The approach uses a whole exome sequencing (WES) constituent device, the Helix® Laboratory Platform (99–101), as a qualitative *in vitro* diagnostics approach covering measurements for approximately 20k genes. The Helix Laboratory Platform has received FDA clearance through a new regulatory approval pathway established by the FDA for WES devices (Regulation 21 CFR 866.6000). Due to the generic applicability of the WES profiling assay used by this platform, called Exome+, the assay has also been applied to find statistically significant gene-based associations for several other phenotypes in large-scale cohort studies (99) and to identify carriers of autosomal dominant diseases by population-based genetic screening (101). Thus, the Helix Laboratory Platform provides a first example for a new approval pathway for omics-based diagnostic tests, in which a clinically approved genomic testing device is not anymore linked to a single diagnostic application or a specific disease type. Instead, the market authorization for diagnostic tests is obtained separately from the device and facilitated and accelerated by the prior approval of the constituent measurement device. For the future development of omics-derived biomarker signatures, this may allow researcher to focus on demonstrating the clinical utility of a new signature, while the analytical validity of the underlying testing device has already been established previously.

Discussion

Statement of principal findings

The scoping review of articles on patient stratification using omics data revealed common limitations in the study design for many published biomarker development projects, such as insufficient and imbalanced sample sizes per study group and inadequate validation methods, but also identified multiple studies that have led to validated diagnostic and prognostic tests. These success stories were investigated in more detail to identify common characteristics in the study design, discovery and validation methods, which may have supported the clinical translation of the initial findings. Fig. 6 outlines key shared aspects that are possible determinants of the study success and could help to guide future biomarker investigations. In particular, they cover the following main features:

(1) A sample size selection, study group and replicate design that provides adequate statistical power for the ML analyses;

2) The application of robust statistical filtering and evaluation schemes (including multiple layers of statistical and ML-based feature selection, combined statistical and biological filters, robust validation schemes that involve multiple cross-validation, bootstrapping and external validation analyses, using multiple suitable and complementary performance metrics, and providing information on the statistical variation and confidence intervals for the performance estimates, see Fig. 7 for an overview of recommended generic steps for robust model building and evaluation);

3) Clarity of the study scope and goals (involving clear inclusion and exclusion criteria, primary and secondary outcomes, and decision processes to make necessary adjustments due to new knowledge gained during the project, such as the adjusted inclusion criteria in the Corus CAD study and the progression from non-targeted microarray technology to higher-sensitivity RT-PCR in the case of the Prosigna test and the Cancer Type ID test);

4) Completeness and reproducibility of the study documentation (covering details on used instruments, parameters and settings, reproducible methods descriptions, and information on data provenance);

5) Interpretability and biological plausibility of the created predictive models (including explainable and justifiable predictions, human-interpretable model descriptions, and biologically plausible models that agree with the current mechanistic understanding of the studied disorder);

6) Integration of prior biological knowledge into the predictive feature selection, model building and validation procedures (e.g., using public data on disease-associated molecular pathways and networks; complementary clinical and real-world data, and relevant multi-omics data).

Strengths and Limitations

The majority of methodological recommendations derived from the study relate to the early planning and study design for biomarker discovery projects, involving considerations associated with the choice of the study group, sampling and blocking design, the measurement technology, and the input and output variables (16,17). These recommendations are therefore mainly applicable to prospective studies. For retrospective biomarker investigations of already collected data, the suggestions derived from the review are limited to guidance on improving analysis workflows, e.g. for filtering and evaluation analyses, the integration of prior knowledge from multi-omics data and public annotation databases, and the choice of robust and interpretable modelling approaches for the generation of biologically plausible and reproducible prediction models. While the focus of the review on studies that have already led to validated biomarker models and that fulfil minimum requirements for sample size and statistical model assessment helps to ensure the quality of the selected articles, no further quality evaluation was performed. The reader should also note the generic limitations of machine learning methods which can affect all biomarker studies: These include the necessity for a representative coverage of the relevant outcomes in the training and validation groups, a sufficiently comprehensive and sensitive coverage of informative predictor variables in the data for the outcomes of interest, which may not be achievable for omics data from tissues and body fluids with limited disease relevance or measurement sensitivity, and a sufficient data quality in terms of the influence of systematic biases and noise. Moreover, for multi-omics biomarker analyses, in addition to adequate pre-processing and machine learning approaches, suitable strategies and methods for the integration of diverse omics data are also needed. These multi-omics data integration strategies were not within the scope of the present review, but have been reviewed in previous publications (102–104). Finally, more recent methodological developments in the machine learning and cross-validation analysis of omics data, such as meta-learning (105) and bolstered cross-validation (106), have only limited coverage among the articles that passed the eligibility criteria, and will therefore require further dedicated study in the future.

Discussing important differences in results

Previous reviews of ML approaches using omics data for patient stratification have focused on domain-specific analyses for specific types of diseases, or specific types of ML methodologies (107–115). By contrast, this scoping review focuses on disease-agnostic workflows with generic applicability across complex human disorders involving multifactorial molecular alterations. The

coverage of statistical and ML approaches for stratification does not aim to provide a detailed discussion of specific algorithms, statistical methods or scoring metrics, but rather at identifying key determinants of success for generic analysis and validation workflows in biomedical stratification studies. Therefore, the results describe general workflow characteristics that distinguish omics biomarker studies with clinical translation from other studies, and cover associated disease-agnostic recommendations for future studies, whereas method recommendations specific to particular disease types or ML analysis types are covered elsewhere in domain-specific reviews (107–115).

Meaning of the study: implications for clinicians and policymakers

The previous clinical translation successes in omics-based biomarker development reviewed in this study, which have mostly been achieved in the field of oncology, highlight the potential for developing similar biomarker signatures for further disease indications. In contrast to conventional statistical biomarker discovery approaches, which focus on identifying single-molecule markers, systems-level analysis of omics data using multivariate ML approaches can identify multifactorial signatures which are robust against noise in individual gene or protein measurements, and more biologically insightful by reflecting disease-associated cellular process alterations in a more comprehensive fashion.

This scoping review has identified common characteristics of omics studies which have led to clinically validated diagnostic and prognostic tests. Thus, the conclusions drawn on recommended practices for sample size selection, biological data filtering and ML, and the implementation of adequate validation schemes may help to guide clinical researchers on study design choices and the selection of analysis methodologies. Additionally, the scoping review results can help to raise awareness of common pitfalls, such as issues associated with batch effects, biases, confounding factors, lack of statistical power, and multiple hypothesis testing, and thus contribute to preventing these failure causes in biomarker development. For policymakers and funding bodies, findings on the distinctive characteristics of studies with successful clinical biomarker translation, e.g. concerning the specific requirements for robust cross-validation and external result validation methods, may provide relevant information for the design of public and private funding schemes for biomedical research. Risks in funded research projects may be addressed upfront through appropriate guidelines and regulations for the study design and validation (e.g. recommendations on power calculations and specific validation and documentation requirements). Finally, the scoping review results can guide clinicians involved in biomarker discovery on how to make better use of available public knowledge and data sources, e.g. cellular pathway and molecular interaction databases, that may allow them to exploit prior knowledge effectively, and create more robust and interpretable biomarker models.

Unanswered questions & future research

Since the recommendations and guidelines identified from the reviewed articles are mostly derived from established biomarker discovery and validation approaches, new methodologies and upcoming trends could only be covered to a limited extent and may lead to changed recommendations in the future. In particular, in the reviewed patient stratification studies, some of more recently introduced ML concepts (e.g. transfer learning, distance metric learning, semi-supervised learning, structured machine learning, meta learning, multi-view learning, and generative models), data processing techniques (e.g. new dimension reduction approaches, outlier removal methods, data augmentation techniques), and model validation methods (e.g. bootstrapping or bolstered cross-validation, uncertainty quantification), are still underrepresented among the eligible studies reviewed, and may provide suitable topics for follow-up research.

Overall, while the currently available literature on validated stratification biomarkers already provides ample information on common pitfalls and established practices, the development of widely accepted standard guidelines on methodologies for omics biomarker discovery will require further knowledge exchange and deliberation among stakeholders in the field. In particular, integration of domain-specific expertise in discussions involving clinicians, experimental and data scientists, and regulatory and legal experts is required as a follow-up effort to derive comprehensive methodological guidelines for future biomarker development.

Acknowledgments

The authors thank Vanna Pistotti for her assistance with search strategy development and conduction.

List of Figures

Figure 1: Keyword based search strategy for the scoping review

Figure 2: Study selection flow diagram

Figure 3: Validation methods used in omics biomarker studies

Figure 4: Map representation of country statistics for the selected articles

Figure 5: Representation of study types among the selected articles

Figure 6: Characteristics of successful omics-based studies

Figure 7: Recommended generic workflow for biomarker development using machine learning analysis of omics data

Definitions (In boxes)

Box 1: *What is Personalised Medicine?*

According to the European Council Conclusion on personalised medicine for patients, personalised medicine is 'a medical model using characterisation of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention (116).

In the context of the Permit project, we applied the following common operational definition of personalised medicine research: a set of comprehensive methods (methodology, statistics, validation, technology) to be applied in the different phases of the development of a personalised approach to treatment, diagnosis, prognosis, or risk prediction. Ideally, robust and reproducible methods should cover all the steps between the generation of the hypothesis (e.g., a given stratum of patients could better respond to a treatment), its validation and pre-clinical development, and up to the definition of its value in a clinical setting (19).

References

1. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49(107739).
2. Goecks J, Jalili V, Heiser LM, Gray JW. How Machine Learning Will Transform Biomedicine. Vol. 181, Cell. 2020. p. 92–101.
3. Jiang Y, Wang M. Personalized medicine in oncology: Tailoring the right drug to the right patient. Biomarkers in Medicine. 2010.
4. Hopp WJ, Li J, Wang G. Big Data and the Precision Medicine Revolution. Prod Oper Manag. 2018;27(9):1647–64.
5. Glaab E. Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification. Brief Bioinform. 2016;
6. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016;375(8):717–29.
7. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;
8. Knezevic D, Goddard AD, Natraj N, Cherbavaz DB, Clark-Langone KM, Snable J, et al. Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genomics. 2013;14(1).
9. Torres A, Alshalalfa M, Tomlins SA, Erho N, Gibb EA, Chelliserry J, et al. Comprehensive Determination of Prostate Tumor ETS Gene Status in Clinical Samples Using the CLIA Decipher Assay. J Mol Diagnostics. 2017;19(3):475–84.
10. Angell TE, Babiarz J, Barth N, Blevins T, Duh Q, Ghossein RA, et al. Clinical validation of the AFIRMA genomic sequencing braf V600E classifier. Thyroid [Internet]. 2017;27:A50. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L624116485
11. Ladapo JA, Budoff MJ, Sharp D, Kuo JZ, Huang L, Maniet B, et al. Utility of a Precision Medicine Test in Elderly Adults with Symptoms Suggestive of Coronary Artery Disease. J Am Geriatr Soc. 2018;66(2):309–15.
12. Tabari E, Lovejoy AF, Lin H, Bolen CR, Saelee SL, Lefkowitz JP, et al. Molecular characteristics and disease burden metrics determined by next-generation sequencing on circulating tumor DNA correlate with progression free survival in previously untreated diffuse large B-cell lymphoma. Blood [Internet]. 2019;134. Available from: http://dx.doi.org/10.1182/blood-2019-123633
13. Deng MC. The AlloMap™ genomic biomarker story: 10 years after. Clin Transplant. 2017;31(3).
14. He J, Abdel-Wahab O, Nahas MK, Wang K, Rampal RK, Intlekofer AM, et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood. 2016;127(24):3004–14.
15. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med. 2006;130(4):465–73.
16. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013.

17. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73.

18. Banzi R, Gerardi C, Fratelli M, Garcia P, Torres T, Abad JMH, et al. Web-page for the Personalized Medicine Trials (PERMIT) project [Internet]. 2020 [cited 2021 Aug 2]. Available from: https://permit-eu.org

19. Banzi R, Gerardi C, Fratelli M, Garcia P, Torres T, Abad JMH, et al. Methodological approaches for personalised medicine: protocol for a series of scoping reviews [Internet]. 10.5281/zenodo.3770937. Available from: https://zenodo.org/record/3770937

20. Perlis RH. Translating biomarkers to clinical practice. Vol. 16, Molecular Psychiatry. 2011. p. 1076–87.

21. Graaf G, Postmus D, Westerink J, Buskens E. The early economic evaluation of novel biomarkers to accelerate their translation into clinical applications. Cost Eff Resour Alloc. 2018;16(1).

22. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. Vol. 4, EPMA Journal. 2013.

23. Williams JK, Anderson CM. Omics research ethics considerations. Nurs Outlook. 2018;

24. Vähäkangas K. Research ethics in the post-genomic era. Environmental and Molecular Mutagenesis. 2013.

25. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: A large benchmark suite for machine learning evaluation and comparison. BioData Min. 2017;

26. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. Nucleic Acids Res [Internet]. 2018;46(20):10546–62. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L626271229

27. Herrmann M, Probst P, Hornung R, Jurinovic V, Boulesteix AL. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform. 2021;

28. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13(3):141–6.

29. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, Mcewen SA. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Res Synth Methods. 2014;5(4):371–85.

30. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143.

31. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018 Oct;169(7):467–73.

32. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.

33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210.

[dataset] 34. Glaab E, Armin, Rita, Chiara, Paula, Jacques, et al. Data from: Selected articles from the scoping review of biomarker discovery studies for the EU project on "Personalised Medicine Trials" (PERMIT). Zenodo, November 4, 2021.

https://doi.org/10.5281/zenodo.5646467

35. Rahman M, Fukui T. Biomedical research productivity: Factors across the countries. Int J Technol Assess Health Care. 2003;19(1):249–52.

36. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle Regulation of Artificial Intelligence- And Machine Learning-Based Software Devices in Medicine. Vol. 322, JAMA - Journal of the American Medical Association. 2019. p. 2285–6.

37. FDA Center for Devices and Radiological Health. Web-page on Nucleic Acid Based Tests by the Food and Drug Administration (FDA) [Internet]. 2021 [cited 2021 Aug 2]. Available from: https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests

38. Winner BS, Sgroi DC, Ryan PD, Bruinsma TJ, Glas AM, Male A, et al. Analysis of the mamma print breast cancer assay in a predominantly postmenopausal cohort. Clin Cancer Res. 2008;14(10):2988–93.

39. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: Another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn. 2009;9(5):417–22.

40. Sapino A, Roepman P, Linn SC, Snel MHJ, Delahaye LJMJ, Van Den Akker J, et al. MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue. J Mol Diagnostics. 2014;16(2):190–7.

41. Mook S, Knauer M, Bueno-De-Mesquita JM, Retel VP, Wesseling J, Linn SC, et al. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. Ann Surg Oncol. 2010;17(5):1406–13.

42. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, et al. Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. Ann Surg. 2013;257(6):1053–8.

43. Kopetz S, Tabernero J, Rosenberg R, Jiang Z, Moreno V, Bachleitner-Hofmann T, et al. Genomic Classifier ColoPrint Predicts Recurrence in Stage II Colorectal Cancer Patients More Accurately Than Clinical Factors. Oncologist. 2015;20(2):127–33.

44. Tan IB, Tan P. Genetics: An 18-gene signature (ColoPrint®) for colon cancer prognosis. Nat Rev Clin Oncol. 2011;8(3):131–3.

45. Rosenberg R, Maak M, Simon I, Nitsche U, Schuster T, Kuenzli B, et al. Independent validation of a prognostic genomic profile (ColoPrint) for stage II colon cancer (CC) patients. J Clin Oncol. 2011;29(4_suppl):358–358.

46. Salazar R, de Waard JW, Glimelius B, Marshall J, Klaase J, Van Der Hoeven J, et al. The PARSC trial, a prospective study for the assessment of recurrence risk in stage II colon cancer (CC) patients using ColoPrint. J Clin Oncol. 2012;30(4_suppl):678–678.

47. Tabernero J, Moreno V, Rosenberg R, Nitsche U, Bachleitner-Hofmann T, Lanza G, et al. Clinical and technical validation of a genomic classifier (ColoPrint) for predicting outcome of patients with stage II colon cancer. J Clin Oncol. 2012;30(4_suppl):384–384.

48. Bachleitner-Hofmann T, Simon I, Salazar R, Tabernero J, Rosenberg R, van der Akker J, et al. Development and Validation of a Robust Molecular Diagnostic Test (COLOPRINT) for Predicting Outcome in Stage II Colon Cancer Patients. Ann Oncol. 2012;23:ix179.

49. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer. 2014;14(1).

50. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and

verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics. 2015;8(1).

51. Alvarado MD, Prasad C, Rothney M, Cherbavaz DB, Sing AP, Baehner FL, et al. A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer. Adv Ther. 2015;32(12):1237–47.

52. Jensen MB, Lænkholm AV, Nielsen TO, Eriksen JO, Wehn P, Hood T, et al. The Prosigna gene expression assay and responsiveness to adjuvant cyclophosphamide-based chemotherapy in premenopausal high-risk patients with breast cancer. Breast Cancer Res. 2018;20(1).

53. Hequet D, Callens C, Gentien D, Albaud B, Mouret-Reynier MA, Dubot C, et al. Prospective, multicenter French study evaluating the clinical impact of the Breast Cancer Intrinsic Subtype-Prosigna® Test in the management of early-stage breast cancers. PLoS One. 2017;12(10).

54. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7.

55. Bartlett JMS, Bayani J, Marshall A, Dunn JA, Campbell A, Cunningham C, et al. Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test is More Equal Than the Others. J Natl Cancer Inst. 2016;108(9).

56. Kelly CM, Krishnamurthy S, Bianchini G, Litton JK, Gonzalez-Angulo AM, Hortobagyi GN, et al. Utility of oncotype DX risk estimates in clinically intermediate risk hormone receptor-positive, HER2-normal, grade II, lymph node-negative breast cancers. Cancer. 2010;116(22):5161–7.

57. Lo SS, Mumby PB, Norton J, Rychlik K, Smerage J, Kash J, et al. Prospective multicenter study of the impact of the 21-gene recurrence score assay on medical oncologist and patient adjuvant breast cancer treatment selection. J Clin Oncol. 2010;28(10):1671–6.

58. Carlson JJ, Roth JA. The impact of the Oncotype Dx breast cancer assay in clinical practice: A systematic review and meta-analysis. Vol. 141, Breast Cancer Research and Treatment. 2013. p. 13–22.

59. Thakur SS, Li H, Chan AMY, Tudor R, Bigras G, Morris D, et al. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One. 2018/01/06. 2018;13(1):e0188983.

60. Pease AM, Riba LA, Gruner RA, Tung NM, James TA. Oncotype DX ® Recurrence Score as a Predictor of Response to Neoadjuvant Chemotherapy. Ann Surg Oncol. 2019;

61. Gianni L, Zambetti M, Clark K, Baker J, Cronin M, Wu J, et al. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J Clin Oncol. 2005;23(29):7265–77.

62. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817–26.

63. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. J Clin Oncol. 2013;31(22):2783–90.

64. Marrone M, Potosky AL, Penson D, Freedman AN. A 22 gene-expression assay, decipher® (GenomeDx biosciences) to predict five-year risk of metastatic prostate cancer in men treated with radical prostatectomy. PLoS Curr. 2015;7(EVIDENCEONGENOMICTESTS).

65. Nguyen PL, Haddad Z, Lam LLC, Ong K, Buerki C, Deheshi S, et al. Evaluation of the

Decipher prostate cancer classifier to predict metastasis and disease-specific mortality from genomic analysis of diagnostic prostate needle biopsy specimens. J Clin Oncol. 2017;35(6_suppl):4–4.

66. Magi-Galluzzi C, Yousefi K, Haddad Z, Palmer-Aronsten B, Lam LLC, Buerki C, et al. Validation of the Decipher prostate cancer classifier for predicting 10-year postoperative metastasis from analysis of diagnostic needle biopsy specimens. J Clin Oncol. 2016;34(2_suppl):59–59.

67. Dalela D, Löppenberg B, Sood A, Sammon J, Abdollah F. Contemporary Role of the Decipher® Test in Prostate Cancer Management: Current Practice and Future Perspectives. Rev Urol. 2016;18(1):1–9.

68. Klein EA, Haddad Z, Yousefi K, Lam LLC, Wang Q, Choeurng V, et al. Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. Urology. 2016;90:148–52.

69. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, et al. Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagnostics. 2013;15(2):263–9.

70. Greco FA, Spigel DR, Yardley DA, Erlander MG, Ma X, Hainsworth JD. Molecular Profiling in Unknown Primary Cancer: Accuracy of Tissue of Origin Prediction. Oncologist. 2010;15(5):500–6.

71. Hainsworth JD, Rubin MS, Spigel DR, Boccia R V., Raby S, Quinn R, et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: A prospective trial of the Sarah cannon research institute. J Clin Oncol. 2013;31(2):217–23.

72. Harrison G, Sosa JA, Jiang X. Evaluation of the Afirma gene expression classifier in repeat indeterminate thyroid nodules. Arch Pathol Lab Med. 2017;141(7):985–9.

73. Chudova D, Wilde JI, Wang ET, Wang H, Rabbee N, Egidio CM, et al. Molecular classification of thyroid nodules using high-dimensionality genomic data. J Clin Endocrinol Metab. 2010;95(12):5296–304.

74. Kim MI, Alexander EK. Diagnostic use of molecular markers in the evaluation of thyroid nodules. Vol. 18, Endocrine Practice. 2012. p. 796–802.

75. Ali SZ, Fish SA, Lanman R, Randolph GW, Sosa JA. Use of the Afirma® gene expression classifier for preoperative identification of benign thyroid nodules with indeterminate fine needle aspiration cytopathology. PLoS Currents. 2013. p. 1–7.

76. McIver B, Castro MR, Morris JC, Bernet V, Smallridge R, Henry M, et al. An independent study of a gene expression classifier (Afirma) in the evaluation of cytologically indeterminate thyroid nodules. J Clin Endocrinol Metab. 2014;99(11):4069–77.

77. Lastra RR, Pramick MR, Crammer CJ, LiVolsi VA, Baloch ZW. Implications of a suspicious Afirma test result in thyroid fine-needle aspiration cytology: An institutional experience. Cancer Cytopathol. 2014;122(10):737–44.

78. Kim JY, Park SC, Lee JK, Choi SJ, Hahm KS, Park Y. Novel Antibacterial Activity of β2-Microglobulin in Human Amniotic Fluid. PLoS One. 2012;

79. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31(11):1023–31.

80. Wang K, Sanchez-Martin M, Wang X, Knapp KM, Koche R, Vu L, et al. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. Leukemia. 2017;31(1):151–8.

81. Tarlock K, He J, Zhong S, Ries RE, Bailey M, Morley S, et al. Distinct age-associated genomic profiles in acute myeloid leukemia (AML) using FoundationOne heme. J Clin Oncol. 2016;34(15_suppl):7041–7041.

82. Lieber DS, Kennedy MR, Johnson DB, Greenbowe JR, Frampton GM, Schrock AB, et al. Abstract B16: Validation and clinical feasibility of a Foundation Medicine assay to identify immunotherapy response potential through tumor mutational burden (TMB). In 2017.

83. Lee Deak K, Jackson JB, Valkenburg KC, Keefer LA, Robinson Gerding KM, Angiuoli SV, et al. Next-Generation Sequencing Concordance Analysis of Comprehensive Solid Tumor Profiling between a Centralized Specialty Laboratory and the Decentralized PGDx Elio Tissue Complete Kitted solution. J Mol Diagnostics [Internet]. 2021 Jul;in press. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1525157821002105

84. Labriola MK, Zhu J, Gupta R, McCall S, Jackson J, Kong EF, et al. Characterization of tumor mutation burden, PD-L1 and DNA repair genes to assess relationship to immune checkpoint inhibitors response in metastatic renal cell carcinoma. J Immunother Cancer [Internet]. 2020 Mar;8(1):e000319. Available from: https://jitc.bmj.com/lookup/doi/10.1136/jitc-2019-000319

85. Yamani MH, Taylor DO, Rodriguez ER, Cook DJ, Zhou L, Smedira N, et al. Transplant Vasculopathy Is Associated With Increased AlloMap Gene Expression Score. J Hear Lung Transplant. 2007;26(4):403–6.

86. Yamani MH, Taylor DO, Haire C, Smedira N, Starling RC. Post-transplant ischemic injury is associated with up-regulated AlloMap gene expression. Clin Transplant. 2007;21(4):523–5.

87. Kobashigawa J, Patel J, Azarbal B, Kittleson M, Chang D, Czer L, et al. Randomized Pilot Trial of Gene Expression Profiling Versus Heart Biopsy in the First Year after Heart Transplant: Early Invasive Monitoring Attenuation Through Gene Expression Trial. Circ Hear Fail. 2015;8(3):557–64.

88. Wingrove JA, Daniels SE, Sehnert AJ, Tingley W, Elashoff MR, Rosenberg S, et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. Circ Cardiovasc Genet. 2008;1(1):31–8.

89. Rosenberg S, Dehais C, Ducray F, Alentron A, Tanguy M, De Reyneis A, et al. OS11.3 Machine learning for better prognostic stratification and driver genes identification in 1p/19q-codeleted grade III gliomas. Neuro Oncol [Internet]. 2017;19(suppl_3):iii22–iii22. Available from: http://dx.doi.org/10.1093/neuonc/nox036

90. Vargas J, Lima JAC, Kraus WE, Douglas PS, Rosenberg S. Use of the Corus® CAD Gene Expression Test for Assessment of Obstructive Coronary Artery Disease Likelihood in Symptomatic Non-Diabetic Patients. PLoS Currents. 2013.

91. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC Med Genomics. 2011;4.

92. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010;153(7):425–34.

93. Brahe CH, Østergaard M, Johansen JS, Defranoux N, Wang X, Bolce R, et al. Predictive value of a multi-biomarker disease activity score for clinical remission and radiographic progression in patients with early rheumatoid arthritis: a post-hoc study of the OPERA trial. Scand J Rheumatol. 2019;

94. Chernoff D, Scott Eastman P, Hwang CC, Flake DD, Wang X, Kivitz A, et al. Determination of the minimally important difference (MID) in multi-biomarker disease activity (MBDA) test

scores: impact of diurnal and daily biomarker variation patterns on MBDA scores. Clin Rheumatol. 2019;38(2):437–45.

95. Curtis JR, Weinblatt ME, Shadick NA, Brahe CH, Østergaard M, Hetland ML, et al. Validation of the adjusted multi-biomarker disease activity score as a prognostic test for radiographic progression in rheumatoid arthritis: a combined analysis of multiple studies. Arthritis Res Ther. 2021;

96. Curtis JR, Xie F, Yang S, Danila MI, Owensby JK, Chen L. Uptake and Clinical Utility of Multibiomarker Disease Activity Testing in the United States. J Rheumatol [Internet]. 2019;46(3):237–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30442830

97. Curtis JR, Van Der Helm-Van Mil AH, Knevel R, Huizinga TW, Haney DJ, Shen Y, et al. Validation of a novel multibiomarker test to assess rheumatoid arthritis disease activity. Arthritis Care Res. 2012;64(12):1794–803.

98. Food and Drug Administration. Helix Genetic Health Risk App For Late-Onset Alzheimer's Disease - FDA Review Decision Summary. 2020; Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K192073

99. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun. 2020;

100. Lu JT, Ferber M, Hagenkord J, Levin E, South S, Kang HP, et al. Evaluation for Genetic Disorders in the Absence of a Clinical Indication for Testing: Elective Genomic Testing. Journal of Molecular Diagnostics. 2019.

101. Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. Nat Med. 2020;

102. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Vol. 8, Frontiers in Genetics. 2017.

103. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics [Internet]. 2016;17(S2):S15. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0857-9

104. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Vol. 19, Computational and Structural Biotechnology Journal. 2021. p. 3735–46.

105. Lemke C, Budka M, Gabrys B. Metalearning: a survey of trends and technologies. Artif Intell Rev. 2015;

106. Sima C, Braga-Neto UM, Dougherty ER. High-dimensional bolstered error estimation. Bioinformatics. 2011;

107. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical Reviews. 2019.

108. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Br Bioinform. 2008/03/04. 2008;9(2):119–28.

109. Grollemund V, Pradat PF, Querin G, Delbot F, Le Chat G, Pradat-Peyre JF, et al. Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions. Front Neurosci [Internet]. 2019;13. Available from: http://dx.doi.org/10.3389/fnins.2019.00135

110. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet [Internet]. 2019;10(MAR). Available from:

http://dx.doi.org/10.3389/fgene.2019.00267

111. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension. Curr Hypertens Rep. 2018/07/08. 2018;20(9):75.

112. Long NP, Yoon SJ, Anh NH, Nghi TD, Lim DK, Hong YJ, et al. A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer. Metabolomics. 2019/03/05. 2018;14(8):109.

113. Martinez BI, Stabenfeldt SE. Current trends in biomarker discovery and analysis tools for traumatic brain injury. J Biol Eng [Internet]. 2019;13(1). Available from: http://dx.doi.org/10.1186/s13036-019-0145-8

114. Patil S, Awan KH, Arakeri G, Seneviratne CJ, Muddur N, Malik S, et al. Machine learning and its potential applications to the genomic study of head and neck cancer-A systematic review. J Oral Pathol Med. 2019;48(9):773–9.

115. Saini G, Mittal K, Rida P, Janssen EAM, Gogineni K, Aneja R. Panoptic view of prognostic models for personalized breast cancer management. Cancers (Basel) [Internet]. 2019;11(9). Available from: http://dx.doi.org/10.3390/cancers11091325

116. Europen Council. Council conclusions on personalised medicine for patients. Off J Eur Union [Internet]. 2015;431(2):1–4. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015XG1217(01)&from=EN

117. Labriola MK, Zhu J, Gupta R, McCall S, Jackson J, Kong EF, et al. Characterization of tumor mutation burden, PD-L1 and DNA repair genes to assess relationship to immune checkpoint inhibitors response in metastatic renal cell carcinoma. J Immunother Cancer. 2020;8(1).

Author Contributions

Study conception and design: EG, AR.

Methodology: CG, RB.

Data collection and analysis: EG, AR

Original draft preparation: EG

Review and editing: AR, EG, PG, CG, JDM, RB.

Project supervision: PG

Funding acquisition: JDM.

All authors have read and revised the manuscript and approved the final version.

The members of the PERMIT group were involved in the preparation or revision of the joint protocol of the four scoping reviews of the PERMIT series, attended the joint workshop (consultation exercise) and are co-authors of the other scoping reviews of the PERMIT series.

Collaborators

PERMIT group:

1. Antonio L. Andreu
2. Florence Bietrix,
3. Florie Brion Bouvier
4. Montserrat Carmona Rodriguez
5. Maria del Mar Polo-de Santos,
6. Maddalena Fratelli,
7. Rainer Girgenrath,
8. Alexander Grundmann,
9. Josep Maria Haro,
10. Frank Hulstaert,
11. Iñaki Imaz-Iglesia,
12. Setefilla Luengo Matos
13. Emmet McCormack,
14. Albert Sanchez Niubo,
15. Emanuela Oldoni,
16. Raphael Porcher,
17. Vibeke Fosse,
18. Luis M. Sánchez-Gómez,
19. Lorena San Miguel,
20. Cecilia Superchi,
21. Teresa Torres,
22. Anna Monistrol Mula

Funding statement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874825.

Competing interests statement

None declared

Ethics approval

This study was based entirely on a scoping review of relevant published literature and did not require an ethics approval.

Patient consent

This study did not require consent from patients, because it does not use individual data.

Permission to reproduce material from other sources

This study has cited all references which are published and publicly available.

Data sharing statement

The study protocol was published on the online platform Zenodo (19). Copies of searches and data extraction sheets will be made publicly available on Zenodo as part of the database collection for all scoping reviews conducted in the PERMIT project.

Figure legends

Fig. 1. Keyword based search strategy for the scoping review. Four categories of keywords were defined to retrieve relevant articles from the biomedical literature on machine learning analyses of omics data for personalised medicine, which include a validation study (highlighted by the coloured boxes in the centre). For each category relevant keywords were determined, including controlled vocabulary terms from the Medical Subject Headings (MeSH) thesaurus by the US National Library of Medicine (upper and lower boxes with frames coloured according to the corresponding category). As indicated by the keyword "AND" in the centre, a conjunctive search was conducted, i.e., every retrieved article had to contain at least one keyword from each category. This strategy was adapted for searching the other databases.

Fig. 2. Study selection flow diagram. Flow diagram of the procedure for the scoping review article identification, screening, eligibility assessment, and final inclusion, according to the PRISMA scheme (31). Reasons for excluding full-text were not mutually exclusive.

Fig. 3. Validation methods used in omics biomarker studies. Stacked bar chart of the number of articles retrieved in the scoping review for different categories of validation methods used in the underlying biomarker studies (covering time periods from 2000 to 2021). The majority of studies use only internal cohort validation approaches, such as cross-validation (CV), training/test set split validation, resampling/bootstrapping-based validation, out-of-bag validation (for tree-based classifiers), and combinations of CV and test set validation within the same cohort. Studies with an external validation on an independent patient cohort (with or without an additional internal cross-validation) are still underrepresented, even in more recent time periods. All filtered full-text articles derived from the scoping review except for review articles were included in the analysis.

Fig. 4. Map representation of country statistics for the selected articles. The number of articles originating from different countries among the studies selected in the full-text review are visualized on a world map representation using a colour gradient from blue (1 article) to red (98 articles = maximum contribution by a single country; using a logarithmic colour gradient scale to highlight differences over a broad value range).

Fig. 5. Representation of study types among the selected articles. The percentage of articles describing case-control studies, therapy/drug response studies, differential diagnosis studies, prognostic and survival prediction studies, as well as review studies and other study types is represented as a pie chart.

Fig. 6. Characteristics of successful omics-based studies. Six main categories of design and implementation aspects that characterize successful omics-based biomarker development studies were identified (starting from the centre left in the figure and proceeding clockwise): 1) Adequacy of the study design & sample size selection; 2) Rigor and robustness of the statistical evaluation; 3) Clarity of scope and goals; 4) Completeness and reproducibility of the study documentation; 5) Interpretability and biological plausibility of the created predictive models; 6) Integration of prior biological knowledge into the model building and validation procedures.

Fig. 7. Recommended generic workflow for biomarker development using machine learning analysis of omics data. The machine learning analysis of omics data for biomarker discovery and validation should ideally involve dedicated quality control and pre-processing analyses, a dimension reduction using unsupervised feature selection (e.g. a variance filtering) or data transformation approaches (e.g. using a Principal Components Analysis), a cross-validation on the discovery cohort, and an external validation on a distinct validation cohort.

Tables

Name	Test approval type	Purpose (Data type used for discovery)	References
MammaPrint®	FDA-cleared Assay	breast cancer risk-of-recurrence assessment (DNA microarray gene expression data)	(6,38–41)
ColoPrint®	LDT	colon cancer development of distant metastasis prediction (DNA microarray gene expression data)	(42–47)
Prosigna® Assay / PAM50	FDA-cleared Assay	breast cancer risk of distant recurrence prediction (DNA microarray gene expression data)	(49–53)
Oncotype DX®	LDT	breast cancer risk-of-recurrence assessment (DNA microarray gene expression data)	(8,56–59)
Decipher®	LDT	prostate cancer metastatic risk prediction (DNA microarray gene expression data)	(9,64–68)
Cancer Type ID®	LDT	predict tumour type for cancers of unknown / uncertain diagnosis (DNA microarray gene expression data)	(15,69–71)
Afirma™ Gene Expression Classifier	LDT	discriminate between benign and cancerous thyroid nodules (DNA microarray gene expression data)	(72–77)
Foundation One® Heme	LDT	test for hematologic malignancies, sarcomas, or solid tumours (RNA and DNA sequencing data)	(14,79–81)
PGDx Elio™ Tissue Complete	FDA-cleared Assay	test to assess somatic mutations and tumour mutation burden for solid tumours (DNA sequencing data)	(83,117)
AlloMap® Heart	FDA-cleared Assay	identifying heart transplant recipients with risk of cellular rejection (DNA microarray gene expression data)	(13,85–87)
Corus® CAD	LDT	identify obstructive coronary artery disease (DNA microarray gene expression data)	(11,88–91)
Vectra® DA	LDT	multi-biomarker blood test for rheumatoid arthritis (immunoassay + clinical data, 396 candidate biomarkers derived from integrative data analysis)	(93–96)
Helix® Laboratory Platform & Health Risk App for Late-onset Alzheimer's	FDA-cleared medical device	whole exome sequencing constituent device based for reporting and interpreting general genetic health risks (DNA sequencing data)	(99–101)

Tab. 1. Examples of clinically approved omics-derived diagnostic or prognostic tests designs applied

to personalised medicine (synonyms for the same test are separated by the "/"-symbol). FDA-approval status was checked on the web-site by the FDA (37) and reflects the status as of July 2021.

“stratified medicine” OR biomarker* OR “precision medicine” OR “personalized medicine” OR “personalised medicine” OR “individualized Medicine“ OR “individualised Medicine“ OR “individualized therapy“ OR “individualised therapy“ OR “patient stratification” OR pharmacogenetics OR “patient specific modeling” OR “personalized clinical decision making” OR “personalised clinical decision making” OR “prediction of response” OR “prediction of responses” OR "Biomarkers"[Mesh] OR "Precision Medicine"[Mesh]

Genomics"[Mesh]) OR "Metabolomics"[Mesh]) OR "Epigenomics"[Mesh]) OR "Microarray Analysis"[Mesh]) OR "Mass Spectrometry"[Mesh] OR Omic* OR “omic based” OR “multi omic” OR “multi omics” OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR “RNA seq” OR “mass spectrometry")

PERSONALISED MEDICINE

MACHINE LEARNING

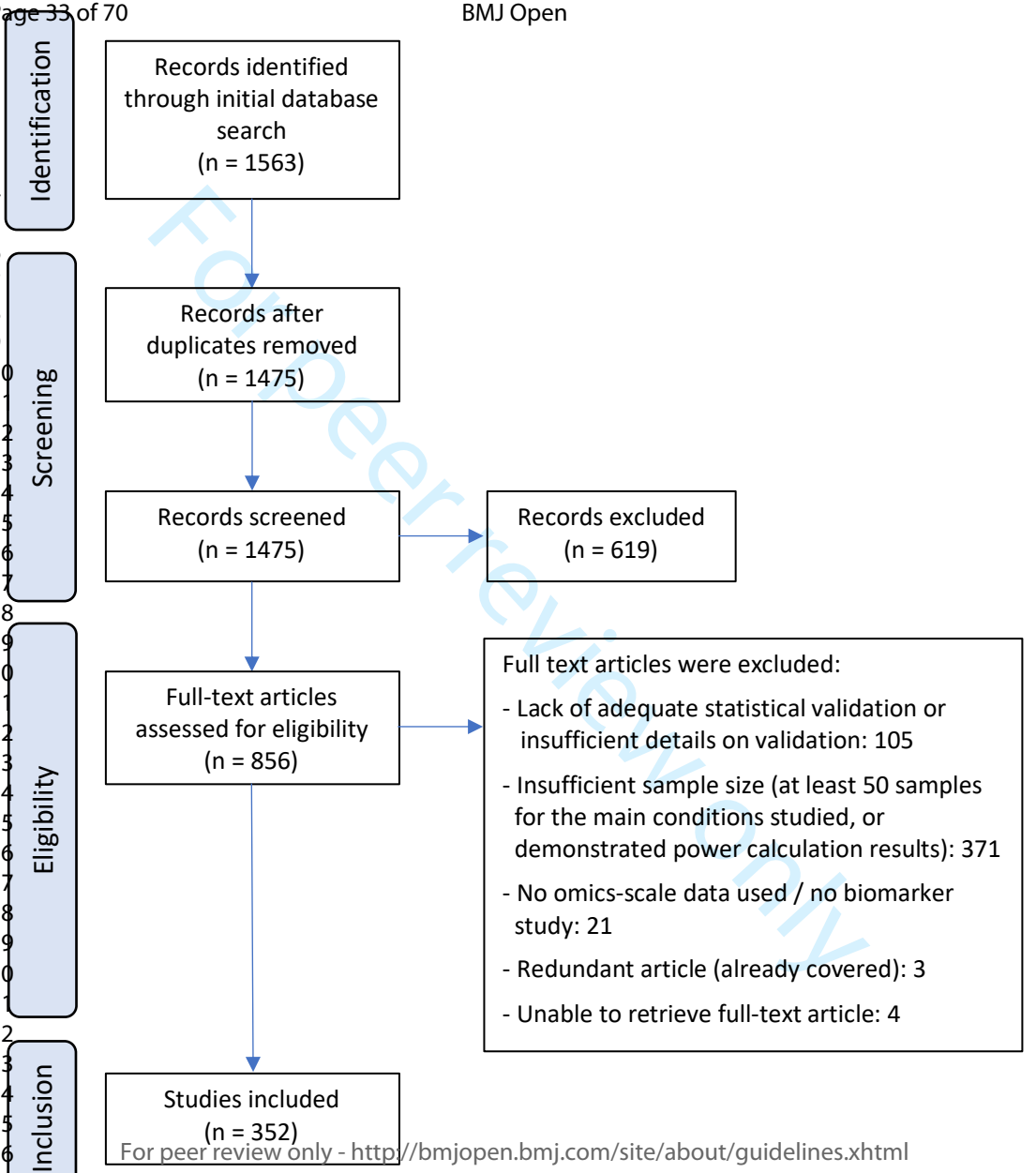
AND

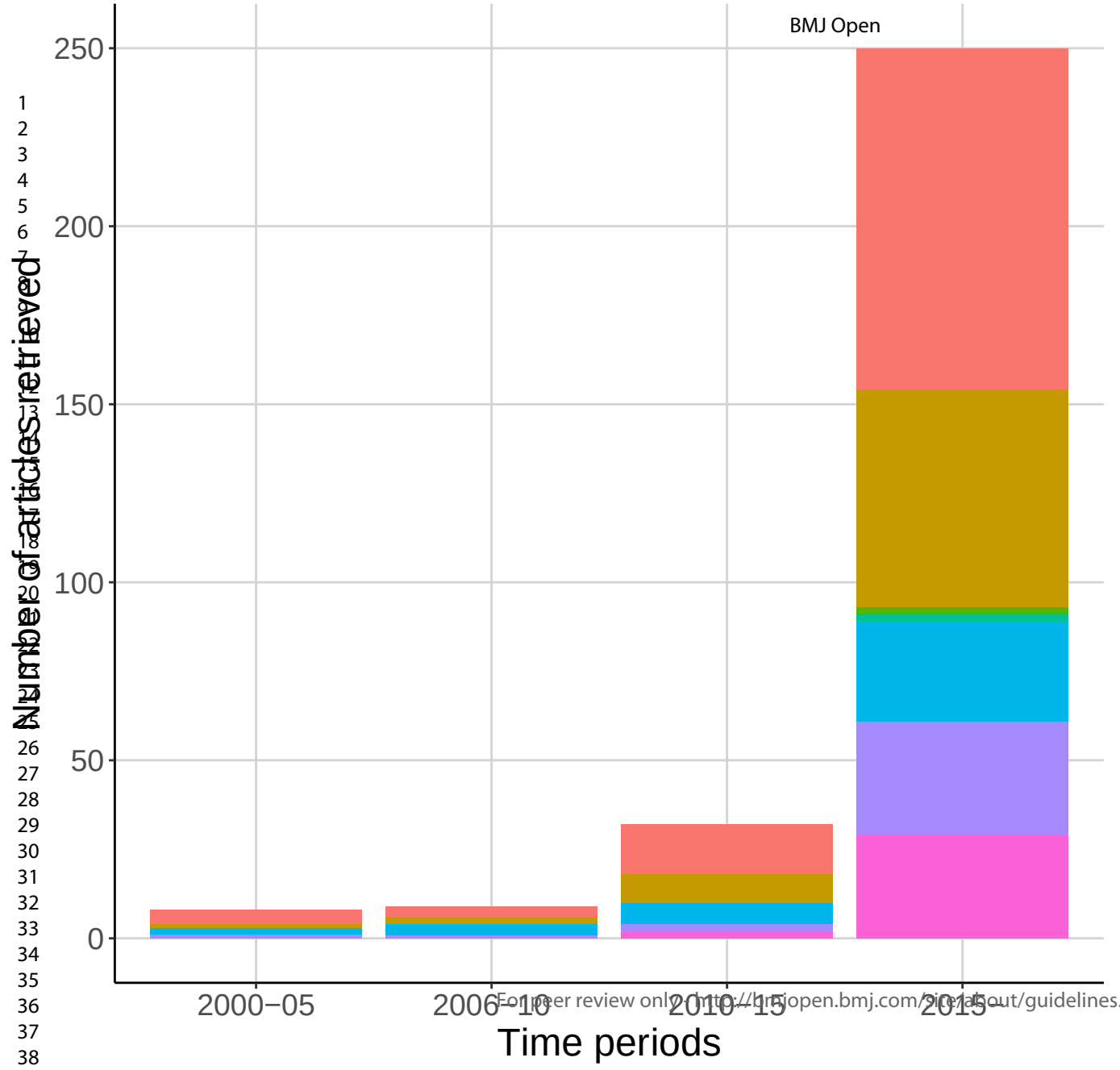
OMICS

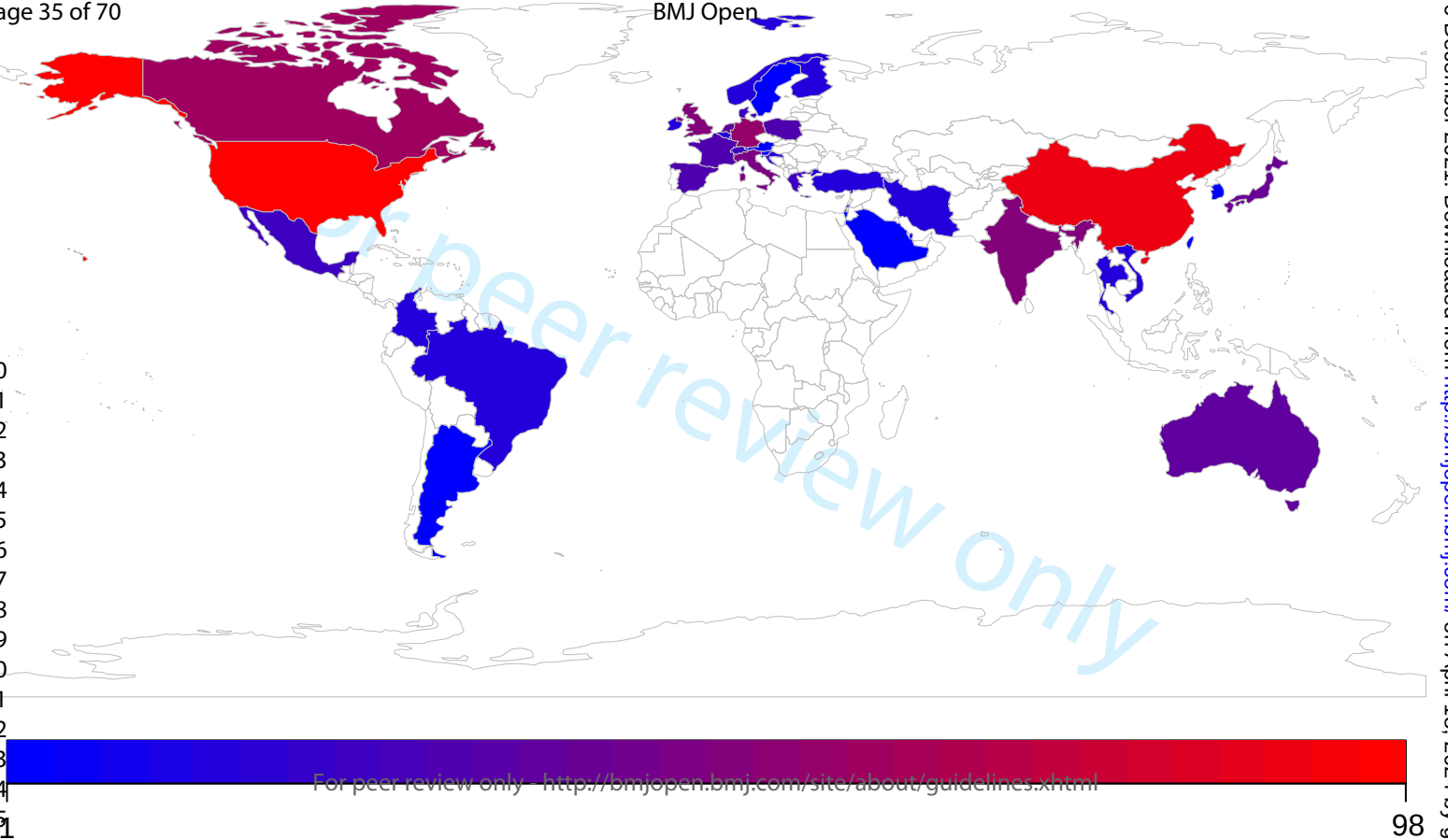
VALIDATION

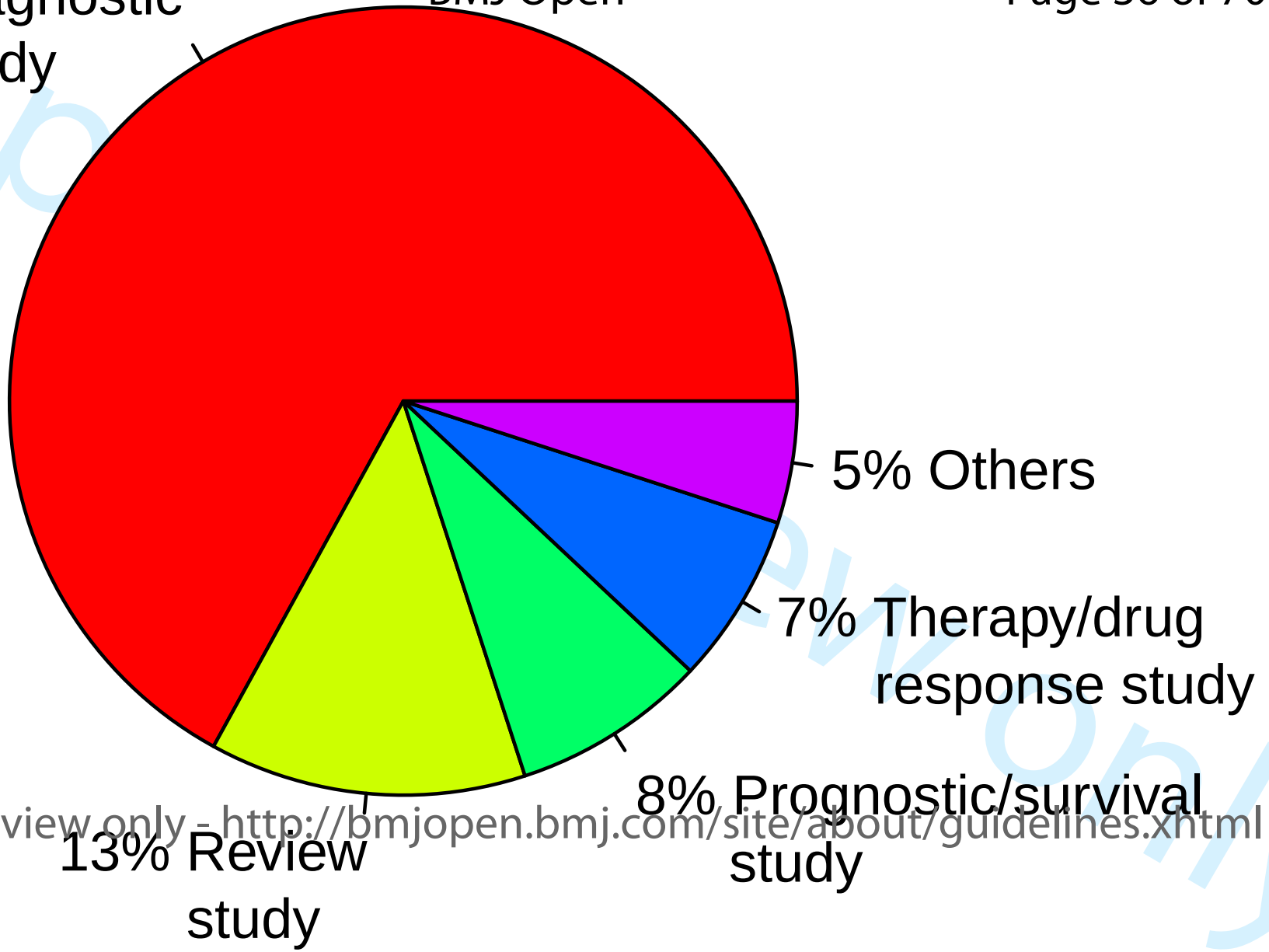
Machine Learning"[Mesh] OR «Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”

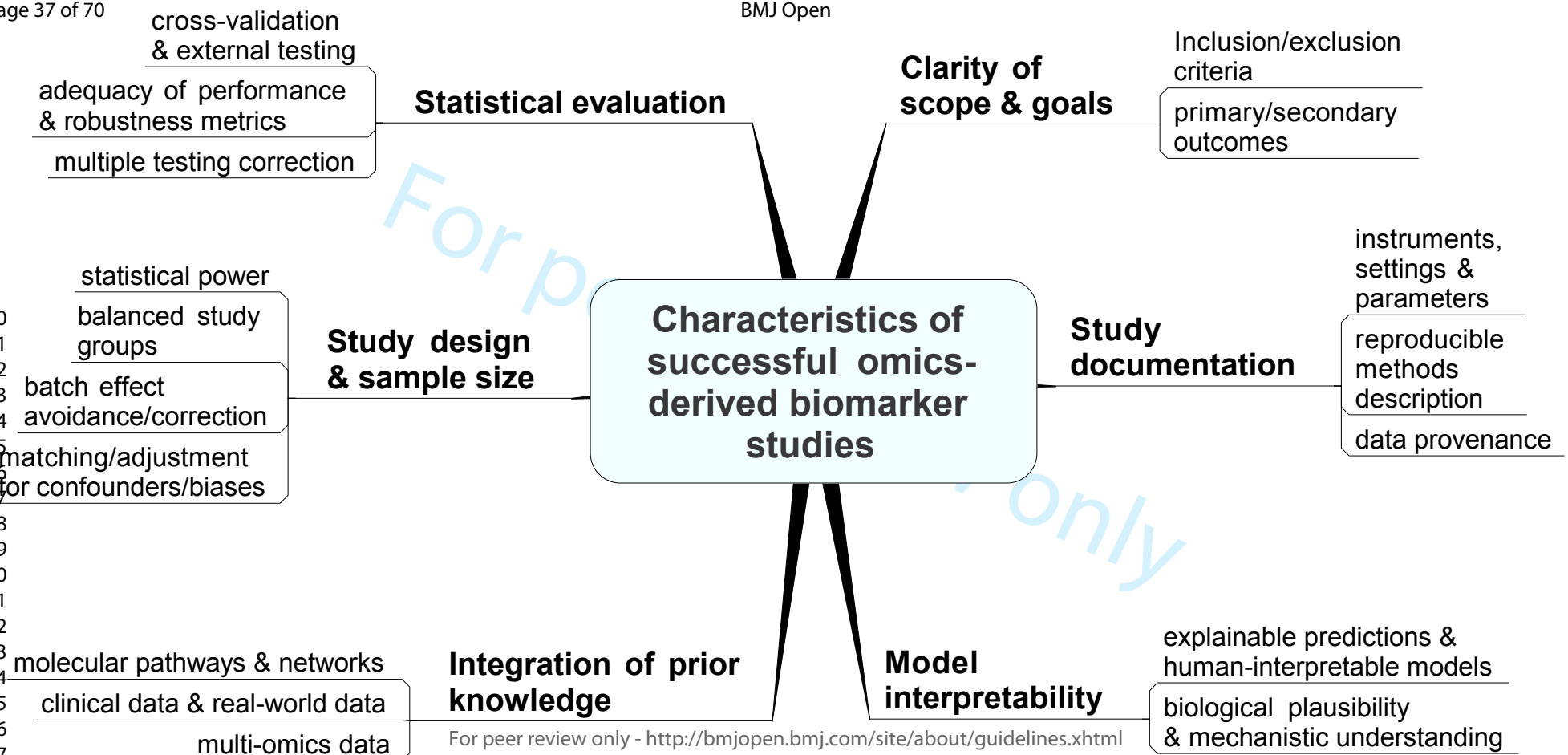
Validation Studies as Topic"[Mesh]) OR "Validation Study" [Publication Type] OR "Sensitivity and Specificity"[Mesh]) OR "Benchmarking"[Mesh]) OR validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR ”cross study” OR ”cross studies”)

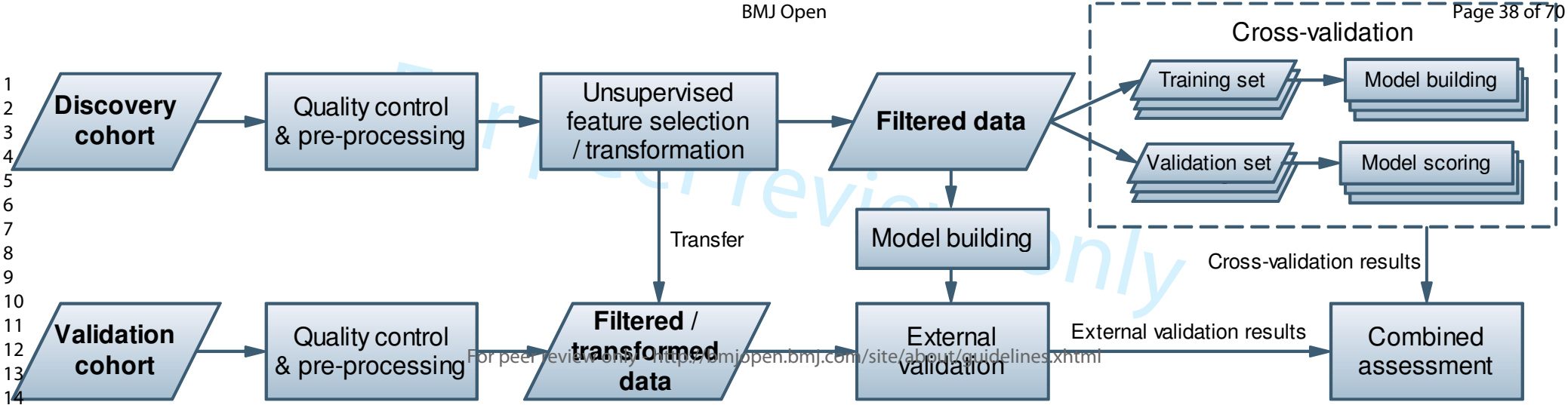












Online Supplementary file 1 – PRISMA-ScR Checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (17).

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	5
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5-6
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	5
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5 (Online Suppl. File 2)
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	6

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	6 (Online Suppl. File 3)
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	6-7
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	7 (Fig. 2)
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	7 (Table 1)
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	7-13
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	7-13
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	13
Limitations	20	Discuss the limitations of the scoping review process.	14
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and	14-16

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
		objectives, as well as potential implications and/or next steps.	
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	25

Online Supplementary file 2 – Search strategy

Keyword searches conducted in the databases *PubMed*, *EMBASE* and *Web of Science* as part of the scoping review.

1) PubMed Query

Search: (((((("Machine learning" OR "statistical learning" OR "supervised learning" OR "unsupervised learning")) OR ("Machine Learning"[Mesh])) AND (("Biomarkers"[Mesh]) OR "Precision Medicine"[Mesh])) AND (((Omic* OR "omic based" OR "multi omic" OR "multi omics" OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR "RNA seq" OR "mass spectrometry")) OR ("Genomics"[Mesh] OR "Metabolomics"[Mesh] OR "Epigenomics"[Mesh] OR "Microarray Analysis"[Mesh] OR "Mass Spectrometry"[Mesh]))) AND ((validation OR validity OR validated OR "cross validation" "cross validated" OR "clinical utility*" OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark* OR bias OR "cross study" OR "cross studies"))
AND
(("2000/01/01"[Date - Entry]: "2021/07/20"[Date - Entry])) Filters: English, French, Italian, Spanish

2) Embase Query

#25: #24 AND [embase]/lim NOT [medline]/lim
#24: #23 AND [2000-2021]/py
#23: #20 AND #21 AND ([english]/lim OR [french]/lim OR [italian]/lim OR [spanish]/lim)
#22: #20 AND #21
#21: omic*:ti,ab OR 'machine learning':ti,ab OR 'personalized medicine':ti,ab OR 'personalised medicine':ti,ab
#20: #4 AND #10 AND #16 AND #19
#19: #17 OR #18
#18: validation:ti,ab OR validity:ti,ab OR validated:ti,ab OR 'cross validation':ti,ab OR 'cross validated':ti,ab OR test*:ti,ab OR 'clinical utility*':ti,ab OR accuracy:ti,ab OR robustness:ti,ab OR reliability*:ti,ab OR sensitivity:ti,ab OR specificity:ti,ab OR benchmark*:ti,ab OR bias:ti,ab OR 'cross study:ti,ab' OR 'cross studies':ti,ab
#17: 'validation study'/exp OR 'reliability'/exp OR 'sensitivity and specificity'/exp OR 'benchmarking'/exp
#16: #14 OR #15
#15: omic*:ti,ab OR 'omic based':ti,ab OR 'multi omic*':ti,ab OR genomic*:ti,ab OR transcriptomic*:ti,ab OR proteomic*:ti,ab OR metabolomic*:ti,ab OR lipidomic*:ti,ab OR epigenomic*:ti,ab OR microarray:ti,ab OR 'rna seq':ti,ab OR 'mass spectrometr*':ti,ab
#14: #11 OR #12 OR #13
#13: 'mass spectrometry'/exp
#12: 'microarray analysis'/exp
#11: 'omics'/exp OR 'genomics'/exp OR 'epigenetics'/exp
#10: #5 OR #6 OR #7 OR #8 OR #9
#9: 'individualized medicine':ti,ab OR 'individualised medicine':ti,ab OR 'individualized therapy':ti,ab OR 'individualised therapy':ti,ab
#8: 'personalised medicine':ti,ab
#7: 'personalized medicine':ti,ab
#6: 'stratified medicine':ti,ab OR cluster*:ti,ab OR 'sub group*':ti,ab OR subgroup*:ti,ab OR biomarker*:ti,ab OR diagnos*:ti,ab OR prognos*:ti,ab OR 'precision medicine':ti,ab
#5: 'biological marker'/exp OR 'personalized medicine'/exp
#4: #1 OR #2 OR #3
#3: 'machine learning'/exp
#2: 'statistical learning'/exp

#1: 'machine learning':ti,ab OR 'statistical learning':ti,ab OR 'supervised learning':ti,ab OR 'unsupervised learning':ti,ab

3) Web of Science Query

(((((#1) AND #1) AND #2) AND #3) AND #4) AND #5
5: (ALL=(((validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark*
OR bias OR ”cross study” OR ”cross studies”)))) AND ALL=(((omic* OR "machine learning" OR
"personalized medicine" OR "personalised Medicine")))
4: ALL=(((validation OR validity OR validated OR “cross validation” “cross validated” OR “clinical utility*” OR accuracy OR robustness OR reliability* OR sensitivity OR specificity OR benchmark*
OR bias OR ”cross study” OR ”cross studies”)))
3: ALL=(TOPIC: ((Omic* OR “omic based” OR “multi omic*” OR genomic* OR transcriptomic* OR proteomic* OR metabolomic* OR lipidomic* OR epigenomic* OR microarray OR “RNA seq”
OR “mass spectrometr*”)))
2: (ALL=(TOPIC: ((“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)))) AND ALL=(TOPIC: ((“stratified medicine” OR cluster* OR “sub group*” OR Subgroup* OR biomarker* OR diagnos* OR prognos* OR “precision medicine” OR
“personalized medicine”OR “personalised medicine” OR “individualized Medicine“ OR “individualised Medicine“ OR “individualized therapy“ OR “individualised therapy“)))
1: ALL=(TOPIC: ((“Machine learning” OR “statistical learning” OR “supervised learning” OR “unsupervised learning”)))

Online Supplementary file 3 – Data extraction form

Data items extracted from each processed article during the full-text scoping review, and associated qualifications for each item.

Item	Qualifications
Authors	
Title	
Journal	
Volume	
Issue	(if applicable)
Pages	(if applicable)
Year	
Location	
URL / DOI	
Type of publication	• Research article • Meeting abstract • Review
Study population and sample size	(if applicable)
Methodology / Study Design	• Case-control study • Cases only stratification study (+ further qualification, e.g. treatment response prediction, tumor subtype categorization, recurrence/relapse prediction, survival prediction, tissue-of-origin prediction)
Outcome assessment	• Performance measures (e.g. accuracy, sensitivity, specificity, Kohen's Kappa, F-score, AUC) • Validation scheme (cross-validation approach, external validation approach, single cohort or multiple cohorts)
Generic machine learning category	• Supervised learning • Unsupervised learning • Other / mixed approaches
Name of specific machine learning approach	(if applicable)

Main results / key findings that relate to the research question	short description

The spreadsheet version of this table has been made available at https://doi.org/10.5281/zenodo.5646467 (navigating the PDF version below will require a PDF viewer with zoom and search functionality)

Number	Authors	Title	Journal	Volume	Issue	Pages	Year	Location	URL / DOI	Type of publication	Study population and sample size if applicable	Methodology Study design	Outcome measures, if applicable	Validation type	Main results	Key findings that relate to the review question
1	Agranoff, D and Fernandez-Reyes, D and Papadopoulos, M C and Rojas, S A and Herbster, M and Loosemore, A and Tarelli, E and Sheldon, J and Schwenk, A and Pollok, R and Rayner, C F and Krishna, S	Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum	Lancet	368	9540	1012-1021	2006	UK	http://dx.doi.org/10.1016/s0140-6736(06)69342-2	article	179 serum samples from patients, 170 serum samples from controls	Case-control study	accuracy, sensitivity, specificity (in k-fold cross-validation + validation set test)	cross-validation + test set	SVM classifier discriminated the proteomic profile of patients with active tuberculosis from that of controls with overlapping clinical features. Diagnostic accuracy was 94% (sensitivity 93·5%, specificity 94·9%) for patients with tuberculosis.	Patients with active tuberculosis can be distinguished from controls with overlapping clinical features using machine learning on proteomics data.
2	Ali, M and Aittokallio, T	Machine learning and feature selection for drug response prediction in precision oncology applications	Biophysical Reviews	11	1	31-39	2019	Finland	http://dx.doi.org/10.1007/s12551-018-0446-z	article	review (not applicable)	Review			Technical variability and frequent missingness in input “big data” require the application of dedicated data preprocessing pipelines that often lead to some loss of information and compressed view of the biological signal. Most of the variability in the drug response levels across the cell lines can be explained by the genome-wide gene expression data, whereas the other omics profiles only marginally improve the prediction performance.(Jang et al. 2014; Costello et al. 2014). However, the use of multiple omics profiles from various biological levels can still improve the prediction results	Technical variability and frequent missingness in input “big data” require the application of dedicated data preprocessing pipelines that often lead to some loss of information and compressed view of the biological signal. Most of the variability in the drug response levels across the cell lines can be explained by the genome-wide gene expression data, whereas the other omics profiles only marginally improve the prediction performance.(Jang et al. 2014; Costello et al. 2014). However, the use of multiple omics profiles from various biological levels can still improve the prediction results
3	Alqudah, A M	Ovarian Cancer Classification Using Serum Proteomic Profiling and Wavelet Features A Comparison of Machine Learning and Features Selection Algorithms	Journal of Clinical Engineering	44	4	165-173	2019		http://dx.doi.org/10.1097/JCE.0000000000000359	article	262 cancer patients, 191 controls	Case-control study	accuracy, sensitivity, and precision (70% training set, 30% test set split)	training + test set	Results show that all the presented ML algorithms performed well for Ovarian Cancer Classification, with different feature selection algorithms all exceeding 90% accuracy.	Results show that all the presented ML algorithms performed well for Ovarian Cancer Classification, with different feature selection algorithms all exceeding 90% accuracy.
4	Ashton, N J and Nevado-Holgado, A J and Lynham, S and Ward, M and Gupta, V B and Chatterjee, P and Gooze, K and Hone, E and Pedrini, S and Bush, A I and Rowe, C C and Villemagne, V L L and Ames, D and Masters, C L and Aarsland, D and Lovestone, S and Martins, R N and Hye, A	A mass spectrometry-based discovery and replication of a multi-analyte classifier for neocortical amyloid pathology	Alzheimer's and Dementia	13	7	P1033-P1033	2017		http://dx.doi.org/10.1016/j.jalz.2017.06.1456	meeting abstract	297 participants	Cases only (predicint neocortical amyloid burden)	accuracy (5-fold cross validation, external testing in other cohort)	cross-validation + external cohort validation	Machine learning analysis revealed a 14-analyte MS proteomics panel that is able to predict NAB in a cognitively normal cohort at an accuracy of 86.6%. Pathway analysis highlights the convergence of pathways involved in coagula- tion, APP processing, neuronal transcription factors and axonal injury to be important in predicting NAB.	Machine learning analysis revealed a 14-analyte MS proteomics panel that is able to predict NAB in a cognitively normal cohort at an accuracy of 86.6%. Pathway analysis highlights the convergence of pathways involved in coagula- tion, APP processing, neuronal transcription factors and axonal injury to be important in predicting NAB.
5	Assawamakin, A and Prueksaaroon, S and Kulawonganunchai, S and Shaw, P J and Varavithya, V and Ruangrajitpakorn, T and Tongsima, S	Biomarker selection and classification of "-omics" data using a two-step bayes classification framework	Biomed Res Int	2013		148014-148014	2013	Thailand	http://dx.doi.org/10.1155/2013/148014	article	more than 45 microarray and proteomics datasets of different sizes used	Case-control study	AUC, accuracy, sensitivity, specificity (10-fold cross-validation)	cross-validation	Current machine learning approaches are either too complex or perform poorly. The proposed two-step Bayes classification framework was equal to and, in some cases, outperformed other classification methods in terms of prediction accuracy, minimum number of classification markers, and computational time.	Low-complexity machine learning models using few features can achieve similar performance as more complex models.
6	Awedat, K and Abdel-Qader, I and Springstead, J R	Prostate cancer recognition based on mass spectrometry sensing data and data fingerprint recovery	Biomedical Signal Processing and Control	33		392-399	2017	USA	http://dx.doi.org/10.1016/j.bspc.2016.12.003	article	237 blood samples from subjects with different PSA levels	Case-control study (cases with different PSA levels)	accuracy, sensitivity, specificity, PPV, NPV (10-fold CV)	cross-validation	The high dimensionality and noisy spectra of Mass Spectrometry (MS) data are two of the main challenges to achieving high accuracy in prostate cancer recognition. The objective of this work is to produce an accurate prediction of class content by employing compressive sensing (CS).	The study highlights the benefits of compressed sensing for dimensionality reduction in high-dimensional omics classification analyses.
7	Baer, C and Walter, W and Stengel, A and Hutter, S and Meggendorfer, M and Kern, W and Haferlach, C and Haferlach, T	Molecular classification of AML-MRC reveals a distinct profile and identifies MRC-like patients with poor overall survival	Blood	134			2019		http://dx.doi.org/10.1182/blood-2019-128234	meeting abstract	619 patients with survival data	Case-control study	accuracy (10-fold CV)	cross-validation	AML with myelodysplasia related changes (AML-MRC) can be diagnosed using patients' history and NGS-derived genetic information instead of morphology, allowing to identify 96-99% of AML-MRC as defined in WHO today.	Using patients' history and genetic information instead of morphology allow to identify 96-99% of AML-MRC as defined in WHO today
8	Barcelo, F and Gomila, R and de Paul, I and Gili, X and Segura, J and Perez-Montana, A and Jimenez-Marco, T and Sampol, A and Portugal, J	MALDI-TOF analysis of blood serum proteome can predict the presence of monoclonal gammopathy of undetermined significance	PLoS One	13	8	e0201793-e0201793	2018	Spain	http://dx.doi.org/10.1371/journal.pone.0201793	article	103 patients clinically diagnosed with MGUS, 108 healthy volunteer donors	Case-control study	accuracy, sensitivity, specificity (20-fold cross-validation)	cross-validation	MALDI-TOF analysis of blood serum proteome using support vector machines can predict the presence of monoclonal gammopathy of undetermined significance	MALDI-TOF analysis of blood serum proteome using support vector machines can predict the presence of monoclonal gammopathy of undetermined significance
9	Barla, A and Jurman, G and Riccadonna, S and Merler, S and Chierici, M and Furlanello, C	Machine learning methods for predictive proteomics	Brief Bioinform	9	2	119-128	2008	Italy	https://doi.org/10.1093/bib/bbn008	article	review (not applicable)	Review			The aim of this review is to explain how to build a general purpose design analysis protocol (DAP) for predictive proteomic profiling: we show how to limit leakage due to parameter tuning and how to organize classification and ranking on large numbers of replicate versions of the original data to avoid selection bias. A procedure for assessing stability and predictive value of the resulting biomarkers’ list is also provided.	The review presents common techniques for avoiding selection bias and assessing the stability and predictie value of biomarkers for proteomics machine learning studies.
10	Baron, D and Ramstein, G and Chesneau, M and Echasseriau, Y and Pallier, A and Paul, C and Degauque, N and Hernandez-Fuentes, M and Sanchez-Fueyo, A and Newell, K and Giral, M and Soulillou, J P and Houlgatte, R and Brouard, S	A common gene signature across multiple studies identifies biomarkers and functional regulation in tolerance to renal allograft	American Journal of Transplantation	15			2015		http://dx.doi.org/10.1038/ki.2014.395	article	96 samples with tolerance to renal allograft	Case-control study	accuracy, sensitivity, specificity (6-fold cross-validation + external validation)	cross-validation + test set	A gene signature derived from blood cell transcriptional data predicts tolerance to renal allograft correctly in 92% of cases.	A gene signature derived from blood cell transcriptional data predicts tolerance to renal allograft correctly in 92% of cases.
11	Bashiri, A and Ghazisaeedi, M and Safdari, R and Shahmoradi, L and Ehtesham, H	Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review	Iranian Journal of Public Health	46	2	165-172	2017	Iran	https://pubmed.ncbi.nlm.nih.gov/28451550/	article	review (not applicable)	Case-control study			By attention to the capabilities of machine learning techniques in proteomics and genomics applications, developing clinical decision support systems based on these methods for analyzing gene expression data can prevent potential errors in survival estimation, provide appropriate and individualized treatments to patients and improve the prognosis of cancers.	By attention to the capabilities of machine learning techniques in proteomics and genomics applications, developing clinical decision support systems based on these methods for analyzing gene expression data can prevent potential errors in survival estimation, provide appropriate and individualized treatments to patients and improve the prognosis of cancers.
12	Baumbach, J	Computational systems medicine-what we can learn from Arnold Schwarzenegger about breast cancer survival	Systems Medicine	2	1	A-29	2019		http://dx.doi.org/10.1089/sysm.2019.29005	article	review (not applicable)	Case-control study			In transcriptomics-based breast cancer survival prediction, a high number of features (>22K genes) and usually a small number of samples (<1K patients) leads to model overfitting and a lack of statistical robustness. To address this problem we introduce computational approaches for network-based medicine.	The authors present novel developments in graph-based machine learning to increase robust in statistics and prediction power for breast cancer survival prediction, validating them by a investigating the drop in performance when applying permutation tests.
13	Baumgartner, C and Bohm, C and Baumgartner, D	Modelling of classification rules on metabolic patterns including machine learning and expert knowledge	J Biomed Inform	38	2	89-98	2005	Austria	https://doi.org/10.1016/j.jbi.2004.08.009	article	PAHD, n = 94 cases 1241 randomly sampled controls)	Case-control study	accuracy, sensitivity, specificity (10-fold cross-validation)	cross-validation	we investigated metabolic patterns of three severe metabolic disorders, PAHD, MCADD, and 3-MCCD, on which we constructed classification models for disease screening and diagnosis using a decision tree paradigm and logistic regression analysis (LRA). For the LRA model-building process we assessed the relevance of established diagnostic flags, which have been developed from the biochemical knowledge of newborn metabolism. Both approaches yielded comparable classification accuracy in terms of sensitivity (>95.2%), while the LRA models built on flags showed significantly enhanced specificity. Both approaches yielded comparable classification accuracy in terms of sensitivity (>95.2%), while the LRA models built on flags showed significantly enhanced specificity.	we investigated metabolic patterns of three severe metabolic disorders, PAHD, MCADD, and 3-MCCD, on which we constructed classification models for disease screening and diagnosis using a decision tree paradigm and logistic regression analysis (LRA). For the LRA model-building process we assessed the relevance of established diagnostic flags, which have been developed from the biochemical knowledge of newborn metabolism. Both approaches yielded comparable classification accuracy in terms of sensitivity (>95.2%), while the LRA models built on flags showed significantly enhanced specificity. Both approaches yielded comparable classification accuracy in terms of sensitivity (>95.2%), while the LRA models built on flags showed significantly enhanced specificity.
14	Best, M G and Sol, N and Kooi, I and Tannous, J and Westerman, B A and Rustenburg, F and Schellen, P and Verschueren, H and Post, E and Koster, J and Ylstra, B and Ameziane, N and Dorsman, J and Smit, E F and Verheul, H M and Noske, D P and Reijneveld, J C and Nilsson, R J A and Tannous, B A and Wesseling, P and Wurdinger, T	RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics	Cancer Cell	28	5	666-676	2015	Netherlands	http://dx.doi.org/10.1016/j.ccell.2015.09.018	article	228 patients with localized and metastasized tumors and 55 healthy individuals	Case-control study	accuracy, sensitivity, specificity (leave-one-out cross-validation)	cross-validation	We determined the diagnostic potential of TEPs by mRNA sequencing of 283 platelet samples. We distinguished 228 patients with localized and metastasized tumors from 55 healthy individuals with 96% accuracy.	A leave-one-out cross-validation support vector machine algorithm (SVM/LOOCV), trained on mRNA profiles of tumor-educated blood platelets (TEPs), can distinguish patients with localized and metastasized tumors from healthy controls.
15	Bhak, Y and Jeong, H O and Cho, Y S and Jeon, S and Cho, J and Gim, J A and Jeon, Y and Blazyte, A and Park, S G and Kim, H M and Shin, E S and Paik, J W and Lee, H W and Kang, W and Kim, A and Kim, Y and Kim, B C and Ham, B J and Bhak, J and Lee, S	Depression and suicide risk prediction models using blood-derived multi-omics data	Translational Psychiatry	9		8-8	2019	United States	http://dx.doi.org/10.1038/s41398-019-0595-2	article	56 suicide attempters (SAs), 39 patients with major depressive disorder (MDD), and 87 healthy controls	Case-control study	accuracy, sensitivity, specificity, PPV, NPV (leave-one-out cross-validation)	cross-validation	We developed machine learning models to predict depression and suicide risk using blood methylome and transcriptome data. Our random forest classifiers showed accuracies of 92.6% in distinguishing SAs from MDD patients, 87.3% in distinguishing MDD patients from controls, and 86.7% in distinguishing SAs from controls.	We developed machine learning models to predict depression and suicide risk using blood methylome and transcriptome data. Our random forest classifiers showed accuracies of 92.6% in distinguishing SAs from MDD patients, 87.3% in distinguishing MDD patients from controls, and 86.7% in distinguishing SAs from controls.
16	Bhanot, G and Alexe, G and Venkataraghavan, B and Levine, A J	A robust meta-classification strategy for cancer detection from MS data	Proteomics	6	2	592-604	2006	USA	http://dx.doi.org/10.1002/pmic.200500192	article	322 serum spectra (63 with normal prostate)	Case-control study (including cases with different PSA levels)	accuracy, sensitivity, specificity (training data: 215 cases, test data: 107 cases)	training + test set	The paper presents a noise analysis and filtering procedure followed by combining the results of several machine learning tools to produce a robust predictor for MS spectra based cancer diagnosis (sensitivity of 90.31% and a speci-ficity of 98.81%.).	The paper presents a noise analysis and filtering procedure followed by combining the results of several machine learning tools to produce a robust predictor for MS spectra based cancer diagnosis.
17	Bhattacharjee, S and Singh, Y J and Ray, D	Comparative Performance Analysis of Machine Learning Classifiers on Ovarian Cancer Dataset	2017 Third Ieee International Conference on Research in Computational Intelligence and Communication Networks			213-218	2017	USA	https://doi.org/10.1109/ICRCICN.2017.8234509	meeting abstract	121 cancer samples, 95 benign	Case-control study	accuracy, sensitivity, specificity (10-fold cross-validation + external test data)	cross-validation + test set	In a comparative evaluation of machine learning methods on mass spectrometry (MS) for diagnosing benign and malignant forms of Ovarian cancer, neural networks performed better than decision trees, support vector machines, nearest neighbor classifiers and boosted trees.	In a comparative evaluation of machine learning methods on mass spectrometry (MS) for diagnosing benign and malignant forms of Ovarian cancer, neural networks performed better than decision trees, support vector machines, nearest neighbor classifiers and boosted trees.
18	Bhorade, S and Bellinger, C and Bernstein, M and Dotson, T and Feller-Kopman, D and Lee, H and Choi, Y and Pankratz, D and Lofaro, L and Walsh, P and Huang, J and Kennedy, G and Wahidi, M and Mazzone, P	IMPROVING INDETERMINANT PULMONARY NODULE MANAGEMENT WITH THE PERCEPTA GENOMIC SEQUENCING CLASSIFIER	Chest	156	4	A2271-A2272	2019		http://dx.doi.org/10.1016/j.chest.2019.08.307	meeting abstract	1600 patients training set, 412 samples validation set	Case-control study	accuracy, NPV, PPV (training/test set split)	training + test set	The authors present a clinically validated 1232 gene transcript signature from whole-transcriptome RNA sequencingt to classify the probability of malignancy for lung nodule patients with nondiagnostic bronchoscopy. The tool is reported to achieve a combined 95%+ sensitivity for low/intermediate pre-test risk groups.	The authors present a clinically validated 1232 gene transcript signature from whole-transcriptome RNA sequencingt to classify the probability of malignancy for lung nodule patients with nondiagnostic bronchoscopy. The tool is reported to achieve a combined 95%+ sensitivity for low/intermediate pre-test risk groups.
19	Bochare, A and Gangopadhyay, A and Yesha, Y and Joshi, A and Yesha, Y and Brady, M and Grasso, M A and Rishe, N	Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer	International Journal of Medical Engineering and Informatics	6	2	87-99	2014		http://dx.doi.org/10.1504/IJMEI.2014.060245	article	1145 cases and 1142 controls	Case-control study	accuracy, sensitivity, specificity (10-fold cross-validation)	cross-validation	The authors combined genomic data with data on family history and age to predict the risk of developing breast cancer for postmenopausal women of European descent. The machine learning model generated using both prior domain knowledge and feature selection performed better compared to a conventional classification approach.	The authors combined genomic data with data on family history and age to predict the risk of developing breast cancer for postmenopausal women of European descent. The machine learning model generated using both prior domain knowledge and feature selection performed better compared to a conventional classification approach.
20	Bom, M J and Levin, E and Driessen, R S and Danad, I and Van Kuijk, C C and van Rossum, A C and Narula, J and Min, J K and Leipsic, J A and Belo Pereira, J P and Taylor, C A and Nieuwdorp, M and Raijmakers, P G and Koenig, W and Groen, A K and Stroes, E S G and Knaapen, P	Predictive value of targeted proteomics for coronary plaque morphology in patients with suspected coronary artery disease	EBioMedicine	39		109-117	2019	Canada	http://dx.doi.org/10.1016/j.ebiom.2018.12.033	article	196 patients with suspected coronary artery disease	Case-control study	AUC (10-fold CV + 20% hold-out test set)	cross-validation + test set	The authors investigated the ability of targeted proteomics to predict presence of high-risk plaque or absence of coronary atherosclerosis in patients with suspected CAD. The developed machine learning model had fair diagnostic performance with an area under the curve (AUC) of 0·79 ± 0·01, outperforming prediction with generally available clinical characteristics (AUC = 0·65 ± 0·04, p < 0·05). Conversely, a different subset of 34 proteins was predictive for the absence of CAD (AUC = 0·85 ± 0·05), again outperforming prediction with generally available characteristics (AUC = 0·70 ± 0·04, p < 0·05).	The authors investigated the ability of targeted proteomics to predict presence of high-risk plaque or absence of coronary atherosclerosis in patients with suspected CAD. The developed machine learning model had fair diagnostic performance with an area under the curve (AUC) of 0·79 ± 0·01, outperforming prediction with generally available clinical characteristics (AUC = 0·65 ± 0·04, p < 0·05). Conversely, a different subset of 34 proteins was predictive for the absence of CAD (AUC = 0·85 ± 0·05), again outperforming prediction with generally available characteristics (AUC = 0·70 ± 0·04, p < 0·05).
21	Bomane, A and Gonçalves, A and Ballester, P J	Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data	Frontiers in Genetics	10			2019	France	http://dx.doi.org/10.3389/fgene.2019.01041	article	1,098 patients	Cases only (treatment response prediction)	AUC (LOOCV)	cross-validation	The authors aimed to predicted Breast Cancer (BC) patient response to the drug paclitaxel using data from the US National Cancer Institute’s Genomic Data Commons, DNA methylation and miRNA profile data. ML algorithms selecting the smallest subset of molecular features generated the most predictive classifiers: a complexity-optimized XGBoost classifier based on CpG island methylation extracted a subset of molecular factors relevant to predict paclitaxel response (AUC = 0.74). A CpG site methylation-based Decision Tree (DT) combining only 2 of the 22,941 considered CpG sites (AUC = 0.89) and a miRNA expression-based DT employing just 4 of the 337 analyzed mature miRNAs (AUC = 0.72) reveal the molecular types associated to paclitaxel-sensitive and resistant BC tumors.	The authors aimed to predicted Breast Cancer (BC) patient response to the drug paclitaxel using data from the US National Cancer Institute’s Genomic Data Commons, DNA methylation and miRNA profile data. ML algorithms selecting the smallest subset of molecular features generated the most predictive classifiers: a complexity-optimized XGBoost classifier based on CpG island methylation extracted a subset of molecular factors relevant to predict paclitaxel response (AUC = 0.74). A CpG site methylation-based Decision Tree (DT) combining only 2 of the 22,941 considered CpG sites (AUC = 0.89) and a miRNA expression-based DT employing just 4 of the 337 analyzed mature miRNAs (AUC = 0.72) reveal the molecular types associated to paclitaxel-sensitive and resistant BC tumors.
22	Bovelstad, H M and Nygard, S and Storvold, H L and Aldrin, M and Borgan, O and Frigessi, A and Lingjaerde, O C	Predicting survival from microarray data--a comparative study	Bioinformatics	23	16	2080-2087	2007	Norway	https://doi.org/10.1093/bioinformatics/btm305	article	Benchmark comparison across several cancer datasets with > 50 samples per condition	Case-control study	Log rank test p-value (10-fold cross-validation)	cross-validation	Statistical learning from subsets should be repeated several times in order to get a fair comparison between methods. Methods using coefficient shrinkage or linear combinations of the gene expression values have much better performance than the simple variable selection methods. For our data sets, ridge regression has the overall best performance.	Statistical learning from subsets should be repeated several times in order to get a fair comparison between methods. Methods using coefficient shrinkage or linear combinations of the gene expression values have much better performance than the simple variable selection methods. For our data sets, ridge regression has the overall best performance.
23	Brooks, J and Modos, D and Sudhakar, P and Fazekas, D and Zoufir, A and Watson, A and Tremelling, M and Verstockt, B and Vermeire, S and Bender, A and Carding, S and Korcsmaros, T	Systems genomics of ulcerative colitis: Combining GWAS and signalling networks for patient stratification and individualised drug targeting in ulcerative colitis	Journal of Crohn's and Colitis	13		S006-S007	2019		http://dx.doi.org/10.1093/ecco-jcc/jjy222.009	meeting abstract	377 UC patients	Cases only (clustering using genomic footprint)	"We validated the workflow on a larger cohort of 941 UC patients from the IBD Biobank in Leuven"	external cohort validation	"Analysing the genomic footprints of the patients, we identified 4 patient clusters, and identified common and differing pathogenic pathways between them. We showed that clusters were related to gender and age of onset of disease, but unrelated to therapeutic upscaling of therapy. With machine learning, we identified a subset of patients from within one of the cohorts, for whom the presence of a regulatory MAML2 SNP was a marker for therapeutic upscaling."	

24	Brown, K K and Choi, Y and Colby, T V and Flaherty, K R and Groshong, S and Imtiaz, U and Lynch, D A and Myers, J L and Steele, M P and Martinez, F J and Pankratz, D G and Walsh, P S and Huang, J and Barth, N M and Raghu, G and Kennedy, G C	Prospective validation of a genomic classifier for usual interstitial pneumonia in transbronchial biopsies	American Journal of Respiratory and Critical Care Medicine	195			2017		https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2017.195.1_MeetingAbstracts.A6792	meeting abstract	354 TBB samples	Case-control study	AUC (training / test set split)	training + test set	A definitive diagnosis of idiopathic pulmonary fibrosis (IPF) requires the presence of a usual interstitial pneumonia (UIP) pattern on chest imaging or surgical lung biopsy (SLB). A genomic classifier based on the gene expression pattern found in tissue obtained by less-invasive transbronchial biopsy (TBB) was evaluated and distinguished UIP from non-UIP in the training set with a receiver-operator characteristic area under the curve (AUC) = 0.87 [CI 0.83-0.91] (specificity = 91% [CI 85-95], sensitivity = 71% [CI 65-77]). In validation, the test achieved an AUC of 0.85 [CI 0.73-0.97], with specificity = 88% [CI 68-97%] and sensitivity = 67% [CI 45-84%].	A definitive diagnosis of idiopathic pulmonary fibrosis (IPF) requires the presence of a usual interstitial pneumonia (UIP) pattern on chest imaging or surgical lung biopsy (SLB). A genomic classifier based on the gene expression pattern found in tissue obtained by less-invasive transbronchial biopsy (TBB) was evaluated and distinguished UIP from non-UIP in the training set with a receiver-operator characteristic area under the curve (AUC) = 0.87 [CI 0.83-0.91] (specificity = 91% [CI 85-95], sensitivity = 71% [CI 65-77]). In validation, the test achieved an AUC of 0.85 [CI 0.73-0.97], with specificity = 88% [CI 68-97%] and sensitivity = 67% [CI 45-84%].
25	Cai, Q and Alvarez, J A and Kang, J and Yu, T	Network Marker Selection for Untargeted LC-MS Metabolomics Data	J Proteome Res	16	3	1261-1269	2017	United States	https://doi.org/10.1021/acs.jproteome.6b00861	article	subjects with available high-resolution plasma metabolomics from the Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being (N = 371)	Cases only (BMI analysis)	AUC (5-fold CV)	cross-validation	In this paper, we propose a flexible network feature selection framework that combines metabolomics data with the genome-scale metabolic network. The method adopts a sequential feature screening procedure and machine learning-based criteria to select important subnetworks and identify the optimal feature matching simultaneously. Simulation studies show that the proposed method has a much higher sensitivity than the commonly used maximal matching approach. For demonstration, we apply the method on a cohort of healthy subjects to detect subnetworks associated with the body mass index (BMI). The method identifies several subnetworks that are supported by the current literature, as well as detects some subnetworks with plausible new functional implications.	In this paper, we propose a flexible network feature selection framework that combines metabolomics data with the genome-scale metabolic network. The method adopts a sequential feature screening procedure and machine learning-based criteria to select important subnetworks and identify the optimal feature matching simultaneously. Simulation studies show that the proposed method has a much higher sensitivity than the commonly used maximal matching approach. For demonstration, we apply the method on a cohort of healthy subjects to detect subnetworks associated with the body mass index (BMI). The method identifies several subnetworks that are supported by the current literature, as well as detects some subnetworks with plausible new functional implications.
26	Cai, Z and Xu, D and Zhang, Q and Zhang, J and Ngai, S M and Shao, J	Classification of lung cancer using ensemble-based feature selection and machine learning methods	Mol Biosyst	11	3	791-800	2015	China	https://doi.org/10.1039/c4mb00659c	article	More than 100 samples available for the main sample groups LADC and SQCLC in both training and test set	Case-control study	accuracy, precision, recall, F-score (LOOCV)	cross-validation	"There are three major types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC is further classified into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer. In the present study, ROC (Receiving Operating Curve), RFs (Random Forests) and mRMR (Maximum Relevancy and Minimum Redundancy) were proposed to capture the unbiased, informative as well as compact molecular signatures followed by machine learning methods to classify LADC, SQCLC and SCLC. As a result, a panel of 16 DNA methylation markers exhibits an ideal classification power with an accuracy of 86.54%, 84.6% and a recall 84.37%, 85.5% in the leave-one-out cross-validation (LOOCV) and independent data set test experiments, respectively."	"There are three major types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC is further classified into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer. In the present study, ROC (Receiving Operating Curve), RFs (Random Forests) and mRMR (Maximum Relevancy and Minimum Redundancy) were proposed to capture the unbiased, informative as well as compact molecular signatures followed by machine learning methods to classify LADC, SQCLC and SCLC. As a result, a panel of 16 DNA methylation markers exhibits an ideal classification power with an accuracy of 86.54%, 84.6% and a recall 84.37%, 85.5% in the leave-one-out cross-validation (LOOCV) and independent data set test experiments, respectively."
27	Casanova, R and Varma, S and Simpson, B and Kim, M and An, Y and Saldana, S and Riveros, C and Moscato, P and Griswold, M and Sonntag, D and Wahrheit, J and Klavins, K and Jonsson, P V and Eiriksdottir, G and Aspelund, T and Launer, L J and Gudnason, V and Legido Quigley, C and Thambisetty, M	Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals	Alzheimers Dement	12	7	815-822	2016	Australia	http://dx.doi.org/10.1016/j.jalz.2015.12.008	article	two cohorts of n=93 and n=100 samples	Case-control study	AUC, sensitivity, specificity (6-fold CV)	cross-validation	"Recently, quantitative metabolomics identified a panel of 10 plasma lipids that were highly pre- dictive of conversion to Alzheimer's disease (AD) in cognitively normal older individuals. We failed to replicate these findings in a substantially larger study from two independent cohorts. These findings underscore the importance of large-scale independent validation of index findings from biomarker studies with relatively small sample sizes."	"Recently, quantitative metabolomics identified a panel of 10 plasma lipids that were highly pre- dictive of conversion to Alzheimer's disease (AD) in cognitively normal older individuals. We failed to replicate these findings in a substantially larger study from two independent cohorts. These findings underscore the importance of large-scale independent validation of index findings from biomarker studies with relatively small sample sizes."
28	Chaiboonchoe, A and Samarasinghe, S and Kulasiri, D	Machine Learning for Childhood Acute Lymphoblastic Leukaemia Gene Expression Data Analysis: A Review	Current Bioinformatics	5	2	118-133	2010		https://www.eurekaselect.com/node/86215/article/-machine-learning-for-childhood-acute-lymphoblastic-leukaemia-gene-expression-data-analysis-a-review	article	review (not applicable)				"We present a comprehensive review of machine learning approaches that have been used on [acute lymphoblastic leukaemia] (ALL) microarray data. These methods have been used in four major areas of microarray data analysis: gene selection, clustering, classification and pathway analysis. Each machine learning algorithm has its own advantages and drawbacks."	"We present a comprehensive review of machine learning approaches that have been used on [acute lymphoblastic leukaemia] (ALL) microarray data. These methods have been used in four major areas of microarray data analysis: gene selection, clustering, classification and pathway analysis. Each machine learning algorithm has its own advantages and drawbacks."
29	Chang, Y and Park, H and Yang, H J and Lee, S and Lee, K Y and Kim, T S and Jung, J and Shin, J M	Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature	Sci Rep	8	1	8857-8857	2018	Australia	http://dx.doi.org/10.1038/s41598-018-27214-6	article	787 human cancer cell lines and structural profiles of 244 drugs were considered	Cases only (drug response study)	Rsquared, AUC (training/test split)	training + test set	"We report Cancer Drug Response profile scan (CDRscan) a novel deep learning model that predicts anticancer drug responsiveness based on a large-scale drug screening assay data encompassing genomic profiles of 787 human cancer cell lines and structural profiles of 244 drugs. We applied CDRscan to 1,487 approved drugs and identified 14 oncology and 23 non-oncology drugs having new potential cancer indications."	"We report Cancer Drug Response profile scan (CDRscan) a novel deep learning model that predicts anticancer drug responsiveness based on a large-scale drug screening assay data encompassing genomic profiles of 787 human cancer cell lines and structural profiles of 244 drugs. We applied CDRscan to 1,487 approved drugs and identified 14 oncology and 23 non-oncology drugs having new potential cancer indications."
30	Chao, S M and Connolly, J and Ng, Y H and Ganesan, I and Bernett, L	Can urinary proteomes be used as non-invasive markers for renal involvement in childhood febrile urinary tract infection (UTI)?	Pediatric Nephrology	31	10	1746-1746	2016		http://dx.doi.org/10.1007/s00467-016-3466-6	meeting abstract	121 patients (68 males, 53 females)	Case-control study	sensitivity, PPV (10-fold CV)	cross-validation	"To investigate if urinary proteomes be used as non-invasive markers for renal involvement in childhood febrile urinary tract infection (UTI), decision trees as classifiers were constructed and were able to predict correctly after 10 fold cross-validation, 69.23% patients' status as no renal scarring. RS (82.5% sensitivity, 78.6% PPV). However only 53.98% patients' status as no acute pyelonephritis (APN) were correctly predicted (47.1% sensitivity, 49.0% PPV). Larger cohort studies are needed to test the validity and reproducibility of these biomarkers."	"To investigate if urinary proteomes be used as non-invasive markers for renal involvement in childhood febrile urinary tract infection (UTI), decision trees as classifiers were constructed and were able to predict correctly after 10 fold cross-validation, 69.23% patients' status as no renal scarring. RS (82.5% sensitivity, 78.6% PPV). However only 53.98% patients' status as no acute pyelonephritis (APN) were correctly predicted (47.1% sensitivity, 49.0% PPV). Larger cohort studies are needed to test the validity and reproducibility of these biomarkers."
31	Chaudhary, K and Poirion, O B and Lu, L and Garmire, L X	Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer	Clin Cancer Res	24	6	1248-1259	2018	United States	https://doi.org/10.1158/1078-0432.ccr-17-0853	article	360 patients included	Cases only (survival prediction)	"We validated this multi-omics model on five external datasets of various omics types: LIRI-JP cohort (n = 230, C-index = 0.75), NCI cohort (n = 221, C-index = 0.67), Chinese cohort (n = 166, C-index = 0.69), E-TABM-36 cohort (n = 40, C-index = 0.77), and Hawaiian cohort (n = 27, C-index = 0.82). "	external cohort validation	"Identifying robust survival subgroups of hepatocellular carcinoma (HCC) will significantly improve patient care. We built the DL-based, survival-sensitive model on 360 HCC patients' data using RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), and methylation data from The Cancer Genome Atlas (TCGA) [...]. This DL-based model provides two optimal subgroups of patients with significant survival differences (P = 7.13e-6) and good model fitness [concordance index (C-index) = 0.68]."	"Identifying robust survival subgroups of hepatocellular carcinoma (HCC) will significantly improve patient care. We built the DL-based, survival-sensitive model on 360 HCC patients' data using RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), and methylation data from The Cancer Genome Atlas (TCGA) [...]. This DL-based model provides two optimal subgroups of patients with significant survival differences (P = 7.13e-6) and good model fitness [concordance index (C-index) = 0.68]."
32	Choung, R S and Khaleghi Rostamkolaei, S and Ju, J M and Marietta, E V and Van Dyke, C T and Rajasekaran, J J and Jayaraman, V and Wang, T and Bei, K and Rajasekaran, K E and Krishna, K and Krishnamurthy, H K and Murray, J A	Synthetic Neoepitopes of the Transglutaminase-Deamidated Gliadin Complex as Biomarkers for Diagnosing and Monitoring Celiac Disease	Gastroenterology	156	3	582-591.e1	2019	United States	http://dx.doi.org/10.1053/j.gastro.2018.10.025	article	serum samples from 90 patients with biopsy-proven Celiac disease and 79 healthy individuals (controls)	Case-control study	AUC, accuracy, sensitivity, specificity ("We validated out findings in 82 patients with newly diagnosed CeD and 217 controls.")	external cohort validation	"We aimed to discover biomarkers of CeD derived from neoepitopes of deami- dated gliadin peptides (DGP) and tTG fragments and to determine if immune reactivity against these epitopes can identify patients with CeD with mucosal healing. A fluorescent peptide microarray platform was used to estimate the antibody-binding intensity of each synthesized tTG-DGP epitope. In the 101 training cohort, the set of neoepitopes derived from the tTG-DGP 102 complex identified patients with CeD with 99% sensitivity and 103 100% specificity. The assay identified patients with mucosa healing status 110 with 84% sensitivity and 95% specificity."	"We aimed to discover biomarkers of CeD derived from neoepitopes of deami- dated gliadin peptides (DGP) and tTG fragments and to determine if immune reactivity against these epitopes can identify patients with CeD with mucosal healing. A fluorescent peptide microarray platform was used to estimate the antibody-binding intensity of each synthesized tTG-DGP epitope. In the 101 training cohort, the set of neoepitopes derived from the tTG-DGP 102 complex identified patients with CeD with 99% sensitivity and 103 100% specificity. The assay identified patients with mucosa healing status 110 with 84% sensitivity and 95% specificity."
33	Chung, W Y and Correa, E and Yoshimura, K and Chang, M C and Dennison, A and Takeda, S and Chang, Y T	Using probe electrospray ionization mass spectrometry and machine learning for detecting pancreatic cancer with high performance	American Journal of Translational Research	12	1	171-179	2020	Japan	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7013221/	article	322 PDAC patients and 265 controls	Case-control study	accuracy, sensitivity, specificity (1000 independent repetitions of a bootstrap cross-validation process)	cross-validation	"A rapid blood-based diagnostic modality to detect pancreatic ductal adenocarcinoma (PDAC) with high accuracy is an unmet medical need. The study aimed to validate a unique diagnosis system using Probe Electrospray Ionization Mass Spectrometry (PESI-MS) and Machine Learning to the diagnosis of PDAC. The sensitivity of the machine learning algorithm using PESI-MS profiles to identify PDAC is 90.8% with specificity of 91.7% (95% CI 83.9%-97.4% and 82.8%-97.7% respectively). Combined PESI-MS profiles with age and CA19-9 as predictors, the accuracy for stage 1 or 2 of PDAC is 92.9% and for stage 3 or 4 is 93% (95% CI 86.3-98.2; 87.9-97.4 respectively). The accuracy and simplicity of the PESI-MS profiles combined with machine learning provide an opportunity to detect PDAC at an early stage and must be applicable to the examination of at-risk populations."	"A rapid blood-based diagnostic modality to detect pancreatic ductal adenocarcinoma (PDAC) with high accuracy is an unmet medical need. The study aimed to validate a unique diagnosis system using Probe Electrospray Ionization Mass Spectrometry (PESI-MS) and Machine Learning to the diagnosis of PDAC. The sensitivity of the machine learning algorithm using PESI-MS profiles to identify PDAC is 90.8% with specificity of 91.7% (95% CI 83.9%-97.4% and 82.8%-97.7% respectively). Combined PESI-MS profiles with age and CA19-9 as predictors, the accuracy for stage 1 or 2 of PDAC is 92.9% and for stage 3 or 4 is 93% (95% CI 86.3-98.2; 87.9-97.4 respectively). The accuracy and simplicity of the PESI-MS profiles combined with machine learning provide an opportunity to detect PDAC at an early stage and must be applicable to the examination of at-risk populations."
34	Clark, O and Safikhani, Z and Smirnov, P and Haibe-Kains, B	Gene isoforms as expression-based biomarkers predictive of drug response in vitro	Irish Journal of Medical Science	187		S348-S348	2018	Canada	https://www.nature.com/articles/s41467-017-01153-8	article	The data comprised 79,903 experiments for 140 different drugs tested on a panel of up to 778 unique cell lines from 30 tissue types	Cases only (drug response prediction in-vitro)	AUC, accuracy (validation in independent breast cancer data and different pharmacological assay)	external cohort validation	"To identify robust transcriptomic biomarkers for drug response across studies, we develop a meta-analytical framework combining the pharmacological data from two large-scale drug screening datasets. We use an independent pan-cancer pharmacogenomic dataset to test the robustness of our candidate biomarkers across multiple cancer types. We further analyze two independent breast cancer datasets and find that specific isoforms of IGF2BP2, NECTIN4, ITGB6, and KLHDC9 are significantly associated with AZD6244, lapatinib, erlotinib, and paclitaxel, respectively. "	"To identify robust transcriptomic biomarkers for drug response across studies, we develop a meta-analytical framework combining the pharmacological data from two large-scale drug screening datasets. We use an independent pan-cancer pharmacogenomic dataset to test the robustness of our candidate biomarkers across multiple cancer types. We further analyze two independent breast cancer datasets and find that specific isoforms of IGF2BP2, NECTIN4, ITGB6, and KLHDC9 are significantly associated with AZD6244, lapatinib, erlotinib, and paclitaxel, respectively. "
35	Croner, L J and Kao, A and Benz, R and Blume, J E and Dillon, R and Wilcox, B and Kairs, S N	A new blood test for colorectal cancer in high-risk subjects	Clinical Chemistry	63		S22-S23	2017	Denmark	http://dx.doi.org/10.11613/bm.2020.030504	meeting abstract	4,435 patient samples (3,066 patients (340 CRC and 2,759 non-CRC) were randomly assigned to the classifier discovery set. The remaining 1,336 samples (147 CRC and 1,189 non-CRC) were assigned to the validation set)	Case-control study	sensitivity, specificity, PPV, NPC (10-fold CV + training / test set split)	cross-validation + test set	"The objective was to develop a blood-based colorectal cancer (CRC) test with clinically useful performance in patients with CRC symptoms. Machine learning was used to build and test candidate classifiers. The final classifier was a logistic regression using 10 predictors : eight proteins, age, and gender. In validation, the indeterminate rate was 23.2%, sensitivity/specificity was 0.80/0.83, the PPV was 36.5%, and the NPV was 97.1%. This performance compares favorably to that from other CRC blood tests."	"The objective was to develop a blood-based colorectal cancer (CRC) test with clinically useful performance in patients with CRC symptoms. Machine learning was used to build and test candidate classifiers. The final classifier was a logistic regression using 10 predictors : eight proteins, age, and gender. In validation, the indeterminate rate was 23.2%, sensitivity/specificity was 0.80/0.83, the PPV was 36.5%, and the NPV was 97.1%. This performance compares favorably to that from other CRC blood tests."
36	Cruz, J A and Wishart, D S	Applications of machine learning in cancer prediction and prognosis	Cancer Informatics	2		59-77	2006	Greece	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675494/	article	review (not applicable)	Case-control study	"In assembling this review we conducted a broad survey of the different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer prediction and prognosis. A number of trends are noted, including a growing dependence on protein biomarkers and microarray data, a strong bias towards applications in prostate and breast cancer, and a heavy reliance on "older" technologies such artificial neural networks (ANNs) instead of more recently developed or more easily interpretable machine learning methods. A number of published studies also appear to lack an appropriate level of validation or testing. Among the better designed and validated studies it is clear that machine learning methods can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality."		"In assembling this review we conducted a broad survey of the different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer prediction and prognosis. A number of trends are noted, including a growing dependence on protein biomarkers and microarray data, a strong bias towards applications in prostate and breast cancer, and a heavy reliance on "older" technologies such artificial neural networks (ANNs) instead of more recently developed or more easily interpretable machine learning methods. A number of published studies also appear to lack an appropriate level of validation or testing. Among the better designed and validated studies it is clear that machine learning methods can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality."	
37	Cugliari, G and Benevenuta, S and Guarrera, S and Sacerdote, C and Panico, S and Krogh, V and Tumino, R and Vineis, P and Fariselli, P and Matullo, G	Improving the prediction of cardiovascular risk with machine-learning and DNA methylation data	2019 16th Ieee International Conference on Computational Intelligence in Bioinformatics and Computational Biology - Cibcb 2019			39-42	2019	USA	https://ieeexplore.ieee.org/document/8791483	article	584 subjects (292 MI cases and 292 matched controls)	Case-control study	AUC, sensitivity, specificity (nested-cross validation)	cross-validation	"Classically, the cardiovascular risk of individual is evaluated using phenomenological variables (PV) such as blood pressure, body mass, smoker status, gender, age etc. Here we show that, on prospective study (after 10-15 years) these PV display a poor agreement with case-control samples. We were able to obtain more accurate predictions using both DNA methylation data and PV as input features of a Random Forest model, achieving a ROC-AUC of 0.74. "	"Classically, the cardiovascular risk of individual is evaluated using phenomenological variables (PV) such as blood pressure, body mass, smoker status, gender, age etc. Here we show that, on prospective study (after 10-15 years) these PV display a poor agreement with case-control samples. We were able to obtain more accurate predictions using both DNA methylation data and PV as input features of a Random Forest model, achieving a ROC-AUC of 0.74. "

38	Das, D and Ito, J and Kadowaki, T and Tsuda, K	An interpretable machine learning model for diagnosis of Alzheimer's disease	PeerJ	2019	3		2019	Japan	http://dx.doi.org/10.7717/peerj.6543	article	97 AD subjects + 54 controls	Case-control study	AUC, accuracy, sensitivity, specificity (cross-validation + test set)	cross-validation + test set	"We present an interpretable machine learning model for medical diagnosis called sparse high-order interaction model with rejection option (SHIMR). A decision tree explains to a patient the diagnosis with a long rule (i.e., conjunction of many intervals), while SHIMR employs a weighted sum of short rules. Using proteomics data of 151 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, SHIMR is shown to be as accurate as other non-interpretable methods (Sensitivity, SN = 0.84 ± 0.1, Specificity, SP = 0.69 ± 0.15 and Area Under the Curve, AUC = 0.86 ± 0.09)."	"We present an interpretable machine learning model for medical diagnosis called sparse high-order interaction model with rejection option (SHIMR). A decision tree explains to a patient the diagnosis with a long rule (i.e., conjunction of many intervals), while SHIMR employs a weighted sum of short rules. Using proteomics data of 151 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, SHIMR is shown to be as accurate as other non-interpretable methods (Sensitivity, SN = 0.84 ± 0.1, Specificity, SP = 0.69 ± 0.15 and Area Under the Curve, AUC = 0.86 ± 0.09)."
39	de Maturana, E L and Alonso, L and Alarcón, P and Martín-Antoniano, I A and Pineda, S and Piorno, L and Calle, M L and Malats, N	Challenges in the integration of omics and non-omics data	Genes	10	3		2019	Spain	http://dx.doi.org/10.3390/genes10030238	article	review (not applicable)	Review			"Only a small number of published studies performed a "real" integration of omics and non-omics (OnO) data, mainly to predict cancer outcomes. Challenges in OnO data integration regard the nature and heterogeneity of non-omics data, the possibility of integrating large-scale non-omics data with high-throughput omics data, the relationship between OnO data (i.e., ascertainment bias), the presence of interactions, the fairness of the models, and the presence of subphenotypes. These challenges demand the development and application of new analysis strategies to integrate OnO data. In this contribution we discuss different attempts of OnO data integration in clinical and epidemiological studies. The integrative strategies used in the identified papers adopted three modeling methods: Independent, conditional, and joint modeling."	"Only a small number of published studies performed a "real" integration of omics and non-omics (OnO) data, mainly to predict cancer outcomes. Challenges in OnO data integration regard the nature and heterogeneity of non-omics data, the possibility of integrating large-scale non-omics data with high-throughput omics data, the relationship between OnO data (i.e., ascertainment bias), the presence of interactions, the fairness of the models, and the presence of subphenotypes. These challenges demand the development and application of new analysis strategies to integrate OnO data. In this contribution we discuss different attempts of OnO data integration in clinical and epidemiological studies. The integrative strategies used in the identified papers adopted three modeling methods: Independent, conditional, and joint modeling."
40	de Ronde, J J and Bonder, M J and Lips, E H and Rodenhuis, S and Wessels, L F	Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes	PLoS One	9	2	e88551-e88551	2014		http://dx.doi.org/10.1371/journal.pone.0088551	article	374 samples were analyzed	Cases only (treatment response prediction)	AUC (nested cross-validation)	cross-validation	"We set out to study if gene expression based predictors of chemotherapy resistance that are specific for breast cancer subtypes can improve upon the performance of generic predictors. For HER2+, ER− breast cancer, subtype specific predictor based on clinical features outperformed the generic, non-specific predictor. This can be explained by the fact that the generic predictor included HER2 and ER status, features that are predictive over the whole set, but not within this subtype. In all other scenarios the generic predictors outperformed the subtype specific predictors or showed equal performance."	"We set out to study if gene expression based predictors of chemotherapy resistance that are specific for breast cancer subtypes can improve upon the performance of generic predictors. For HER2+, ER− breast cancer, subtype specific predictor based on clinical features outperformed the generic, non-specific predictor. This can be explained by the fact that the generic predictor included HER2 and ER status, features that are predictive over the whole set, but not within this subtype. In all other scenarios the generic predictors outperformed the subtype specific predictors or showed equal performance."
41	Di Camillo, B and Sanavia, T and Martini, M and Jurman, G and Sambo, F and Barla, A and Squillario, M and Furlanello, C and Toffolo, G and Cobelli, C	Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment	PLoS One	7	3	e32200-e32200	2012	Italy	http://dx.doi.org/10.1371/journal.pone.0032200	article	3 different datasets (more than 50 samples per group in total)	Case-control study	AUC, sensitivity, specificity (cross validation + Monte Carlo bootstrap resampling)	cross-validation	"The identification of robust lists of molecular biomarkers related to a disease is a fundamental step for early diagnosis and treatment. However, methodologies for the discovery of biomarkers using microarray data often provide results with limited overlap. These differences are imputable to 1) dataset size (few subjects with respect to the number of features); 2) heterogeneity of the disease; 3) heterogeneity of experimental protocols and computational pipelines employed in the analysis. In this paper, we focus on the first two issues and assess, both on simulated (through an in silico regulation network model) and real clinical datasets, the consistency of candidate biomarkers provided by a number of different methods. The simulated data allowed us to outline advantages and drawbacks of different methods across multiple studies and varying number of samples and to evaluate precision of feature selection on a benchmark with known biomarkers. Although comparable classification accuracy was reached by different methods, the use of external cross-validation loops is helpful in finding features with a higher degree of precision and stability."	"The identification of robust lists of molecular biomarkers related to a disease is a fundamental step for early diagnosis and treatment. However, methodologies for the discovery of biomarkers using microarray data often provide results with limited overlap. These differences are imputable to 1) dataset size (few subjects with respect to the number of features); 2) heterogeneity of the disease; 3) heterogeneity of experimental protocols and computational pipelines employed in the analysis. In this paper, we focus on the first two issues and assess, both on simulated (through an in silico regulation network model) and real clinical datasets, the consistency of candidate biomarkers provided by a number of different methods. The simulated data allowed us to outline advantages and drawbacks of different methods across multiple studies and varying number of samples and to evaluate precision of feature selection on a benchmark with known biomarkers. Although comparable classification accuracy was reached by different methods, the use of external cross-validation loops is helpful in finding features with a higher degree of precision and stability."
42	Diaz-Cano, S and Sutherland, R and Moorhead, J and Blanes, A and Dobson, R	Growth pattern analysis in low grade clear cell renal cell carcinomas: Prognostic value and biologic significance	Laboratory Investigation	96		226A-226A	2016		http://dx.doi.org/10.1038/labinvest.2016.10	meeting abstract	low FG (1-2, 174 cases) vs. high FG (3-4, 139 cases) grade	Subtype comparison	AUC (50-fold cross-validation)	cross-validation	"The Pan-Cancer Analysis Project aimed to identify the genomic changes in cancer types from the Cancer Genome Atlas (TCGA). The meaning of architectural features in clear cell renal cell carcinomas (ccRCC) by Fuhrman grade (FG) has not been investigated at clinic-pathologic or genetic levels in this set. Clinical data were also collected (gender, age, and stage). We used a Random Forest machine learning approach comparing low FG (1-2, 174 cases) vs. high FG (3-4, 139 cases) grade. The age, gender, proteins and variants model performed with an AUC of 0.803."	"The Pan-Cancer Analysis Project aimed to identify the genomic changes in cancer types from the Cancer Genome Atlas (TCGA). The meaning of architectural features in clear cell renal cell carcinomas (ccRCC) by Fuhrman grade (FG) has not been investigated at clinic-pathologic or genetic levels in this set. Clinical data were also collected (gender, age, and stage). We used a Random Forest machine learning approach comparing low FG (1-2, 174 cases) vs. high FG (3-4, 139 cases) grade. The age, gender, proteins and variants model performed with an AUC of 0.803."
43	Diggans, J and Kim, S Y and Hu, Z Z and Pankratz, D and Wong, M and Reynolds, J and Tom, E and Pagan, M and Monroe, R and Rosai, J and Livolsi, V A and Lanman, R B and Kloos, R T and Walsh, P S and Kennedy, G C	MACHINE LEARNING FROM CONCEPT TO CLINIC: RELIABLE DETECTION OF BRAF V600E DNA MUTATIONS IN THYROID NODULES USING HIGH-DIMENSIONAL RNA EXPRESSION DATA	Pacific Symposium on Biocomputing 2015			371-382	2015	USA	https://pubmed.ncbi.nlm.nih.gov/25592597/	article	training (n=181) and independent test (n=535) sets	Case-control study	AUC (10-fold CV + external test set)	cross-validation + test set	Here we report the development, clinical validation, and diagnostic accuracy of a pre-operative molecular test (Afirma BRAF) to identify BRAF V600E mutations using mRNA expression in thyroid fine needle aspirate biopsies (FNABs). The resulting 128-gene linear support vector machine was compared to qPCR in the independent test set. Clinical sensitivity and specificity for malignancy were evaluated in a subset of test set samples (n=213) with expert-derived histopathology. We observed high positive- (PPA, 90.4%) and negative (NPA, 99.0%) percent agreement with qPCR on the test set. Clinical sensitivity for malignancy was 43.8% (consistent with published prevalence of BRAF V600E in this neoplasm) and specificity was 100%, identical to qPCR on the same samples."	Here we report the development, clinical validation, and diagnostic accuracy of a pre-operative molecular test (Afirma BRAF) to identify BRAF V600E mutations using mRNA expression in thyroid fine needle aspirate biopsies (FNABs). The resulting 128-gene linear support vector machine was compared to qPCR in the independent test set. Clinical sensitivity and specificity for malignancy were evaluated in a subset of test set samples (n=213) with expert-derived histopathology. We observed high positive- (PPA, 90.4%) and negative (NPA, 99.0%) percent agreement with qPCR on the test set. Clinical sensitivity for malignancy was 43.8% (consistent with published prevalence of BRAF V600E in this neoplasm) and specificity was 100%, identical to qPCR on the same samples."
44	Ding, M Q and Chen, L and Cooper, G F and Young, J D and Lu, X	Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics	Mol Cancer Res	16	2	269-278	2018	United States	http://dx.doi.org/10.1158/1541-7786.mcr-17-0378	article	transcriptomics data from 727 cell lines was used	Cases only (cancer cell line drug response prediction)	accuracy, sensitivity, specificity (25-fold cross-validation)	cross-validation	"In this study, machine learning methods (e.g., deep learning) were used to identify informative features from genome-scale omics data and to train classifiers for predicting the effectiveness of drugs in cancer cell lines. The methodology introduced here can accurately predict the efficacy of drugs, regardless of whether they are molecularly targeted or nonspecific chemotherapy drugs. This approach, on a per-drug basis, can identify sensitive cancer cells with an average sensitivity of 0.82 and specificity of 0.82; on a per-cell line basis, it can identify effective drugs with an average sensitivity of 0.80 and specificity of 0.82."	"In this study, machine learning methods (e.g., deep learning) were used to identify informative features from genome-scale omics data and to train classifiers for predicting the effectiveness of drugs in cancer cell lines. The methodology introduced here can accurately predict the efficacy of drugs, regardless of whether they are molecularly targeted or nonspecific chemotherapy drugs. This approach, on a per-drug basis, can identify sensitive cancer cells with an average sensitivity of 0.82 and specificity of 0.82; on a per-cell line basis, it can identify effective drugs with an average sensitivity of 0.80 and specificity of 0.82."
45	Djebbari, A and Labbe, A	Refining gene signatures: a Bayesian approach	BMC Bioinformatics	10		410-410	2009	Canada	http://dx.doi.org/10.1186/1471-2105-10-410	article	the approach was applied to multiple cancer microarray datasets with > 50 samples per group in total	Case-control study	AUC, sensitivity, specificity (nested 10-fold CV + external testing)	cross-validation + test set	"In this paper, we are interested in the question of how many and which genes should be selected for a disease class prediction. Our work consists of a Bayesian supervised statistical learning approach to refine gene signatures with a regularization which penalizes for the correlation between the variables selected. Our novel Bayesian approach includes a prior which penalizes highly correlated features in model selection and is able to extract key genes in the highly correlated context of microarray data. On real microarray datasets, we show that our approach can refine gene signatures to obtain either the same or better predictive performance than other existing methods with a smaller number of genes."	"In this paper, we are interested in the question of how many and which genes should be selected for a disease class prediction. Our work consists of a Bayesian supervised statistical learning approach to refine gene signatures with a regularization which penalizes for the correlation between the variables selected. Our novel Bayesian approach includes a prior which penalizes highly correlated features in model selection and is able to extract key genes in the highly correlated context of microarray data. On real microarray datasets, we show that our approach can refine gene signatures to obtain either the same or better predictive performance than other existing methods with a smaller number of genes."
46	Dougherty, E R and Hua, J and Bittner, M L	Validation of computational methods in genomics	Current Genomics	8	1	1-19	2007	USA	http://dx.doi.org/10.2174/138920207780076956	article	review (not applicable)	review	"This paper treats the validation issue as it appears in two classes of inference algorithms relating to genomics – classification and clustering. It formulates the problem and reviews salient results."		The manuscript covers several commonly used approaches to evalute classification and clustering methods (e.g. cross-validation and bolstered resubstitution)	
47	Drouin, A and Giguère, S and Déraspe, M and Marchand, M and Tyers, M and Loo, V G and Bourgault, A M and Laviolette, F and Corbeil, J	Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons	BMC Genomics	17	1		2016	Canada	http://dx.doi.org/10.1186/s12864-016-2889-6	article	17 datasets in which the number of examples ranged from 111 to 556	Antibiotic resistance prediction	error rate (5-fold CV, test evaluation)	cross-validation + test set	"The identification of genomic biomarkers is a key step towards improving diagnostic tests and therapies. We present a reference-free method for this task that relies on a k-mer representation of genomes and a machine learning algorithm that produces intelligible models. The method was validated by generating models that predict the antibiotic resistance of C. difficile, M. tuberculosis, P. aeruginosa, and S. pneumoniae for 17 antibiotics. The obtained models are accurate, faithful to the biological pathways targeted by the antibiotics, and they provide insight into the process of resistance acquisition. The method is not limited to predicting antibiotic resistance in bacteria and is applicable to a variety of organisms and phenotypes."	"The identification of genomic biomarkers is a key step towards improving diagnostic tests and therapies. We present a reference-free method for this task that relies on a k-mer representation of genomes and a machine learning algorithm that produces intelligible models. The method was validated by generating models that predict the antibiotic resistance of C. difficile, M. tuberculosis, P. aeruginosa, and S. pneumoniae for 17 antibiotics. The obtained models are accurate, faithful to the biological pathways targeted by the antibiotics, and they provide insight into the process of resistance acquisition. The method is not limited to predicting antibiotic resistance in bacteria and is applicable to a variety of organisms and phenotypes."
48	Drozdov, I and Kidd, M and Modlin, I	Graph-theoretic definition of neuroendocrine disease-a tumor specific mathematical toolbox for assessing neoplastic behaviour	Neuroendocrinology	103		45-46	2016		http://dx.doi.org/10.1159/000448725	meeting abstract	130 blood samples (NENs: n = 63)	Case-control study	AUC, sensitivity, specificity, PPN, NPV (The model was validated in two independent sets (Set 1 [n = 115, NENs: n = 72]; Set 2 [n = 120, NENs: n = 58]))	training + test set	GEP NENs (gastroentero-pancreatic neuroendocrine neoplasms) were investigated by reverse-engineering intracellular signalling networks and identifying hub genes using degree (number of interactions) and betweenness (number of shortest paths) statistics. A random forest algorithm was used to assess hub gene expression in 130 blood samples (NENs: n = 63) and to differentiate healthy controls and GEP-NENs. Gene-based classifiers detected NENs in independent sets with high sensitivity (85-98%), specificity (93-97%), PPV (95-96%) and NPV (87-98%). Additionally, multi-transcript based machine learning algorithms substantially outperform single analyte assays such as CgA."	GEP NENs (gastroentero-pancreatic neuroendocrine neoplasms) were investigated by reverse-engineering intracellular signalling networks and identifying hub genes using degree (number of interactions) and betweenness (number of shortest paths) statistics. A random forest algorithm was used to assess hub gene expression in 130 blood samples (NENs: n = 63) and to differentiate healthy controls and GEP-NENs. Gene-based classifiers detected NENs in independent sets with high sensitivity (85-98%), specificity (93-97%), PPV (95-96%) and NPV (87-98%). Additionally, multi-transcript based machine learning algorithms substantially outperform single analyte assays such as CgA."
49	Elsebakhi, E and Lee, F and Schendel, E and Haque, A and Kathireason, N and Pathare, T and Syed, N and Al-Ali, R	Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms	Journal of Computational Science	11		69-81	2015		https://doi.org/10.1016/j.jocs.2015.09.008	article	Stage1 (population of 127 exacerbation cases and 290 non-exacerbation controls) and Stage2 (population of 50 exacerbation cases and 114 non-exacerbation controls)	Case-control study	AUC, sensitivity, specificity (training/test set split)	training + test set	The goal of the study was to identify severe asthma exacerbation children using phenotypic and SNP Data. We concluded that the new classifier with the Newton-Raphson iterative processes and propensity scores have reliable performance with the increase in AUC values in all cases: (i) phenotypic data only; (ii) phenotypic data with the top ten significant SNPs; and (iii) phenotypic data with the top 160 to 302 significant SNPs.	The goal of the study was to identify severe asthma exacerbation children using phenotypic and SNP Data. We concluded that the new classifier with the Newton-Raphson iterative processes and propensity scores have reliable performance with the increase in AUC values in all cases: (i) phenotypic data only; (ii) phenotypic data with the top ten significant SNPs; and (iii) phenotypic data with the top 160 to 302 significant SNPs.
50	Fan, X J and Wan, X B and Huang, Y and Cai, H M and Fu, X H and Yang, Z L and Chen, D K and Song, S X and Wu, P H and Liu, Q and Wang, L and Wang, J P	Epithelial-mesenchymal transition biomarkers and support vector machine guided model in preoperatively predicting regional lymph node metastasis for rectal cancer	Br J Cancer	106	11	1735-1741	2012	China	http://dx.doi.org/10.1038/bjc.2012.82	article	193 RC patients	Cases only (predicting lymph node metastasis)	accuracy, sensitivity, specificity (training/test set split)	training + test set	Current imaging modalities are inadequate in preoperatively predicting regional lymph node metastasis (RLNM) status in rectal cancer (RC). Here, we designed support vector machine (SVM) model to address this issue by integrating epithelial–mesenchymal-transition (EMT)-related biomarkers along with clinicopathological variables. The sensitivity, specificity and overall accuracy of SVM in predicting RLNM were 68.3%, 81.1% and 72.3%, respectively. Importantly, multivariate logistic regression analysis showed that SVM model was indeed an independent predictor of RLNM status (odds ratio, 11.536; 95% confidence interval, 4.113–32.361; P<0.0001).	Current imaging modalities are inadequate in preoperatively predicting regional lymph node metastasis (RLNM) status in rectal cancer (RC). Here, we designed support vector machine (SVM) model to address this issue by integrating epithelial–mesenchymal-transition (EMT)-related biomarkers along with clinicopathological variables. The sensitivity, specificity and overall accuracy of SVM in predicting RLNM were 68.3%, 81.1% and 72.3%, respectively. Importantly, multivariate logistic regression analysis showed that SVM model was indeed an independent predictor of RLNM status (odds ratio, 11.536; 95% confidence interval, 4.113–32.361; P<0.0001).
51	Fang, Y and Xu, P and Yang, J and Qin, Y	A quantile regression forest based method to predict drug response and assess prediction reliability	PLoS One	13	10	e0205155-e0205155	2018	China	http://dx.doi.org/10.1371/journal.pone.0205155	article	data from 947 cell lines (CCLE dataset)	Cases only (drug response prediction in-vitro)	Pearson correlation of observed and predicted drug response (out-of-bag validation)	outofbag	"Drug response prediction is a critical step for personalized treatment of cancer patients and ultimately leads to precision medicine. In this paper, we proposed a method based on quantile regression forest and applied it to the CCLE dataset. Through the out-of-bag validation, our method achieved much higher prediction accuracy of drug response than other available tools."	"Drug response prediction is a critical step for personalized treatment of cancer patients and ultimately leads to precision medicine. In this paper, we proposed a method based on quantile regression forest and applied it to the CCLE dataset. Through the out-of-bag validation, our method achieved much higher prediction accuracy of drug response than other available tools."
52	Farmakis, D and Koeck, T and Mullen, W and Parissis, J and Gogas, B D and Nikolaou, M and Lekakis, J and Mischak, H and Filippatos, G	Urine proteome analysis in heart failure with reduced ejection fraction complicated by chronic kidney disease: feasibility, and clinical and pathogenetic correlates	Eur J Heart Fail	18	7	822-829	2016	Germany	http://dx.doi.org/10.1002/ejhf.544	article	126 individuals, 59 HFrEF patients and 67 controls	Case-control study	AUC, accuracy, sensitivity, specificity (cross-validation + test set)	cross-validation + test set	"Urine proteome analysis (UPA) has already provided accurate discriminatory patterns of urinary peptides for renal disease, coronary artery disease, and asymptomatic LV diastolic dysfunction. UPA has now been used to characterize a discriminatory peptide biomarker pattern and establish a diagnostic classifier for heart failure patients with reduced ejection fraction (HFrEF) in the presence of chronic kidney disease (CKD). In total, 107 significant discriminatory peptides were identified and used to establish a support vector machine-based classifier that was successfully applied to a test set of 25 HFrEF patients and 33 controls, achieving 84% sensitivity and 91% specificity."	"Urine proteome analysis (UPA) has already provided accurate discriminatory patterns of urinary peptides for renal disease, coronary artery disease, and asymptomatic LV diastolic dysfunction. UPA has now been used to characterize a discriminatory peptide biomarker pattern and establish a diagnostic classifier for heart failure patients with reduced ejection fraction (HFrEF) in the presence of chronic kidney disease (CKD). In total, 107 significant discriminatory peptides were identified and used to establish a support vector machine-based classifier that was successfully applied to a test set of 25 HFrEF patients and 33 controls, achieving 84% sensitivity and 91% specificity."

53	Fassbender, A and Waelkens, E and Kyama, C and Bokor, A and Vodolazkaia, A and Verbeeck, N and Van De Plas, R and Ojeda, F and Gevaert, O and Meuleman, C and Peeraer, K and Tomassetti, C and De Moor, B and D'Hooghe, T	Biomarkers in plasma or serum: Pitfalls in data processing	Reproductive Sciences	18	3	191A-191A	2011	India	http://dx.doi.org/10.1177/1933719120111183s067	meeting abstract	254 plasma samples from women with (n=165) & without (n=89) endometriosis	Case-control study	accuracy (data were divided randomly (100 times) into training set (70%) and test set (30%))	training + test set	The goal of the study was to test the hypothesis that specific proteins/peptides are differentially expressed in plasma of women with & without endometriosis at specific stages of the disease, during follicular, luteal & menstrual phases. Using the standard practices in the field, a highly correct classification could be reached on both training and test set, however good machine learning practice dictates the use of robust testing by randomizing the data. Using this method, the high classification attained on the training set could not be confirmed. The classification performance increased by separating the different phases, especially the menstrual phase, compared to combining the different phases together. Randomization of the data set should become an integral part of this type of analyses in the future."	The goal of the study was to test the hypothesis that specific proteins/peptides are differentially expressed in plasma of women with & without endometriosis at specific stages of the disease, during follicular, luteal & menstrual phases. Using the standard practices in the field, a highly correct classification could be reached on both training and test set, however good machine learning practice dictates the use of robust testing by randomizing the data. Using this method, the high classification attained on the training set could not be confirmed. The classification performance increased by separating the different phases, especially the menstrual phase, compared to combining the different phases together. Randomization of the data set should become an integral part of this type of analyses in the future."
54	Fillmore, N and Ramos-Cejudo, J and Cheng, D and Tuck, D P and Sheikh, A R and Chen, D and Elbers, D and Sung, F C and Johnson, B and Shannon, C and Pierce-Murray, K and Gaynor, K and Dedomenico, C and Schiller, S and Ajjarapu, S and Hall, R and Ayandeh, S and Meng, F and Brophy, M T and Do, N	A predictive model for survival in non-small cell lung cancer (NSCLC) based on electronic health record (EHR) and tumor sequencing data at the Department of Veterans Affairs (VA)	Journal of Clinical Oncology	37			2019		https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.109	meeting abstract	356 VA patients newly diagnosed with NSCLC	Case-control study	Precision, recall, and area under the ROC curve (AUC) (5-fold CV)	cross-validation	"Here, we combine real-world EHR and tumor sequencing data from the VA Precision Oncology Data Repository (PODR) to build accurate individualized survival predictions in newly-diagnosed NSCLC patients. Our predictive model for 1-year survival achieves strong results. Cross-validated AUC is 0.79 (SD 0.08), precision is 0.79 (SD 0.07), recall is 0.74 (SD 0.07)."	"Here, we combine real-world EHR and tumor sequencing data from the VA Precision Oncology Data Repository (PODR) to build accurate individualized survival predictions in newly-diagnosed NSCLC patients. Our predictive model for 1-year survival achieves strong results. Cross-validated AUC is 0.79 (SD 0.08), precision is 0.79 (SD 0.07), recall is 0.74 (SD 0.07)."
55	Firoozbakht, F and Rezaeian, I and D'Agnillo, M and Porter, L and Rueda, L and Ngom, A	An Integrative Approach for Identifying Network Biomarkers of Breast Cancer Subtypes Using Genomic, Interactomic, and Transcriptomic Data	J Comput Biol	24	8	756-766	2017	Canada	http://dx.doi.org/10.1089/cmb.2017.0010	article	"We have used the METABRIC data set (Curtis et al., 2012), which contains the copy number values and GE levels of 2000 primary breast tumors with long-term clinical follow-up"	Tumour subtype categorization	AUC, sensitivity, specificity (10-fold CV)	cross-validation	"Breast cancer is a complex disease that can be classified into at least 10 different molecular subtypes. Appropriate diagnosis of specific subtypes is critical for ensuring the best possible patient treatment and response to therapy. Cancer network bio-markers are subnetworks of functionally related genes that ''work in concert'' to perform functions associated with a tumorigenic. We propose a machine learning framework that can be used to identify network biomarkers and driver genes for each specific breast cancer subtype. Our results show that the resulting network biomarkers can separate one subtype from the others with very high accuracy."	"Breast cancer is a complex disease that can be classified into at least 10 different molecular subtypes. Appropriate diagnosis of specific subtypes is critical for ensuring the best possible patient treatment and response to therapy. Cancer network bio-markers are subnetworks of functionally related genes that ''work in concert'' to perform functions associated with a tumorigenic. We propose a machine learning framework that can be used to identify network biomarkers and driver genes for each specific breast cancer subtype. Our results show that the resulting network biomarkers can separate one subtype from the others with very high accuracy."
56	Fong, F and Bar, H Y and Shedden, K and Saiya-Cork, K and Ouillette, P and Campagne, F and Melnick, A and Malek, S and Shaknovich, R	Epigenetic profiling of primary CLL reveals novel DNA methylation-based clusters and novel mechanisms of leukemogenesis	Blood	120	21		2012		https://doi.org/10.1182/blood.V120.21.3877.3877	meeting abstract	"DNA methylation of over 240 patients with CLL"	Case-control study	AUC (10-fold CV)	cross-validation	"Chronic Lymphocytic Leukemia (CLL) is the most common leukemia in the Western world with nearly 15,000 new cases diagnosed every year in the USA. We hypothesized that DNA methylation profiling would allow us to identify new, biologically significant CLL subtypes and yield greater insight into the biology of this disease. We performed unsupervised analysis on the most variable probesets (standard deviation > 1.3) using K-means consensus clustering. inally, we divided the cohort into training and testing cohorts and used a machine learning BDVAL algorithm to identify DNA methylation outcome classifiers. This procedure identified a 40-probeset classifier that accurately predicted outcome (Area Under the ROC Curve of 0.77."	"Chronic Lymphocytic Leukemia (CLL) is the most common leukemia in the Western world with nearly 15,000 new cases diagnosed every year in the USA. We hypothesized that DNA methylation profiling would allow us to identify new, biologically significant CLL subtypes and yield greater insight into the biology of this disease. We performed unsupervised analysis on the most variable probesets (standard deviation > 1.3) using K-means consensus clustering. inally, we divided the cohort into training and testing cohorts and used a machine learning BDVAL algorithm to identify DNA methylation outcome classifiers. This procedure identified a 40-probeset classifier that accurately predicted outcome (Area Under the ROC Curve of 0.77."
57	Gal, O and Auslander, N and Fan, Y and Meerzaman, D	Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression	Cancer Informatics	18			2019	USA	http://dx.doi.org/10.1177/1176935119835544	article	473 bone marrow specimens from 473 patients	Case-control study	AUC (5-fold CV + test set)	cross-validation + test set	In this study, we explored the question of whether ML algorithms designed to analyze gene-expression patterns obtained through RNA sequencing (RNA-seq) can be used to accurately predict the likelihood of complete remission (CR) in pediatric AML patients who have received induction therapy. We tuned classifier hyperparameters to optimize performance and used multiple methods to guide our feature selection as well as our assessment of algorithm performance. To identify the model which performed best within the context of this study, we plotted receiver operating characteristic (ROC) curves. Using the top 75 genes from the k-nearest neighbors algorithm (K-NN) model (K = 27) yielded the best area-under-the-curve (AUC) score that we obtained: 0.84. When we finally tested the previously unseen test data set, the top 50 genes yielded the best AUC = 0.81."	In this study, we explored the question of whether ML algorithms designed to analyze gene-expression patterns obtained through RNA sequencing (RNA-seq) can be used to accurately predict the likelihood of complete remission (CR) in pediatric AML patients who have received induction therapy. We tuned classifier hyperparameters to optimize performance and used multiple methods to guide our feature selection as well as our assessment of algorithm performance. To identify the model which performed best within the context of this study, we plotted receiver operating characteristic (ROC) curves. Using the top 75 genes from the k-nearest neighbors algorithm (K-NN) model (K = 27) yielded the best area-under-the-curve (AUC) score that we obtained: 0.84. When we finally tested the previously unseen test data set, the top 50 genes yielded the best AUC = 0.81."
58	Gamberger, D and Lavrac, N and Zelezny, F and Tolar, J	Induction of comprehensible models for gene expression datasets by subgroup discovery methodology	J Biomed Inform	37	4	269-284	2004	Croatia	http://dx.doi.org/10.1016/j.jbi.2004.07.007	article	the approach was applied to multiple cancer microarray datasets with > 50 samples per group in total	Case-control study	sensitivity, specificity, precision (training/test set split)	training + test set	"The goal of this study is to [construct| simple yet robust logic-based classifiers amenable to direct expert interpretation. On two well-known, publicly available gene expression classification problems, the paper shows the feasibility of this approach, employing a recently developed subgroup discovery methodology. Some of the discovered classifiers allow for novel biological interpretations."	"The goal of this study is to [construct| simple yet robust logic-based classifiers amenable to direct expert interpretation. On two well-known, publicly available gene expression classification problems, the paper shows the feasibility of this approach, employing a recently developed subgroup discovery methodology. Some of the discovered classifiers allow for novel biological interpretations."
59	Gao, H and Zheng, Z and Yue, Z and Liu, F and Zhou, L and Zhao, X	Evaluation of serum diagnosis of pancreatic cancer by using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry	Int J Mol Med	30	5	1061-1068	2012	China	http://dx.doi.org/10.3892/ijmm.2012.1113	article	serum samples from 132 patients with PCa and 67 healthy controls	Case-control study	sensitivity, specificity (leave-one-out cross-validation)	cross-validation	"The aim of this study was to discover potential biomarkers for pancreatic cancer (PCa) using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS). Support vector machine (SVM) analysis of the spectra was used to generate a predictive algorithm based on proteins that were maximally differentially expressed between patients with PCa and the HCs in the training cohort. This algorithm was tested using leave-one-out cross-validation in the test cohort. The classifier was challenged with all samples achieving 96.67% sensitivity and 100% specificity in the training cohort and 93.1% sensitivity and 78.57% specificity in the test cohort."	"The aim of this study was to discover potential biomarkers for pancreatic cancer (PCa) using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS). Support vector machine (SVM) analysis of the spectra was used to generate a predictive algorithm based on proteins that were maximally differentially expressed between patients with PCa and the HCs in the training cohort. This algorithm was tested using leave-one-out cross-validation in the test cohort. The classifier was challenged with all samples achieving 96.67% sensitivity and 100% specificity in the training cohort and 93.1% sensitivity and 78.57% specificity in the test cohort."
60	Gerl, M J and Klose, C and Surma, M A and Fernandez, C and Melander, O and Mannisto, S and Borodulin, K and Havulinna, A S and Salomaa, V and Ikonen, E and Cannistraci, C V and Simons, K	Machine learning of human plasma lipidomes for obesity estimation in a large population cohort	Plos Biology	17	10	25-25	2019	China	http://dx.doi.org/10.1371/journal.pbio.3000443	article	Samples of the FINRISK 2012 underwent lipidomics measurements (1,141 randomly selected individuals) of which 1,061 were used	Case-control study	R-squared of obesity indicator variables (5x repeated 10-fold CV)	cross-validation	"We aimed at predicting different measures of obesity based on the plasma lipidome in a large population cohort using advanced machine learning modeling. Multiple machine intelligence models were trained to predict obesity estimates, i.e., body mass index (BMI), waist circumference (WC), waist-hip ratio (WHR), and body fat percentage (BFP) and validated in 250 randomly chosen participants of the Malmö Diet and Cancer Cardiovascular Cohort (MDC-CC). Comparison of the different models revealed that the lipidome predicted BFP the best (R2 = 0.73), based on a Lasso model. In this model, the strongest positive and the strongest negative predictor were sphingomyelin molecules, which differ by only 1 double bond, implying the involvement of an unknown desaturase in obesity-related aberrations of lipid metabolism. Moreover, we used this regression to probe the clinically relevant information contained in the plasma lipidome and found that the plasma lipidome also contains information about body fat distribution, because WHR (R2 = 0.65) was predicted more accurately than BMI (R2 = 0.47)."	
61	Gong, I Y and Fox, N S and Huang, V and Boutros, P C	Prediction of early breast cancer patient survival using ensembles of hypoxia signatures	PLoS One	13	9	e0204123-e0204123	2018	Canada	http://dx.doi.org/10.1371/journal.pone.0204123	article	1,564 early breast cancer patients	Cases only (survival prediction)	AUC, accuracy, sensitivity, specificity (10-fold cross-validation + test set)	cross-validation + test set	"We risk stratified breast cancer patients into either low-risk or high-risk groups based on four published hypoxia signatures (Buffa, Winter, Hu, and Sorensen), using 24 different preprocessing approaches for microarray normalization. The 24 binary risk profiles determined for each hypoxia signature were combined using a random forest to evaluate the efficacy of a preprocessing ensemble classifier. We demonstrate that the best way of merging preprocessing methods varies from signature to signature, and that there is likely no 'best' preprocessing pipeline that is universal across datasets, highlighting the need to evaluate ensembles of preprocessing algorithms. Further, we developed novel signatures for each preprocessing method and the risk classifications from each were incorporated in a meta-random forest model. Interestingly, the classification of these biomarkers and its ensemble show striking consistency, demonstrating that similar intrinsic biological information are being faithfully represented. As such, these classification patterns further confirm that there is a subset of patients whose prognosis is consistently challenging to predict."	"We risk stratified breast cancer patients into either low-risk or high-risk groups based on four published hypoxia signatures (Buffa, Winter, Hu, and Sorensen), using 24 different preprocessing approaches for microarray normalization. The 24 binary risk profiles determined for each hypoxia signature were combined using a random forest to evaluate the efficacy of a preprocessing ensemble classifier. We demonstrate that the best way of merging preprocessing methods varies from signature to signature, and that there is likely no 'best' preprocessing pipeline that is universal across datasets, highlighting the need to evaluate ensembles of preprocessing algorithms. Further, we developed novel signatures for each preprocessing method and the risk classifications from each were incorporated in a meta-random forest model. Interestingly, the classification of these biomarkers and its ensemble show striking consistency, demonstrating that similar intrinsic biological information are being faithfully represented. As such, these classification patterns further confirm that there is a subset of patients whose prognosis is consistently challenging to predict."
62	Graim, K and Friedl, V and Houlahan, K E and Stuart, J M	PLATYPUS: A Multiple-View Learning Predictive Framework for Cancer Drug Sensitivity Prediction	Pac Symp Biocomput	24		136-147	2019	USA	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417802/	article	At the time of download the Cancer Cell Line Encyclopedia (CCLE) contained genomic, phenotype, clinical, and other annotation data for 1,037 cancer cell lines	Cases only (drug sensitivity prediction)	AUC (cross-validation + external test set)	cross-validation + external cohort validation	"We introduce a multi-view machine-learning strategy called PLATYPUS that builds 'views' from multiple data sources that are all used as features for predicting patient outcomes. We show that a learning strategy that finds agreement across the views on unlabeled data increases the performance of the learning methods over any single view."	"The paper presents a multi-view machine-learning strategy using co-training. Co–training works by training a separate classifier using each view as a separate feature set to make independent predictions, then incorporating disagreement into the loss function. Each view trains on the labeled data then predicts labels for the common unlabeled set. High confidence labels are passed as truth in the next iteration. Co–training methods iterate until either convergence, some threshold (a minimal change in label definition on the unlabeled samples) is attained, or a maximum number of iterations is reached."
63	Grollemund, V and Pradat, P F and Querin, G and Delbot, F and Le Chat, G and Pradat-Peyre, J F and Bede, P	Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions	Frontiers in Neuroscience	13			2019	France	http://dx.doi.org/10.3389/fnins.2019.00135	article	review (not applicable)	review			"Despite considerable sample size limitations, ML techniques have already been successfully applied to ALS data sets and a number of promising diagnosis models have been proposed. Prognostic models have been tested using core clinical variables, biological, and neuroimaging data. These models also offer patient stratification opportunities for future clinical trials. Despite the enormous potential of ML in ALS research, statistical assumptions are often violated, the choice of specific statistical models is seldom justified, and the constraints of ML models are rarely enunciated. rom a mathematical perspective, the main barrier to the development of validated diagnostic, prognostic, and monitoring indicators stem from limited sample sizes. The combination of multiple clinical, biofluid, and imaging biomarkers is likely to increase the accuracy of mathematical modeling and contribute to optimized clinical trial designs."	"Despite considerable sample size limitations, ML techniques have already been successfully applied to ALS data sets and a number of promising diagnosis models have been proposed. Prognostic models have been tested using core clinical variables, biological, and neuroimaging data. These models also offer patient stratification opportunities for future clinical trials. Despite the enormous potential of ML in ALS research, statistical assumptions are often violated, the choice of specific statistical models is seldom justified, and the constraints of ML models are rarely enunciated. rom a mathematical perspective, the main barrier to the development of validated diagnostic, prognostic, and monitoring indicators stem from limited sample sizes. The combination of multiple clinical, biofluid, and imaging biomarkers is likely to increase the accuracy of mathematical modeling and contribute to optimized clinical trial designs."
64	Guan, N N and Zhao, Y and Wang, C C and Li, J Q and Chen, X and Piao, X	Anticancer Drug Response Prediction in Cell Lines Using Weighted Graph Regularized Matrix Factorization	Molecular Therapy - Nucleic Acids	17		164-174	2019	China	http://dx.doi.org/10.1016/j.omtn.2019.05.017	article	"The first dataset we used was collected from Genomic of Drug Sensitivity in Cancer project (release-5.0, https://www.cancerrxgene.org/downloads), including 652 cancer cell lines, 135 drugs, and 70,676 known response values. The second dataset was collected from the CCLE (https://portals.broadinstitute.org/ccle), which contains 23 drugs and 491 cell lines with 10,870 known responses"	Cases only (drug response prediction in-vitro)	Pearson correlation coefficient (PCC), root-mean-square error (RMSE), PCCsr, and RMSEsr averaged over all drugs (10-fold CV)	cross-validation	"In this work, we presented a novel method to utilize weighted graph regularized matrix factorization (WGRMF) for inferring anticancer drug response in cell lines. We constructed a p-nearest neighbor graph to sparsify drug similarity matrix and cell line similarity matrix, respectively. The results on the Genomics of Drug Sensitivity in Cancer (GDSC) dataset are 0.64 ± 0.16, 1.37 ± 0.35, 0.73 ± 0.14, and 1.71 ± 0.44 for PCC, RMSE, PCCsr, and RMSEsr in turn. And for the Cancer Cell Line Encyclopedia (CCLE) dataset, WGRMF got results of 0.72 ± 0.09, 0.56 ± 0.19, 0.79 ± 0.07, and 0.69 ± 0.19, respectively."	"In this work, we presented a novel method to utilize weighted graph regularized matrix factorization (WGRMF) for inferring anticancer drug response in cell lines. We constructed a p-nearest neighbor graph to sparsify drug similarity matrix and cell line similarity matrix, respectively. The results on the Genomics of Drug Sensitivity in Cancer (GDSC) dataset are 0.64 ± 0.16, 1.37 ± 0.35, 0.73 ± 0.14, and 1.71 ± 0.44 for PCC, RMSE, PCCsr, and RMSEsr in turn. And for the Cancer Cell Line Encyclopedia (CCLE) dataset, WGRMF got results of 0.72 ± 0.09, 0.56 ± 0.19, 0.79 ± 0.07, and 0.69 ± 0.19, respectively."
65	Guo, S and Guo, D and Chen, L and Jiang, Q	A centroid-based gene selection method for microarray data classification	J Theor Biol	400		32-41	2016	China	http://dx.doi.org/10.1016/j.jtbi.2016.03.034	article	multiple microarray datasets for different cancers with > 50 sample per group in total were used	Case-control study	accuracy + standard deviation (repeated 5-fold CV)	cross-validation	"In this paper, a new gene selection method is proposed to choose the best subset of features for microarray data with the irrelevant and redundant features removed. We formulate the selection problem as a L1-regularized optimization problem, based on a newly defined linear discriminant analysis criterion. The experimental results on ten publicly available microarray da- tasets demonstrate that the proposed method performs effectively and competitively compared with state-of-the-art methods."	"In this paper, a new gene selection method is proposed to choose the best subset of features for microarray data with the irrelevant and redundant features removed. We formulate the selection problem as a L1-regularized optimization problem, based on a newly defined linear discriminant analysis criterion. The experimental results on ten publicly available microarray da- tasets demonstrate that the proposed method performs effectively and competitively compared with state-of-the-art methods."
66	Guo, Y and Yu, H and Chen, D Q and Zhao, Y Y	Machine learning distilled metabolite biomarkers for early stage renal injury	Metabolomics	16	1	10-10	2019	China	http://dx.doi.org/10.1007/s11306-019-1624-0	article	"Serum samples were collected from all participants, including 587 CKD patients (CKD1 = 120, CKD2 = 104, CKD3 = 110, CKD4 = 119, CKD5 = 134) and 110 age-matched normal healthy controls."	Case-control study	AUC (10-fold CV)	cross-validation	"Shotgun high throughput was applied to the detection of serum metabolites biomarker discovery for early stages of chronic kidney disease (CKD) from 703 participants. We discovered a set of metabolite biomarkers using a series of classic and neural network based machine learning techniques. This set of metabolites can separate early CKD stage patents from normal subjects with high accuracy. Our study illustrates the power of machine learning methods in metabolite biomarker study."	"Shotgun high throughput was applied to the detection of serum metabolites biomarker discovery for early stages of chronic kidney disease (CKD) from 703 participants. We discovered a set of metabolite biomarkers using a series of classic and neural network based machine learning techniques. This set of metabolites can separate early CKD stage patents from normal subjects with high accuracy. Our study illustrates the power of machine learning methods in metabolite biomarker study."

67	Han, H	Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery	BMC Bioinformatics	11		S1-S1	2010	USA	http://dx.doi.org/10.1186/1471-2105-11-s1-s1	article	4 different case/control cancer MS serum profile datasets were analyzed with > 50 samples per group on average	Case-control study	accuracy (LOOCV and 100 trials of 50% holdout cross validations (HOCV))	cross-validation	"In this study, we develop a nonnegative principal component analysis algorithm and present a nonnegative principal component analysis based support vector machine algorithm with sparse coding to conduct a high-performance proteomic pattern classification. Moreover, we also propose a nonnegative principal component analysis based filter-wrapper biomarker capturing algorithm for mass spectral serum profiles. We demonstrate the superiority of the proposed algorithm by comparison with six peer algorithms on four benchmark datasets."
68	Hao, Y Y and Duh, Q Y and Kloos, R T and Babiarz, J and Harrell, R M and Traweek, S T and Kim, S Y and Fedorowicz, G and Walsh, P S and Sadow, P M and Huang, J and Kennedy, G C	Identification of Hurthle cell cancers: solving a clinical challenge with genomic sequencing and a trio of machine learning algorithms	Bmc Systems Biology	13		14-14	2019	USA	http://dx.doi.org/10.1186/s12918-019-0693-z	article	318 samples, including 119 Hürthle cell-negative and 199 Hürthle cell-positive samples	Case-control study	AUC, sensitivity, specificity (10-fold nested CV)	cross-validation	"Identification of Hürthle cell cancers by non-operative fine-needle aspiration biopsy (FNAB) of thyroid nodules is challenging. It is highly desirable to avoid surgery and to provide accurate classification of benignity versus malignancy from FNAB preoperatively. In our first-generation algorithm, Gene Expression Classifier (GEC), we achieved this goal by using machine learning (ML) on gene expression features. The classifier is sensitive, but not specific due in part to the presence of non-neoplastic benign Hürthle cells in many FNAB. We sought to overcome this low-specificity limitation by expanding the feature set for ML using next-generation whole transcriptome RNA sequencing and called the improved algorithm the Genomic Sequencing Classifier (GSC). The accurate algorithmic depiction of this complex biological system among Hürthle subtypes results in a dramatic improvement of classification performance; specificity among Hürthle cell neoplasms increases from 11.8% with the GEC to 58.8% with the GSC, while maintaining the same sensitivity of 89%."
69	Heard, B J and Rosvold, J M and Fritzler, M J and El-Gabalawy, H and Wiley, J P and Krawetz, R J	A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers	J R Soc Interface	11	97	20140428-20140428	2014	Canada	http://dx.doi.org/10.1098/rsif.2014.0428	article	"normal individuals (normal, n = 100), patients with osteoarthritis (OA, n = 100), and rheumatoid arthritis (RA, n = 100)"	Case-control study	accuracy, sensitivity, specificity (training, validation and test set split)	cross-validation + test set	"The objective of this study was to develop a method for categorizing normal individuals (normal, n = 100) as well as patients with osteoarthritis (OA, n = 100), and rheumatoid arthritis (RA, n = 100) based on a panel of inflammatory cytokines expressed in serum samples. We show that ANNs can be used for the accurate differentiation between serum samples of patients with OA, a diagnosed RA patient comparator cohort and normal/control cohort. Using neural network and systems biology approaches to manage large datasets derived from high-throughput proteomics should be further explored and considered for diagnosing diseases with complex pathologies."
70	Hernández, B and Pennington, S R and Parnell, A C	Bayesian methods for proteomic biomarker development	EuPA Open Proteomics	9		54-64	2015		http://dx.doi.org/10.1016/j.euprot.2015.08.001	article	review (not applicable)	review			"Bayesian methods offer some advantages over traditional statistical and machine learning methods. In particular they can incorporate external information into current experiments so as to guide biomarker selection. Further, they can be more robust to over-fitting than other approaches, especially when the number of samples used for discovery is relatively small. In this review we provide an introduction to Bayesian inference and demonstrate some of the advantages of using a Bayesian framework."
71	Hira, Z M and Gillies, D F	Identifying significant features in cancer methylation data using gene pathway segmentation	Cancer Informatics	15		189-198	2016	UK	http://dx.doi.org/10.4137/CIN.S39859	article	"Two datasets were used in this study. The first was the Cancer Genome Atlas (TCGA) methylation brain lower grade glioma (LGG) dataset (http://cancergenome.nih.gov). In this, there are 370,203 probes and in total 531 samples. The second dataset was an as yet unpublished study of chronic myelogenous leukemia (CML) with 429,231 probes and 91 samples. The data supplied for this research was fully anonymized. In this set, 60 samples were responsive to CML treatment, while 31 samples were not."	Case-control study	AUC, accuracy (stratified 10-fold CV)	cross-validation	"In order to provide the most effective therapy for cancer, it is important to be able to diagnose whether a patient's cancer will respond to a proposed treatment. Methylation profiling could contain information from which such predictions could be made. we have investigated a way of utilizing prior knowledge to segment microarray datasets in such a way that machine learning can be used to identify candidate sets of genes for hypothesis testing. A methylation dataset is divided into subsets, where each subset contains only the probes that relate to a known gene pathway. Each of these pathway subsets is used independently for classification. The classification method is AdaBoost with decision trees as weak classifiers. Since each pathway subset contains a relatively small number of gene probes, it is possible to train and test its classification accuracy quickly and determine whether it has valuable diagnostic information."
72	Ho, D S W and Schierding, W and Wake, M and Saffery, R and O'Sullivan, J	Machine learning SNP based prediction for precision medicine	Frontiers in Genetics	10			2019	Australia	http://dx.doi.org/10.3389/fgene.2019.00267	article	review (not applicable)	review			"Polygenic risk scoring is a straightforward model for assigning genetic risk to individual outcomes, but has achieved only limited success in complex disease predictions due to its dependency on linear regression. The polygenic risk scoring method is ineffective in modeling highly dimensional genotype data with complex interactions. By contrast, the strength of machine learning data modeling in complex disease prediction lies in its handling of interactive high-dimensional data. Coupled with large new population datasets with high-quality phenotyping at different stages in the lifecourse, machine learning models are capable of classifying individual disease risks with high precision. Notably, machine learning predictors that include tissue-specific disease risks for individuals show even greater promise of insights that could ultimately provide cost-effective and proactive healthcare with great efficacy."
73	Honda, K and Hayashida, Y and Umaki, T and Okusaka, T and Kosuge, T and Kikuchi, S and Endo, M and Tsuchida, A and Aoki, T and Itoi, T and Moriyasu, F and Hirohashi, S and Yamada, T	Possible detection of pancreatic cancer by plasma protein profiling	Cancer Res	65	22	10613-10622	2005	Japan	http://dx.doi.org/10.1158/0008-5472.can-05-1851	article	71 pancreatic cancer patients and 71 healthy controls	Case-control study	AUC, sensitivity, specificity (LOOCV, external test set)	cross-validation + external cohort validation	"The survival rate of pancreatic cancer patients is the lowest among those with common solid tumors, and early detection is one of the most feasible means of improving outcomes. We compared plasma proteomes between pancreatic cancer patients and sex- and age-matched healthy controls using surface-enhanced laser desorption/ionization coupled with hybrid quadrupole time-of-flight mass spectrometry. We recognized a set of four mass peaks at 8,766, 17,272, 28,080, and 14,779 m/z, whose mean intensities differed significantly (Mann-Whitney U test, P < 0.01), as most accurately discriminating cancer patients from healthy controls in the training cohort [sensitivity of 97.2% (69 of 71), specificity of 94.4% (67 of 71), and area under the curve value of 0.978]. This set discriminated cancer patients in the first validation cohort with a sensitivity of 90.9% (30 of 33) and a specificity of 91.1% (41 of 45), and its discriminating capacity was further validated in an independent cohort at a second institution. When combined with CA19-9, 100% (29 of 29 patients) of pancreatic cancers, including early-stage (stages I and II) tumors, were detected."
74	Howrylak, J A and Walter, V and Wasserman, E and Moll, M and Dolinay, T and Schultz, L and Ma, Z and Choi, A M and Baron, R M and Thomas, N J and Wong, H R and Broach, J R and Chinchilli, V M	Genomics analysis of gene expression profiles demonstrates a distinct ARDS signature	American Journal of Respiratory and Critical Care Medicine	197			2018		https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2018.197.1_MeetingAbstracts.A7533	meeting abstract	training cohort (n = 318, 75%), validation cohort (n = 105, 25%)	Case-control study	sensitivity, specificity (training/test set split)	training + test set	"The acute respiratory distress syndrome (ARDS) is a complication of critical illness and accurate diagnosis remains challenging. Lack of objective molecular biomarkers for ARDS hampers the selection of study subjects for large-scale clinical trials as well as the development of novel therapeutic options. We pooled gene expression profiles from five publicly available datasets from critically ill adult and pediatric patients with and without ARDS and applied machine learning techniques to a discovery cohort of subjects to develop a predictive model of ARDS that we subsequently validated in an independent cohort of subjects. [...] The sensitivity and specificity of this model at identifying ARDS patients in the validation cohort were 73.7% and 85.1%, respectively.
75	Huang, C and Clayton, E A and Matyunina, L V and McDonald, L D and Benigno, B B and Vannberg, F and McDonald, J F	Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy	Sci Rep	8	1	16444-16444	2018	USA	http://dx.doi.org/10.1038/s41598-018-34753-5	article	175 cancer patients	Cases only (treatment response prediction)	accuracy, sensitivity, specificity (LOOCV)	cross-validation	"We report here on the application of our previously established open-source support vector machine (SVM)-based algorithm to predict the responses of 175 individual cancer patients to a variety of standard-of-care chemotherapeutic drugs from the gene-expression profiles (RNA-seq or microarray) of individual patient tumors. The models were found to predict patient responses with >80% accuracy."
76	Huang, Y C and Chung, H H and Dutkiewicz, E P and Chen, C L and Hsieh, H Y and Chen, B R and Wang, M Y and Hsu, C C	Predicting Breast Cancer by Paper Spray Ion Mobility Spectrometry Mass Spectrometry and Machine Learning	Analytical Chemistry	92	2	1653-1657	2020	United States	http://dx.doi.org/10.1021/acs.analchem.9b03966	article	breast core needle biopsies: 29+177 benign, 14+42 malignant	Case-control study	accuracy, sensitivity, specificity (cross-validation + external validation)	cross-validation + external cohort validation	"Here, we demonstrated that by utilizing paper spray ionization-mass spectrometry (PSI-MS) coupled with field asymmetric waveform ion mobility spectrometry (FAIMS), predictive metabolic and lipidomic profiles of routine breast core needle biopsies could be obtained effectively. By the combination of machine learning algorithms and pathological examination reports, we developed a classification model, which has an overall accuracy of 87.5% for an instantaneous differentiation between cancerous and noncancerous breast tissues utilizing metabolic and lipidomic profiles."
77	Huang, Y H and Kuo, H C and Li, S C and Cai, X Y and Liu, S F and Kuo, H C	HAMP promoter hypomethylation and increased hepcidin levels as biomarkers for Kawasaki disease	J Mol Cell Cardiol	117		82-87	2018	England	http://dx.doi.org/10.1016/j.yjmcc.2018.02.017	article	"241 cases, including 18 KD patients, who were tested both prior to receiving intravenous immunoglobulin (IVIG) and at least 3 weeks after IVIG treatment, and 18 febrile controls, who were observed in the Illumina HumanMethylation450 BeadChip study for their CpG markers. The remaining cases consisted of another 92 KD patients and 113 controls that were used for validation by pyrosequencing"	Case-control study	AUC, sensitivity, specificity (5-fold CV, external test set)	cross-validation + external cohort validation	"Kawasaki disease (KD) is the most common coronary vasculitis to appear in children with anemia and has been associated with elevated plasma hepcidin levels. We recruited a total of 241 cases, including 18 KD patients, who were tested both prior to receiving intravenous immunoglobulin (IVIG) and at least 3 weeks after IVIG treatment, and 18 febrile controls, who were observed in the Illumina HumanMethylation450 BeadChip study for their CpG markers. The remaining cases consisted of another 92 KD patients and 113 controls that were used for validation by pyrosequencing. We performed a genetic functional study using Luciferase assays. A support vector machine (SVM) classification model was adopted to identify KD patients and control subjects. We developed a SVM classification model with a 90.9% sensitivity, a 91.9% specificity, and 0.94 auROC in the training set. An independent blind cohort also had good performance (96.1% sensitivity and 89.7% specificity)."
78	Huang, Z C and Shi, Y Y and Cai, B and Wang, L L and Wu, Y K and Ying, B W and Feng, W H and Hu, C J and Li, Y Z	[Promising diagnostic model for systemic lupus erythematosus using proteomic fingerprint technology]	Sichuan Da Xue Xue Bao Yi Xue Ban	40	3	499-503	2009	China	https://europepmc.org/article/med/19627014	article	Blood samples were collected from 64 cases of SLE, 30 cases of rheumatoid arthritis (RA), 30 cases of Sjogren's syndrome (SS), 25 cases of systemic sclerosis (SSc), as well as 83 healthy controls (segregating SLE from non-SLE)	Case-control study	sensitivity, specificity (training/test set split)	training + test set	"The goal of the study was to establish a diagnostic model for systemic lupus erythematosus (SLE) using proteiomic fingerprint techology. The algorithm identified a cluster pattern segregating SLE from non-SLE with sensitivity of 91% and specificity of 92%. The discriminatory diagnostic pattern correctly identified SLE. A sensitivity of 78% and specificity of 96% for the blinded test were obtained when comparing SLE vs non-SLE."
79	Iniesta, R and Hodgson, K and Malki, K and Maier, W and Rietschel, M and Mors, O and Hauser, J and Henigsberg, N and Dernovsek, M Z and Souery, D and Stahl, D and Farmer, A and Lewis, C and McGuffin, P and Uher, R	Combining clinical and genetic variables to predict antidepressant treatment response: A machine learning approach	European Neuropsychopharmacology	27		S353-S354	2017		http://dx.doi.org/10.1016/j.euroneuro.2015.09.010	meeting abstract	430 patients with unipolar depression	Cases only (treatment response prediction)	Accuracy for remission prediction (10-fold CV)	cross-validation	"In this study, we use regularized models to test the predictive ability of a combination of Genome Wide single nucleotide polymorphisms (SNPs), transcriptomic and clinical variables to predict antidepressant treatment response in a total of 430 patients with unipolar depression from the Genome-based Therapeutic Drugs for Depression pharmacogenetic study (GENDEP). A model including 14 variables (4 clinical and 10 SNPs) explained a 3.76% of the variance of the percentage of symptoms improvement in the whole sample. By drug, a model including 20 SNPs showed a outcome variance explained (measured by R2) of 16.03% in the nortriptyline-treated group and of 15.36% among patients in the escitalopram arm using 17 variables (5 clinical and 12 SNPs). Accuracy for remission prediction was 0.68 (pval =0.026), 0.79 (pval<0.001) and 0.70 (pval =0.004) for whole, nortriptyline and escitalopram samples respectively."

80	Ishii, H and Saitoh, M and Sakamoto, K and Sakamoto, K and Saigusa, D and Kasai, H and Ashizawa, K and Miyazawa, K and Takeda, S and Masuyama, K and Yoshimura, K	Lipidome-based rapid diagnosis with machine learning for detection of TGF-beta signalling activated area in head and neck cancer	British Journal of Cancer			10-10		Japan	http://dx.doi.org/10.1038/s41416-020-0732-y	article	A total of 240 and 90 mass spectra were obtained from TGF-β-unstimulated and -stimulated HNSCC cells, respectively	Case-control study	acuracy (LOOCV)	cross-validation	"We established a rapid diagnostic system based on the combination of probe electrospray ionisation-mass spectrometry (PESI-MS) and machine learning without the aid of immunohistological and biochemical procedures to identify tumour areas with heterogeneous TGF-β signalling status in head and neck squamous cell carcinoma (HNSCC). This discriminant algorithm achieved 98.79% accuracy in discrimination of TGF-β1-stimulated cells from untreated cells."	"We established a rapid diagnostic system based on the combination of probe electrospray ionisation-mass spectrometry (PESI-MS) and machine learning without the aid of immunohistological and biochemical procedures to identify tumour areas with heterogeneous TGF-β signalling status in head and neck squamous cell carcinoma (HNSCC). This discriminant algorithm achieved 98.79% accuracy in discrimination of TGF-β1-stimulated cells from untreated cells."
81	Ivancic, M M and Megna, B W and Sverchkov, Y and Craven, M and Reichelderfer, M and Pickhardt, P J and Sussman, M R and Kennedy, G D	Noninvasive Detection of Colorectal Carcinomas Using Serum Protein Biomarkers	J Surg Res	246		160-169	2020	United States	http://dx.doi.org/10.1016/j.jss.2019.08.004	article	"Blood was drawn from individuals (n = 213) before colonoscopy or from patients with nonmetastatic CRC (n = 50)"	Case-control study	AUC, sensitivity, specificity (training / test set split)	training + test set	"A major roadblock to reducing the mortality of colorectal cancer (CRC) is prompt detection and treatment, and a simple blood test is likely to have higher compliance than all of the current methods. The purpose of this report is to examine the utility of a mass spectrometry–based blood serum protein biomarker test for detection of CRC. A five-marker panel consisting of leucine-rich alpha-2-glycoprotein 1, epidermal growth factor receptor, inter-alpha-trypsin inhibitor heavy-chain family member 4, hemopexin, and superoxide dismutase 3 performed the best with 70% specificity at over 89% sensitivity (area under the curve = 0.86) in the validation set."	"A major roadblock to reducing the mortality of colorectal cancer (CRC) is prompt detection and treatment, and a simple blood test is likely to have higher compliance than all of the current methods. The purpose of this report is to examine the utility of a mass spectrometry–based blood serum protein biomarker test for detection of CRC. A five-marker panel consisting of leucine-rich alpha-2-glycoprotein 1, epidermal growth factor receptor, inter-alpha-trypsin inhibitor heavy-chain family member 4, hemopexin, and superoxide dismutase 3 performed the best with 70% specificity at over 89% sensitivity (area under the curve = 0.86) in the validation set."
82	Jalali, A and Pfeifer, N	Interpretable per case weighted ensemble method for cancer associations	BMC Genomics	17		501-501	2016	Germany	http://dx.doi.org/10.1186/s12864-016-2647-9	article	AML dataset: "194 samples contain methylation data and we use the part of the data measured by JHU-USC HumanMethylation450 arrays. 173 samples contain mRNA data measured by HG-U133 arrays.", "BRCA dataset: This data set includes 993 samples with clinical data. Only very few samples in this data set are indicated as having metastasized (8 samples). Hence the data are analyzed according to “tumor size”, “affected nearby lymph nodes”, “stage”, and “estrogen receptor”."	Cases only (risk & severity stratification)	AUC (training / test set split)	training + test set	"Molecular measurements from cancer patients such as gene expression and DNA methylation can be influenced by several external factors. If a model does not take potential biases in the data into account, this can lead to problems when trying to predict the stage of a certain cancer type. This is especially true when these biases differ between the training and test set. We introduce a method that can estimate this bias on a per-feature level and incorporate calculated feature confidences into a weighted combination of classifiers with disjoint feature sets. Moreover, we show how to visualize the learned classifiers to display interesting associations with the target label."	"Molecular measurements from cancer patients such as gene expression and DNA methylation can be influenced by several external factors. If a model does not take potential biases in the data into account, this can lead to problems when trying to predict the stage of a certain cancer type. This is especially true when these biases differ between the training and test set. We introduce a method that can estimate this bias on a per-feature level and incorporate calculated feature confidences into a weighted combination of classifiers with disjoint feature sets. Moreover, we show how to visualize the learned classifiers to display interesting associations with the target label."
83	Jones, J J and Wilcox, B E and Benz, R W and Babbar, N and Boragine, G and Burrell, T and Christie, E B and Croner, L J and Cun, P and Dillon, R and Kairs, S N and Kao, A and Preston, R and Schreckengaust, S R and Skor, H and Smith, W F and You, J and Hillis, W D and Agus, D B and Blume, J E	A Plasma-Based Protein Marker Panel for Colorectal Cancer Detection Identified by Multiplex Targeted Mass Spectrometry	Clin Colorectal Cancer	15	2	186-194.e13	2016	United States	http://dx.doi.org/10.1016/j.clcc.2016.02.004	article	the present study used 274 individual patient blood plasma samples, 137 with biopsy-confirmed colorectal cancer and 137 age- and gender-matched controls.	Case-control study	AUC, sensitivity, specificity (cross-validation + external test set)	cross-validation + external cohort validation	"Early colon cancer detection in patient populations ineligible for testing, such as the elderly or those with significant comorbidities, could have clinical benefit. A multiplex assay was developed for 187 candidate marker proteins, using 337 peptides monitored through 674 simultaneously measured MRM transitions in a 30-minute liquid chromatography-mass spectrometry analysis of immunodepleted blood plasma. To evaluate the combined candidate marker performance, the present study used 274 individual patient blood plasma samples, 137 with biopsy-confirmed colorectal cancer and 137 age- and gender-matched controls. Using one half of the data as a discovery set (69 disease cases and 69 control cases), the elastic net feature selection and random forest classifier assembly were used in cross-validation to identify a 15-transition classifier. The mean training receiver operating characteristic area under the curve was 0.82. After final classifier assembly using the entire discovery set, the 136-sample (68 disease cases and 68 control cases) validation set was evaluated. The validation area under the curve was 0.91."	"Early colon cancer detection in patient populations ineligible for testing, such as the elderly or those with significant comorbidities, could have clinical benefit. A multiplex assay was developed for 187 candidate marker proteins, using 337 peptides monitored through 674 simultaneously measured MRM transitions in a 30-minute liquid chromatography-mass spectrometry analysis of immunodepleted blood plasma. To evaluate the combined candidate marker performance, the present study used 274 individual patient blood plasma samples, 137 with biopsy-confirmed colorectal cancer and 137 age- and gender-matched controls. Using one half of the data as a discovery set (69 disease cases and 69 control cases), the elastic net feature selection and random forest classifier assembly were used in cross-validation to identify a 15-transition classifier. The mean training receiver operating characteristic area under the curve was 0.82. After final classifier assembly using the entire discovery set, the 136-sample (68 disease cases and 68 control cases) validation set was evaluated. The validation area under the curve was 0.91."
84	Jurmeister, P and Bockmayr, M and Seegerer, P and Bockmayr, T and Treue, D and Montavon, G and Vollbrecht, C and Arnold, A and Teichmann, D and Bressem, K and Schuller, U and von Laffert, M and Muller, K R and Capper, D and Klauschen, F	Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases	Science Translational Medicine	11	509	10-10	2019	Germany	http://dx.doi.org/10.1126/scitranslmed.aaw8513	article	408 patients with a history of primary HNSC and a synchronous or metachronous squamous lung tumor	Case-control study	AUC, accuracy (5-fold CV + external test set)	cross-validation + external cohort validation	"Head and neck squamous cell carcinoma (HNSC) patients are at risk of suffering from both pulmonary metastases or a second squamous cell carcinoma of the lung (LUSC). Differentiating pulmonary metastases from primary lung cancers is of high clinical importance, but not possible in most cases with current diagnostics. To address this, we performed DNA methylation profiling of primary tumors and trained three different machine learning methods to distinguish metastatic HNSC from primary LUSC. We developed an artificial neural network that correctly classi- fied 96.4% of the cases in a validation cohort of 279 patients with HNSC and LUSC as well as normal lung controls, outperforming support vector machines (95.7%) and random forests (87.8%)."	"Head and neck squamous cell carcinoma (HNSC) patients are at risk of suffering from both pulmonary metastases or a second squamous cell carcinoma of the lung (LUSC). Differentiating pulmonary metastases from primary lung cancers is of high clinical importance, but not possible in most cases with current diagnostics. To address this, we performed DNA methylation profiling of primary tumors and trained three different machine learning methods to distinguish metastatic HNSC from primary LUSC. We developed an artificial neural network that correctly classi- fied 96.4% of the cases in a validation cohort of 279 patients with HNSC and LUSC as well as normal lung controls, outperforming support vector machines (95.7%) and random forests (87.8%)."
85	Karimpour-Fard, A and Epperson, L E and Hunter, L E	A survey of computational tools for downstream analysis of proteomic and other omic datasets	Human Genomics	9		11-11	2015	USA	http://dx.doi.org/10.1186/s40246-015-0050-2	article	review (not applicable)	review			"In this paper, we review the well-known and ready-to-use tools for classification, clustering and validation, interpretation, and generation of biological information from experimental data. We suggest some rules of thumb for the reader on choosing the best suitable learning method for a particular dataset and conclude with pathway and functional analysis and then provide information about submitting final results to a repository."	"In this paper, we review the well-known and ready-to-use tools for classification, clustering and validation, interpretation, and generation of biological information from experimental data. We suggest some rules of thumb for the reader on choosing the best suitable learning method for a particular dataset and conclude with pathway and functional analysis and then provide information about submitting final results to a repository."
86	Khan, S R and Mohan, H and Liu, Y and Batchuluun, B and Gohil, H and Al Rijjal, D and Manialawy, Y and Cox, B J and Gunderson, E P and Wheeler, M B	The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes	Diabetologia	62	4	687-703	2019	Canada	https://link.springer.com/article/10.1007%2Fs00125-018-4800-2	article	55 incident cases matched to 85 non-case control participants	Case-control study	AUC, accuracy, sensitivity, specificity (45-fold cross-validation)	cross-validation	"Gestational diabetes mellitus (GDM) affects up to 20% of pregnancies, and almost half of the women affected progress to type 2 diabetes later in life, making GDM the most s+P182ignificant risk factor for the development of future type 2 diabetes. We used a well-characterised prospective cohort of women with a history of GDM pregnancy, all of whom were enrolled at 6–9 weeks postpartum (baseline), were confirmed not to have diabetes via 2 h 75 g OGTT and tested anually for type 2 diabetes on an ongoing basis (2 years of follow-up). A large-scale targeted lipidomic study was implemented to analyse ~1100 lipid metabolites in baseline plasma samples [...]. Machine learning optimisation in a decision tree format revealed a seven-lipid metabolite type 2 diabetes predictive signature with a discriminating power (AUC) of 0.92 (87% sensitivity, 93% specificity and 91% accuracy)."	"Gestational diabetes mellitus (GDM) affects up to 20% of pregnancies, and almost half of the women affected progress to type 2 diabetes later in life, making GDM the most s+P182ignificant risk factor for the development of future type 2 diabetes. We used a well-characterised prospective cohort of women with a history of GDM pregnancy, all of whom were enrolled at 6–9 weeks postpartum (baseline), were confirmed not to have diabetes via 2 h 75 g OGTT and tested anually for type 2 diabetes on an ongoing basis (2 years of follow-up). A large-scale targeted lipidomic study was implemented to analyse ~1100 lipid metabolites in baseline plasma samples [...]. Machine learning optimisation in a decision tree format revealed a seven-lipid metabolite type 2 diabetes predictive signature with a discriminating power (AUC) of 0.92 (87% sensitivity, 93% specificity and 91% accuracy)."
87	Khusial, R D and Cioffi, C E and Caltharp, S A and Krasinskas, A M and Alazraki, A and Knight-Scott, J and Cleeton, R and Castillo-Leon, E and Jones, D P and Pierpont, B and Caprio, S and Santoro, N and Akil, A and Vos, M B	Development of a Plasma Screening Panel for Pediatric Nonalcoholic Fatty Liver Disease Using Metabolomics	Hepatology Communications	3	10	1311-1321	2019	United States	http://dx.doi.org/10.1002/hep4.1417	article	subjects with NAFLD (n = 222) and without NAFLD (n = 337)	Case-control study	AUC (training set: 2/3 of data, test set 1/3 of data)	training + test set	"Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease in children, but diagnosis is challenging due to limited availability of noninvasive biomarkers. Machine learning applied to high-resolution metabolomics and clinical phenotype data offers a novel framework for developing a NAFLD screening panel in youth. Here, untargeted metabolomics by liquid chromatography–mass spectrometry was performed on plasma samples from a combined cross-sectional sample of children and adolescents ages 2-25 years old with NAFLD (n = 222) and without NAFLD (n = 337) [...]. The highest performing classification model was random forest, which had an area under the receiver operating characteristic curve (AUROC) of 0.94, sensitivity of 73%, and specificity of 97% for detecting NAFLD cases. A second classification model was developed using the homeostasis model assessment of insulin resistance substituted for the WBISI. Similarly, the highest performing classification model was random forest, which had an AUROC of 0.92, sensitivity of 73%, and specificity of 94%"	"Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease in children, but diagnosis is challenging due to limited availability of noninvasive biomarkers. Machine learning applied to high-resolution metabolomics and clinical phenotype data offers a novel framework for developing a NAFLD screening panel in youth. Here, untargeted metabolomics by liquid chromatography–mass spectrometry was performed on plasma samples from a combined cross-sectional sample of children and adolescents ages 2-25 years old with NAFLD (n = 222) and without NAFLD (n = 337) [...]. The highest performing classification model was random forest, which had an area under the receiver operating characteristic curve (AUROC) of 0.94, sensitivity of 73%, and specificity of 97% for detecting NAFLD cases. A second classification model was developed using the homeostasis model assessment of insulin resistance substituted for the WBISI. Similarly, the highest performing classification model was random forest, which had an AUROC of 0.92, sensitivity of 73%, and specificity of 94%"
88	Kim, J and Da Rosa, J C and Lee, J and Tomalin, L and Lowes, M A and Fitz, L and Bernstein, G and Valdez, H and Wolk, R and Krueger, J G and Suárez-Fariñas, M	Precision medicine in psoriasis: Machine learning and proteomics join forces to develop a blood-based test to predict response to tofacitinib or Etanercept in psoriasis patients	Experimental Dermatology	25		49-50	2016	United States	http://dx.doi.org/10.1111/exd.13200	meeting abstract	259 serum samples from a phase 3 study in adults with moderate-to- severe psoriasis	Case-control study	AUC, accuracy (training / test set split: 80% / 20%)	training + test set	"Despite various recent FDA-approved treatments 20%-30% of psoriatic patients fail to respond to biologics. [...] we applied machine-learning algorithms to proteomic data obtained using a proximity extension assay [...] to develop a blood-based test to predict response to tofacitinib or Etanercept in psoriasis patients. he elastic-net algorithm, using only pre-treatment data, was the best performer among the methods evaluated, with average AUC values (over 500 random 20/80 data splits) of 90.8% and 91% and accuracies of 88% and 87% for tofacitinib and Etanercept, respectively."	"Despite various recent FDA-approved treatments 20%-30% of psoriatic patients fail to respond to biologics. [...] we applied machine-learning algorithms to proteomic data obtained using a proximity extension assay [...] to develop a blood-based test to predict response to tofacitinib or Etanercept in psoriasis patients. he elastic-net algorithm, using only pre-treatment data, was the best performer among the methods evaluated, with average AUC values (over 500 random 20/80 data splits) of 90.8% and 91% and accuracies of 88% and 87% for tofacitinib and Etanercept, respectively."
89	Kim, M and Oh, I and Ahn, J	An improved method for prediction of cancer prognosis by network learning	Genes	9	10	1.-11	2018	Switzerland	http://dx.doi.org/10.3390/genes9100478	article	"First, we downloaded gene mRNA data, CNV data, DNA methylation data, SNP data, and clinical data for PAAD, BRCA, KIRC, LGG, and STAD from The Cancer Genome Atlas (TCGA)" (more than 50 samples per group for multiple datasets)	Case-control study	AUC, accuracy (10-fold CV)	cross-validation	"In this paper, we propose a novel machine learning-based method for more accurate identification of prognostic biomarker genes and use them for prediction of cancer prognosis. The proposed method specifies the candidate prognostic gene module by graph learning using the generative adversarial networks (GANs) model, and scores genes using a PageRank algorithm. We applied the proposed method to multiple-omics data that included copy number, gene expression, DNA methylation, and somatic mutation data for five cancer types. The proposed method showed better prediction accuracy than did existing methods."	"In this paper, we propose a novel machine learning-based method for more accurate identification of prognostic biomarker genes and use them for prediction of cancer prognosis. The proposed method specifies the candidate prognostic gene module by graph learning using the generative adversarial networks (GANs) model, and scores genes using a PageRank algorithm. We applied the proposed method to multiple-omics data that included copy number, gene expression, DNA methylation, and somatic mutation data for five cancer types. The proposed method showed better prediction accuracy than did existing methods."
90	Kim, S and Jhong, J H and Lee, J and Koo, J Y	Meta-analytic support vector machine for integrating multiple omics data	BioData Mining	10	1		2017	South Korea	http://dx.doi.org/10.1186/s13040-017-0126-8	article	"we apply the Meta-SVM methods to two real examples of idiopathic pulmonary fibrosis expression profiles (IPF; 221 samples in four studies of binary outcome (i.e., case and control)) and breast cancer expression profiles provided by The Cancer Genome Atlas (TCGA) including mRNA, copy number variation (CNV) and epigenetic DNA methylation (http://cancergenome.nih.gov/; 300 samples of estrogen receptor binary outcome (i.e., ER+ and ER-))"	Case-control study	sensitivity, specificity (cross-validation)	cross-validation	"We propose a meta-analytic support vector machine (Meta-SVM) that can accommodate multiple omics data, making it possible to detect consensus genes associated with diseases across studies. Experimental studies show that the Meta-SVM is superior to the existing meta-analysis method in detecting true signal genes. In real data applications, diverse omics data of breast cancer (TCGA) and mRNA expression data of lung disease (idiopathic pulmonary fibrosis; IPF) were applied. As a result, we identified gene sets consistently associated with the diseases across studies."	"We propose a meta-analytic support vector machine (Meta-SVM) that can accommodate multiple omics data, making it possible to detect consensus genes associated with diseases across studies. Experimental studies show that the Meta-SVM is superior to the existing meta-analysis method in detecting true signal genes. In real data applications, diverse omics data of breast cancer (TCGA) and mRNA expression data of lung disease (idiopathic pulmonary fibrosis; IPF) were applied. As a result, we identified gene sets consistently associated with the diseases across studies."
91	Kim, S and Lin, C W and Tseng, G C	MetaKTSP: a meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis	Bioinformatics	32	13	1966-1973	2016	South Korea	http://dx.doi.org/10.1093/bioinformatics/btw115	article	"we demonstrate application of MetaKTSP methods to three real omics examples of breast cancer expression profiles (1658 samples in seven studies), IPF expression profiles (IPF; 291 samples in six studies) and The Cancer Genome Atlas multi-cancer methylation profiles (TCGA, http://cancergenome.nih.gov/; 1785 samples in six studies)"	Case-control study	Youden index (5-fold cross-validation)	cross-validation	"The purpose of this article is to develop a meta-analytic top scoring pair (MetaKTSP) framework that combines multiple transcriptomic studies and generates a robust prediction model applicable to independent test studies. We proposed two frameworks, by averaging TSP scores or by combining P-values from individual studies, to select the top gene pairs for model construction. We applied the proposed methods in simulated data sets and three large-scale real applications in breast cancer, idiopathic pulmonary fibrosis and pan-cancer methylation. The result showed superior performance of cross-study validation accuracy and biomarker selection for the new meta-analytic framework."	"The purpose of this article is to develop a meta-analytic top scoring pair (MetaKTSP) framework that combines multiple transcriptomic studies and generates a robust prediction model applicable to independent test studies. We proposed two frameworks, by averaging TSP scores or by combining P-values from individual studies, to select the top gene pairs for model construction. We applied the proposed methods in simulated data sets and three large-scale real applications in breast cancer, idiopathic pulmonary fibrosis and pan-cancer methylation. The result showed superior performance of cross-study validation accuracy and biomarker selection for the new meta-analytic framework."
92	Kim, S Y and Diggans, J and Pankratz, D and Huang, J and Pagan, M and Sindy, N and Tom, E and Anderson, J and Choi, Y and Lynch, D A and Steele, M P and Flaherty, K R and Brown, K K and Farah, H and Bukstein, M J and Pardo, A and Selman, M and Wolters, P J and Nathan, S D and Colby, T V and Myers, J L and Katzenstein, A L A and Raghu, G and Kennedy, G C	Classification of usual interstitial pneumonia in patients with interstitial lung disease: assessment of a machine learning approach using high-dimensional transcriptional data	Lancet Respiratory Medicine	3	6	473-482	2015	Mexico	http://dx.doi.org/10.1016/s2213-2600(15)00140-x	article	"125 surgical lung biopsies from 86 patients. 58 samples were identified by the expert panel as usual interstitial pneumonia, 23 as non-specific interstitial pneumonia, 16 as hypersensitivity pneumonitis, four as sarcoidosis, four as respiratory bronchiolitis, two as organising pneumonia, and 18 as subtypes other than usual interstitial pneumonia"	Case-control study	sensitivity, specificity (training/test set split)	training + test set	"Idiopathic pulmonary fibrosis is a progressive fibrotic lung disease that distorts pulmonary architecture, leading to hypoxia, respiratory failure, and death. Diagnosis is difficult because other interstitial lung diseases have similar radiological and histopathological characteristics. We aimed to develop a molecular test that distinguishes usual interstitial pneumonia from other interstitial lung diseases in surgical lung biopsy samples. [...] The microarray classifier was trained on 77 samples and was assessed in a test set of 48 samples, for which it had a specificity of 92% (95% CI 81–100) and a sensitivity of 82% (64–95)."	"Idiopathic pulmonary fibrosis is a progressive fibrotic lung disease that distorts pulmonary architecture, leading to hypoxia, respiratory failure, and death. Diagnosis is difficult because other interstitial lung diseases have similar radiological and histopathological characteristics. We aimed to develop a molecular test that distinguishes usual interstitial pneumonia from other interstitial lung diseases in surgical lung biopsy samples. [...] The microarray classifier was trained on 77 samples and was assessed in a test set of 48 samples, for which it had a specificity of 92% (95% CI 81–100) and a sensitivity of 82% (64–95)."

93	Kim, Y and Bismeijer, T and Zwart, W and Wessels, L F A and Vis, D J	Genomic data integration by WON-PARAFAC identifies interpretable factors for predicting drug-sensitivity in vivo	Nat Commun	10	1	5034-5034	2019	Netherlands	http://dx.doi.org/10.1038/s41467-019-13027-2	article	1815 genes by 935 cell line	Cases only (drug sensitivity prediction)	AUC (10-fold CV)	cross-validation	"We introduced Weighted Orthogonal Nonnegative parallel factor analysis (WON-PARAFAC), a data integration method that identifies sparse and interpretable factors. WON-PARAFAC summarizes the GDSC1000 cell line compendium in 130 factors. We interpret the factors based on their association with recurrent molecular alterations, pathway enrichment, cancer type, and drug-response. Crucially, the cell line derived factors capture the majority of the relevant biological variation in Patient-Derived Xenograft (PDX) models, strongly suggesting our factors capture invariant and generalizable aspects of cancer biology. Furthermore, drug response in cell lines is better and more consistently translated to PDXs using factor-based predictors as compared to raw feature-based predictors."
94	Kim, Y R and Kim, D and Kim, S Y	Prediction of Acquired Taxane Resistance Using a Personalized Pathway-Based Machine Learning Method	Cancer Res Treat	51	2	672-684	2019	Korea (South)	http://dx.doi.org/10.4143/crt.2018.137	article	more than 50 samples per group for most human cancer cell line datasets considered	Cases only (drug response prediction in-vitro)	AUC (LOOCV)	cross-validation	"Taxanes, notably paclitaxel (PTX) and docetaxel (DTX), are cytotoxic microtubule-stabilizing agents used in various types of cancers, including gynaecological cancers (ovarian, cervical, and endometrial cancer) and breast cancers, with proven survival benefits [1]. Intrinsic or acquired resistance (AR) to chemotherapy are major clinical obstacles, resulting in poor response and lower overall survival rates. In this study, we developed and validated a highly accurate multi-study–derived, multivariable predictive model for ATR using personalized pathways and sophisticated machine learning algorithms."
95	Kirkgoz, T and Kilic, S and Abali, Z Y and Yaman, A and Kaygusuz, S B and Eltan, M and Turan, S and Haklar, G and Sagiroglu, M S and Bereket, A and Guran, T	Simplifying the interpretation of steroid metabolome data by a machine-learning approach	Hormone Research in Paediatrics	91		128-128	2019		http://dx.doi.org/10.1159/000501868	article	500 healthy controls and 427 treatment-naive children with a disorder of adrenal steroidogenesis	Case-control study	sensitivity, specificity (10-fold cross-validation)	cross-validation	"Liquid chromatography-mass spectrometry (LC-MS) based panels of steroid hormones and their precursors offer a distinct pattern of steroid metabolome for various disorders of adrenal and gonadal steroidogenesis. We have implemented a machine-learning algorithm for a time-saving and experience-independent review and interpretation of analytical results. We have tested the performance of this algorithm using our archived data of quantitation of 16 steroid hormones and precursors by an LC-MS/MS based panel in 500 healthy controls and 427 treatment-naive children with a disorder of adrenal steroidogenesis. For discrimination of patients from the healthy controls; the sensitivity and specificity of the RUSBoost algorithm was 97.7% and 92.6%, respectively. The differentiation of each disorder could be achieved with overall accuracy of up to 95% independent of age and sex."
96	Kitazawa, H and Muramatsu, H and Murakami, N and Okuno, Y and Wakamatsu, M and Yoshida, T and Inaya, M and Yamamori, A and Ishiwata, S and Narita, K and Hamada, M and Ichikawa, D and Taniguchi, R and Kawashima, N and Nishikawa, E and Narita, A and Nishio, N and Kojima, S and Takahashi, Y	Genome-wide methylation analysis using the digital restriction enzyme analysis of methylation for stratification of patients with juvenile myelomonocytic leukemia	Blood	134			2019		http://dx.doi.org/10.1182/blood-2019-127792	meeting abstract	99 children (67 boys and 32 girls) with JMML.	Case-control study	accuracy (training / test set split)	training + test set	"Juvenile myelomonocytic leukemia (JMML) is a rare myelodysplastic/myeloproliferative neoplasm that occurs during infancy and early childhood. We studied 99 children (67 boys and 32 girls) with JMML. We performed unsupervised consensus clustering with DREAM methylation data of 7,704 CpG sites within ±1 kb from TSS on autosomal chromosomes detected in ≥95% of the samples with imputation of the missing data using the median of each CpG site methylation level. Clustering identified two distinct subgroups, the HM subgroup (n = 35) and the LM subgroup (n = 64), matching 95% (94 of 99) with the 450K clustering results. The HM subgroup patients showed significantly poorer 5-year OS than the LM subgroup patients (41.9% [95% confidence interval (CI)], 25.3%-57.6%) vs. 71.4% [95% CI, 56.2%-82.1%); P = 0.0034). Then, we developed a prediction model of the methylation subgroups using a machine-learning program. Both the unsupervised clustering analysis and SVM model could repeat the result of 450K-based methylation classification, i.e., the HM and LM subgroups."
97	Kong, A and Azencott, R	Binary Markov Random Fields and interpretable mass spectra discrimination	Statistical Applications in Genetics and Molecular Biology	16	1	13-30	2017	Germany	http://dx.doi.org/10.1515/sagmb-2016-0019	article	"A dataset of 238 MALDI colorectal mass spectra and two datasets of 216 and 253 SELDI ovarian mass spectra respectively were used to test our approach."	Case-control study	accuracy (LOOCV)	cross-validation	"For mass spectra acquired from cancer patients by MALDI or SELDI techniques, automated discrimination between cancer types or stages has often been implemented by machine learning algorithms. Nevertheless, these techniques typically lack interpretability in terms of biomarkers. In this paper, we propose a new mass spectra discrimination algorithm by parameterized Markov Random Fields to automatically generate interpretable classifiers with small groups of scored biomarkers. The results show that our approach reaches accuracies of 81% to 100% to discriminate between patients from different colorectal and ovarian cancer stages, and performs as well or better than previous studies on similar datasets. Moreover, our approach enables efficient planar-displays to visualize mass spectra discrimination and has good asymptotic performance for large datasets."
98	Krawczuk, J and Łukaszuk, T	The feature selection bias problem in relation to high-dimensional gene data	Artif Intell Med	66		63-71	2016	Poland	http://dx.doi.org/10.1016/j.artmed.2015.11.001	article	seven microarray datasets with > 50 samples group for multiple datasets were used	Case-control study	accuracy (double LOOCV)	cross-validation	"In this paper, we consider feature selection for the classification of gene datasets. For this kind of data, it is easy to find a model that fits the learning data. However, it is not easy to find one that will simultaneously evaluate new data equally well as learning data. This overfitting issue is well known as regards classification and regression, but it also applies to feature selection. We address this problem and investigate its importance in an empirical study of four feature selection methods applied to six high-dimensional gene datasets. Our main result reveals the existence of positive feature selection bias in all 28 experiments (7 datasets and 4 feature selection methods). This work provides evidence that using the same dataset for feature selection and learning is not appropriate. We recommend using cross-validation for feature selection in order to reduce selection bias."
99	Krittanawong, C and Bomback, A S and Baber, U and Bangalore, S and Messerli, F H and Wilson Tang, W H	Future Direction for Using Artificial Intelligence to Predict and Manage Hypertension	Curr Hypertens Rep	20	9	75-75	2018	Poland	http://dx.doi.org/10.1007/s11906-018-0875-x	article	review (not applicable)	review			"This review summarizes recent advances in the computer science and medical field, illustrating the innovative AI approach for potential prediction of early stages of hypertension. Additionally, we review ongoing research and future implications of AI in hypertension management and clinical trials, with an eye towards personalized medicine."
100	Kuo, C H S and Pavlidis, S and Loza, M and Baribaud, F and Rowe, A and Pandis, I and Rossios, C and Wilson, S and Djukanovic, R and Sterk, P and Chung, K F and Adcock, I M and Guo, Y	Asthma phenotypes from semi-supervised machine-learning approach of bronchial biopsy and brush transcriptomics in U-biopred	American Journal of Respiratory and Critical Care Medicine	191			2015		http://www.atsjournals.org/doi/pdf/10.1164/ajrccm-conference.2015.191.1_MeetingAbstracts.A2205	meeting abstract	"Subjects with moderate-to-severe asthma recruited in the U-BIOPRED study underwent fiberoptic bronchoscopy for bronchial biopsy (91) and brush (105) samples"	Case-control study	accuracy (cross-validation)	cross-validation	"In this work, four state-of-the-art Random Forest-based feature selection methods were compared in a gene selection context. The analysis focused on the stability of selection because, although it is necessary for determining the significance of results, it is often ignored in similar studies. The comparison of post-selection accuracy of a validation of Random Forest classifiers revealed that all investigated methods were equivalent in this context. However, the methods substantially differed with respect to the number of selected genes and the stability of selection. Of the analysed methods, the Boruta algorithm predicted the most genes as potentially important. When the number of consistently selected genes was considered, the Boruta algorithm was clearly the best. Although it was also the most computationally intensive method, the Boruta algorithm's computational demands could be reduced to levels comparable to those of other algorithms by replacing the Random Forest importance with a comparable measure from Random Ferns (a similar but simplified classifier). Despite their design assumptions, the minimal optimal selection methods, were found to select a high fraction of false positives."
101	Kursa, M B	Robustness of Random Forest-based gene selection methods	BMC Bioinformatics	15		8-8	2014	Poland	http://dx.doi.org/10.1186/1471-2105-15-8	article	4 microarray datasets were used, one contained > 50 samples per group	Case-control study	error-rate (training / test set split)	training + test set	"As the worldwide prevalence of colorectal cancer (CRC) is increasing, it is of vital importance to reduce its morbidity and mortality by early detection. Saliva is a noninvasively accessible fluid that potentially reflects both oral and systemic diseases. [...] Biomarkers for distinguishing subjects with CRC from the others, as well as metabolites that normalize whole saliva concentration were identified. Analysis of these metabolites using machine learning-based artificial intelligence showed a high area under the receiver operating characteristic curve (AUC 1/4 0.876; P < 0.0001) in the training data- set. This combination also showed high AUC values using the validation dataset (AUC 1/4 0.861; P < 0.0001)."
102	Kuwabara, H and Iwabuthi, A and Soya, R and Enomoto, M and Ishizaki, T and Tsuchida, A and Nagakawa, Y and Katsumata, K and Sugimoto, M	Salivary metabolomics for colorectal cancer detection	Annals of Oncology	30		v46-v46	2019		http://dx.doi.org/10.1093/annonc/mdz239.058	article	"231 subjects with CRC, 99 subjects with polyps, and 2272 subjects with healthy controls"	Case-control study	AUC (training / test set split)	training + test set	"We applied a fuzzy logic selection on seven breast cancer microarray databases to obtain new gene signatures. To validate these gene signatures, we designed probes for the selected genes on Nimblegen custom microarrays and tested them on a series of 151 consecutive invasive breast carcinomas displaying clinicopathological features similar to those observed in routine practice. Results: Using fuzzy logic selection, we identified new gene lists that were included in our Nimblegen in-house microarray (1100 genes,17000 probes). Analysis in the training public sets showed good performance for the 12 newly generated gene signatures for grade (sensitivity ranging from 80% to 100%, specificity 80% to 97%, error rate 3 to 12%). Similar results were obtained in our validation set of 151 breast cancer samples, with an error rate of less than 10% in the majority of the gene lists tested."
103	Lacroix-Triki, M and Kempowsky-Hamon, T and Valle, C and Hedjazi, L and Lamarre, S and Trouilh, L and Puydenus, L and Mhamdi, L and Dalenc, F and Filleron, T and Favre, G and Le Lann, M V and Le Berre-Anton, V	Fuzzy logic selection as a new reliable tool to identify gene signatures in breast cancer - the INNODIAG Study	Laboratory Investigation	93		51A-51A	2013		http://dx.doi.org/10.1038/labinvest.2013.14	meeting abstract	7 breast cancer microarray datasets + 151 consecutive invasive breast carcinomas	Case-control study	sensitivity, specificity, error rate (training / test set split)	training + test set	"The differential diagnosis between primary epithelial ovarian and endometrial cancers is often unresolved because the histologic subtypes of these two tumor types can have very similar histology and immunohistochemical appearance. Here we report the development and validation of a gene expression profile (GEP) diagnostic test (Pathwork Tissue of Origin Endometrial Test) that distinguishes ovarian and endometrial cancers in formalin-fixed, paraffin-embedded (FFPE) specimens using a 316-gene classification model. The Tissue of Origin Endometrial Test accurately identified the primary site for 94.7% (95% CI, 87% to 99%) of ovarian and endometrial cancers. Other measures of test performance include an area under the ROC curve of 0.997 and a diagnostic odds ratio of 406. Test performance did not change significantly when stratified by specimens from metastases (90.5%) or by poorly differentiated and undifferentiated primary tumors (96.3%)."
104	Lal, A and Panos, R and Marjanovic, M and Walker, M and Fuentes, E and Kapp, D S and Henner, W D and Buturovic, L J and Miller, M H	A gene expression profile test that distinguishes ovarian from endometrial cancers	Journal of Clinical Oncology	30	15		2012		https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326651	meeting abstract	75 metastatic, poorly differentiated or undifferentiated primary FFPE tumor specimens	Differential diagnosis prediction	AUC (training / test set split)	training + test set	"The differential diagnosis between primary epithelial ovarian and endometrial cancers is often unresolved because the histologic subtypes of these two tumor types can have very similar histology and immunohistochemical appearance. Here we report the development and validation of a gene expression profile (GEP) diagnostic test (Pathwork Tissue of Origin Endometrial Test) that distinguishes ovarian and endometrial cancers in formalin-fixed, paraffin-embedded (FFPE) specimens using a 316-gene classification model. The Tissue of Origin Endometrial Test accurately identified the primary site for 94.7% (95% CI, 87% to 99%) of ovarian and endometrial cancers. Other measures of test performance include an area under the ROC curve of 0.997 and a diagnostic odds ratio of 406. Test performance did not change significantly when stratified by specimens from metastases (90.5%) or by poorly differentiated and undifferentiated primary tumors (96.3%)."

105	Lawton, K A and Brown, M V and Alexander, D and Li, Z and Wulff, J E and Lawson, R and Jaffa, M and Milburn, M V and Ryals, J A and Bowser, R and Cudkowicz, M E and Berry, J D	Plasma metabolomic biomarker panel to distinguish patients with amyotrophic lateral sclerosis from disease mimics	Amyotroph Lateral Scler Frontotemp oral Degener	15	5	362-370	2014	England	http://dx.doi.org/10.3109/21678421.2014.908311	article	172 patients recently diagnosed with ALS, 50 healthy controls, and 73 neurological disease mimics	Case-control study	AUC, sensitivity, specificity (training / test set split)	training + test set	"Our objective was to identify plasma biomarkers of ALS that can aid in distinguishing patients with ALS from those with disease mimics. Using all identified biochemicals detected in > 50% of all samples in the metabolomics analysis, samples were classified as ALS or mimic with 65% sensitivity and 81% specificity by LASSO analysis (AUC of 0.76). A subset panel of 32 candidate biomarkers classified these diagnosis groups with a specificity of 90%/sensitivity 58% (AUC of 0.81)."	"Our objective was to identify plasma biomarkers of ALS that can aid in distinguishing patients with ALS from those with disease mimics. Using all identified biochemicals detected in > 50% of all samples in the metabolomics analysis, samples were classified as ALS or mimic with 65% sensitivity and 81% specificity by LASSO analysis (AUC of 0.76). A subset panel of 32 candidate biomarkers classified these diagnosis groups with a specificity of 90%/sensitivity 58% (AUC of 0.81)."
106	Le, T T and Blackwood, N O and Taroni, J N and Fu, W and Breitenstein, M K	Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients	AMIA Annu Symp Proc	2018		1358-1367	2018	USA	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371296/	article	The SLE compendium contained 15,497 gene expression measurements with observations from healthy control (n=160) samples, treatment-naïve SLE (n=1,290) samples, and SLE samples exposed to various treatments (n =126)	Case-control + treatment response	balanced accuracy (cross-validation + 20% hold-out test set)	cross-validation + test set	"Within a compendium of systemic lupus erythematosus (SLE) patients, we applied the Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB) to profile de novo gene expression features affecting CD20, CD22 and CD30 gene aberrance. Utilizing carefully aggregated secondary data and leveraging a priori hypotheses, i-mAB fostered robust biomarker profiling among interdependent biological features."	"Within a compendium of systemic lupus erythematosus (SLE) patients, we applied the Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB) to profile de novo gene expression features affecting CD20, CD22 and CD30 gene aberrance. Utilizing carefully aggregated secondary data and leveraging a priori hypotheses, i-mAB fostered robust biomarker profiling among interdependent biological features."
107	Leclercq, M and Vittrant, B and Martin-Magniette, M L and Scott Boyer, M P and Perin, O and Bergeron, A and Fradet, Y and Droit, A	Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data	Frontiers in Genetics	10			2019	Canada	http://dx.doi.org/10.3389/fgene.2019.00452	article	five microarray datasets were used, including datasets with > 50 samples per group	Case-control study	accuracy (ACC), balanced error rate (BER), Matthew's correlation coefficient (MCC), area under the curve (AUC), sensitivity, specificity, Root Mean Squared Error (RMSE), Correlation Coefficient (CC) (10-fold CV)	cross-validation	"The identification of biomarker signatures in omics molecular profiling is usually performed to predict outcomes in a precision medicine context, such as patient disease susceptibility, diagnosis, prognosis, and treatment response. To identify these signatures, we have developed a biomarker discovery tool, called BioDiscML. From a collection of samples and their associated characteristics, i.e., the biomarkers (e.g., gene expression, protein levels, clinico-pathological data), BioDiscML exploits various feature selection procedures to produce signatures associated to machine learning models that will predict efficiently a specified outcome. To this purpose, BioDiscML uses a large variety of machine learning algorithms to select the best combination of biomarkers for predicting categorical or continuous outcomes from highly unbalanced datasets."	"The identification of biomarker signatures in omics molecular profiling is usually performed to predict outcomes in a precision medicine context, such as patient disease susceptibility, diagnosis, prognosis, and treatment response. To identify these signatures, we have developed a biomarker discovery tool, called BioDiscML. From a collection of samples and their associated characteristics, i.e., the biomarkers (e.g., gene expression, protein levels, clinico-pathological data), BioDiscML exploits various feature selection procedures to produce signatures associated to machine learning models that will predict efficiently a specified outcome. To this purpose, BioDiscML uses a large variety of machine learning algorithms to select the best combination of biomarkers for predicting categorical or continuous outcomes from highly unbalanced datasets."
108	Lee, S S and Attwood, K and Roder, H and Asmellash, S and Meyer, K and Kakolyris, S and Oliveira, C and Roder, J and Grigorieva, J and Chelis, L and Iyer, R and Mahalingam, D	An independent validation of a screening test using mass spectrometry for detection of hepatocellular carcinoma	Cancer Research	79	13		2019		http://dx.doi.org/10.1158/1538-7445.SABCS18-4530	meeting abstract	156 pts (97 HCC, 59 non-HCC healthy controls)	Case-control study	AUC (training and validation cohort)	external cohort validation	"Early detection is critical to improve outcome in hepatocellular carcinoma (HCC). A test to detect HCC in a high-risk population from 10 uL serum, combining MALDI mass spectrometry and AFP data was developed using a dropout-regularized hierarchical machine learning approach designed to minimize overfitting in small development sets. It was previously validated in 293 high risk pts (158 HCC, 135 non-HCC) with SS/SP of 83%/84% in development and 81%/79% in validation across various etiologies and Child-Pugh classification [...]. In independent validation, AUC for the test output prior to thresholding was 0.979, significantly better than AFP AUC 0.915 (P<0.001)."	"Early detection is critical to improve outcome in hepatocellular carcinoma (HCC). A test to detect HCC in a high-risk population from 10 uL serum, combining MALDI mass spectrometry and AFP data was developed using a dropout-regularized hierarchical machine learning approach designed to minimize overfitting in small development sets. It was previously validated in 293 high risk pts (158 HCC, 135 non-HCC) with SS/SP of 83%/84% in development and 81%/79% in validation across various etiologies and Child-Pugh classification [...]. In independent validation, AUC for the test output prior to thresholding was 0.979, significantly better than AFP AUC 0.915 (P<0.001)."
109	Lin, X and Afsari, B and Marchionni, L and Cope, L and Parmigiani, G and Naiman, D and Geman, D	The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations	BMC Bioinformatics	10		256-256	2009	USA	http://dx.doi.org/10.1186/1471-2105-10-256	article	118 samples for BRCA1 breast cancers + three datasets used for the ER status cross-study validation, including a dataset with > 50 samples per group	Case-control study	accuracy, sensitivity, specificity (LOOCV, cross-study validation)	cross-validation	"We present a three-gene version of "relative expression analysis" (RXA), a rigorous and systematic comparison with earlier approaches in a variety of cancer studies, a clinically relevant application to predicting germline BRCA1 mutations in breast cancer and a cross-study validation for predicting ER status. In the BRCA1 study, RXA yields high accuracy with a simple decision rule: in tumors carrying mutations, the expression of a "reference gene" falls between the expression of two differentially expressed genes, PPP1CB and RNF14."	"We present a three-gene version of "relative expression analysis" (RXA), a rigorous and systematic comparison with earlier approaches in a variety of cancer studies, a clinically relevant application to predicting germline BRCA1 mutations in breast cancer and a cross-study validation for predicting ER status. In the BRCA1 study, RXA yields high accuracy with a simple decision rule: in tumors carrying mutations, the expression of a "reference gene" falls between the expression of two differentially expressed genes, PPP1CB and RNF14."
110	Liu, F and Xing, L and Zhang, X and Zhang, X	A Four-Pseudogene Classifier Identified by Machine Learning Serves as a Novel Prognostic Marker for Survival of Osteosarcoma	Genes (Basel)	10	6		2019	China	http://dx.doi.org/10.3390/genes10060414	article	94 osteosarcoma patients	Cases only (survival prediction)	AUC (10-fold CV)	cross-validation	"Osteosarcoma is a common malignancy with high mortality and poor prognosis due to lack of predictive markers. The aim of this study was to identify a prognostic pseudogene signature of osteosarcoma by machine learning. A sample of 94 osteosarcoma patients' RNA-Seq data with clinical follow-up information was involved in the study. The survival-related pseudogenes were screened and related signature model was constructed by cox-regression analysis (univariate, lasso, and multivariate). In total, 125 survival-related pseudogenes were identified and a four-pseudogene (RPL11-551L14.1, HR: 0.65 (95% CI: 0.44–0.95); RPL7AP28, HR: 0.32 (95% CI: 0.14–0.76); RP4-706A16.3, HR: 1.89 (95% CI: 1.35–2.65); RP11-326A19.5, HR: 0.52(95% CI: 0.37–0.74)) signature effectively distinguished the high- and low-risk patients, and predicted prognosis with high sensitivity and specificity (AUC: 0.878)."	"Osteosarcoma is a common malignancy with high mortality and poor prognosis due to lack of predictive markers. The aim of this study was to identify a prognostic pseudogene signature of osteosarcoma by machine learning. A sample of 94 osteosarcoma patients' RNA-Seq data with clinical follow-up information was involved in the study. The survival-related pseudogenes were screened and related signature model was constructed by cox-regression analysis (univariate, lasso, and multivariate). In total, 125 survival-related pseudogenes were identified and a four-pseudogene (RPL11-551L14.1, HR: 0.65 (95% CI: 0.44–0.95); RPL7AP28, HR: 0.32 (95% CI: 0.14–0.76); RP4-706A16.3, HR: 1.89 (95% CI: 1.35–2.65); RP11-326A19.5, HR: 0.52(95% CI: 0.37–0.74)) signature effectively distinguished the high- and low-risk patients, and predicted prognosis with high sensitivity and specificity (AUC: 0.878)."
111	Liu, L and Liu, Y and Liu, C and Zhang, Z and Du, Y and Zhao, H	Analysis of gene expression profile identifies potential biomarkers for atherosclerosis	Mol Med Rep	14	4	3052-3058	2016	China	http://dx.doi.org/10.3892/mmr.2016.5650	article	A total of 118 samples from the peripheral blood of females, including 47 atherosclerotic and 71 non-atherosclerotic patients, was used for expression profiling.	Case-control study	AUC (5-fold CV)	cross-validation	"The present study aimed to identify potential biomarkers for atherosclerosis via analysis of gene expression profiles. [...] the RFE algorithm was used to identify 11 biomarkers, whose receiver operating characteristic curves had an area under curve of 0.92, indicating that the identified 11 biomarkers were representative."	"The present study aimed to identify potential biomarkers for atherosclerosis via analysis of gene expression profiles. [...] the RFE algorithm was used to identify 11 biomarkers, whose receiver operating characteristic curves had an area under curve of 0.92, indicating that the identified 11 biomarkers were representative."
112	Liu, M C and Jamshidi, A and Venn, O and Fields, A P and Maher, M C and Cann, G and Amini, H and Gross, S and Bredno, J and Miller, M and Schellenberger, J and Kurtzman, K N and Fung, E T and Maddala, T and Oxnard, G R and Klein, E A and Spigel, D R and Hartman, A R and Aravanis, A and Seiden, M	Genome-wide cell-free DNA (cfDNA) methylation signatures and effect on tissue of origin (TOO) performance	Journal of Clinical Oncology	37			2019		https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.3049	meeting abstract	811 cancer cell methylomes representing 21 tumor types	Tissue-of-origin prediction	accuracy (training / test set split)	training + test set	"For multi-cancer detection using cfDNA, TOO determination is critical to enable safe and efficient diagnostic follow-up. Previous array-based studies captured < 2% of genomic CpGs. Here, we report genome-wide fragment-level methylation patterns across 811 cancer cell methylomes representing 21 tumor types (97% of SEER cancer incidence), and define effects of this methylation database on TOO prediction within a machine learning framework. Improvement was observed across all cancer types and was consistent in early-stage cancers (stage I-III). Respective performances in breast cancer (n = 23) were 87% vs 96%; in lung cancer (n = 32) were 85% vs 88%; in hepatobiliary (n = 10) were 70% vs 90%; and in pancreatic cancer (n = 17) were 94% vs 100%."	"For multi-cancer detection using cfDNA, TOO determination is critical to enable safe and efficient diagnostic follow-up. Previous array-based studies captured < 2% of genomic CpGs. Here, we report genome-wide fragment-level methylation patterns across 811 cancer cell methylomes representing 21 tumor types (97% of SEER cancer incidence), and define effects of this methylation database on TOO prediction within a machine learning framework. Improvement was observed across all cancer types and was consistent in early-stage cancers (stage I-III). Respective performances in breast cancer (n = 23) were 87% vs 96%; in lung cancer (n = 32) were 85% vs 88%; in hepatobiliary (n = 10) were 70% vs 90%; and in pancreatic cancer (n = 17) were 94% vs 100%."
113	Liu, W T and Wang, Y and Zhang, J and Ye, F and Huang, X H and Li, B and He, Q Y	A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma	Cancer Lett	425		43-53	2018	China	http://dx.doi.org/10.1016/j.canlet.2018.03.043	article	5 different microarray datasets that included 330 samples	Case-control study	accuracy (LOOCV, external test set)	cross-validation + external cohort validation	"Lung adenocarcinoma (LAC) is the most lethal cancer and the leading cause of cancer-related death worldwide. To determine potential indicators of LAC, we performed genome-wide relative significance (GWRS), genome-wide global significance (GWGS) and support vector machine (SVM) analyses progressively to identify robust gene biomarker signatures from 5 different microarray datasets that included 330 samples. In conclusion, our integrated microarray analysis demonstrated that CENPA, CDK1 and CDC20 might serve as novel cluster of prognostic biomarkers for LAC, and the cooperative unit of three genes provides a technically simple approach for identification of LAC patients."	"Lung adenocarcinoma (LAC) is the most lethal cancer and the leading cause of cancer-related death worldwide. To determine potential indicators of LAC, we performed genome-wide relative significance (GWRS), genome-wide global significance (GWGS) and support vector machine (SVM) analyses progressively to identify robust gene biomarker signatures from 5 different microarray datasets that included 330 samples. In conclusion, our integrated microarray analysis demonstrated that CENPA, CDK1 and CDC20 might serve as novel cluster of prognostic biomarkers for LAC, and the cooperative unit of three genes provides a technically simple approach for identification of LAC patients."
114	Liu, Y and Yieh, L and Yang, T and Drinkenburg, W and Peeters, P and Steckler, T and Narayan, V A and Wittenberg, G and Ye, J	Metabolomic biosignature differentiates melancholic depressive patients from healthy controls	BMC Genomics	17		669-669	2016	Belgium	http://dx.doi.org/10.1186/s12864-016-2953-2	article	"the data set consists of 97 healthy control and 90 MDD subjects"	Case-control study	accuracy, sensitivity, specificity (10-fold CV)	cross-validation	"Here we report results on the Janssen-BRC metabolomics data set, consisting of 97 healthy control and 90 MDD subjects, of which 21 suffer melancholic depression and 58 from anxious depression. In this work, our goals are three-fold. First, we test the hypothesis that more clinically homogeneous groups of MDD patients are easier to predict from healthy controls than the entire MDD group using blood metabolomics data. Second, we develop a novel method for building maximally predictive and robust machine-learning classifiers that retain information on the correlation structure of the metabolomics data to ease biological interpretation. Third, we use this framework to describe the metabolomics biosignature of melancholic depression."	"Here we report results on the Janssen-BRC metabolomics data set, consisting of 97 healthy control and 90 MDD subjects, of which 21 suffer melancholic depression and 58 from anxious depression. In this work, our goals are three-fold. First, we test the hypothesis that more clinically homogeneous groups of MDD patients are easier to predict from healthy controls than the entire MDD group using blood metabolomics data. Second, we develop a novel method for building maximally predictive and robust machine-learning classifiers that retain information on the correlation structure of the metabolomics data to ease biological interpretation. Third, we use this framework to describe the metabolomics biosignature of melancholic depression."
115	Long, N P and Jung, K H and Yoon, S J and Anh, N H and Nghi, T D and Kang, Y P and Yan, H H and Min, J E and Hong, S S and Kwon, S W	Systematic assessment of cervical cancer initiation and progression uncovers genetic panels for deep learning-based early diagnosis and proposes novel diagnostic and prognostic biomarkers	Oncotarget	8	65	109436-109456	2017	Vietnam	http://dx.doi.org/10.18632/oncotarget.22689	article	202 cancer, 115 cervical intraepithelial neoplasia (CIN), and 105 normal samples	Case-control study	accuracy, sensitivity, specificity (10-fold CV, external test set)	cross-validation + external cohort validation	"In this study, eight different gene expression data sets containing 202 cancer, 115 cervical intraepithelial neoplasia (CIN), and 105 normal samples were utilized for an integrative systems biology assessment in a multi-stage carcinogenesis manner. Deep learning-based diagnostic models were established based on the genetic panels of intrinsic genes of cervical carcinogenesis as well as on the unbiased variable selection approach. [...] the 168-gene deep learning model for the differentiation of cancer from normalcy achieved an externally validated accuracy of 97.96% (99.01% sensitivity and 95.65% specificity). Survival analysis revealed that ZNF281 and EPHB6 were the two most promising prognostic genetic markers for CxCa among others."	"In this study, eight different gene expression data sets containing 202 cancer, 115 cervical intraepithelial neoplasia (CIN), and 105 normal samples were utilized for an integrative systems biology assessment in a multi-stage carcinogenesis manner. Deep learning-based diagnostic models were established based on the genetic panels of intrinsic genes of cervical carcinogenesis as well as on the unbiased variable selection approach. [...] the 168-gene deep learning model for the differentiation of cancer from normalcy achieved an externally validated accuracy of 97.96% (99.01% sensitivity and 95.65% specificity). Survival analysis revealed that ZNF281 and EPHB6 were the two most promising prognostic genetic markers for CxCa among others."
116	Long, N P and Nghi, T D and Kang, Y P and Anh, N H and Kim, H M and Park, S K and Kwon, S W	Toward a standardized strategy of clinical metabolomics for the advancement of precision medicine	Metabolites	10	2		2020	USA	http://dx.doi.org/10.3390/metabo10020051	article	review (not applicable)	review			"Despite the tremendous success, pitfalls have been observed in every step of a clinical metabolomics workflow, which impedes the internal validity of the study. n this conceptual review, we will cover inclusive barriers of a metabolomics-based clinical study and suggest potential solutions in the hope of enhancing study robustness, usability, and transferability. The importance of quality assurance and quality control procedures is discussed, followed by a practical rule containing five phases, including two additional "pre-pre-" and "post-post-" analytical steps. Besides, we will elucidate the potential involvement of machine learning and demonstrate that the need for automated data mining algorithms to improve the quality of future research is undeniable. Consequently, we propose a comprehensive metabolomics framework, along with an appropriate checklist refined from current guidelines and our previously published assessment, in the attempt to accurately translate achievements in metabolomics into clinical and epidemiological research."	"Despite the tremendous success, pitfalls have been observed in every step of a clinical metabolomics workflow, which impedes the internal validity of the study. n this conceptual review, we will cover inclusive barriers of a metabolomics-based clinical study and suggest potential solutions in the hope of enhancing study robustness, usability, and transferability. The importance of quality assurance and quality control procedures is discussed, followed by a practical rule containing five phases, including two additional "pre-pre-" and "post-post-" analytical steps. Besides, we will elucidate the potential involvement of machine learning and demonstrate that the need for automated data mining algorithms to improve the quality of future research is undeniable. Consequently, we propose a comprehensive metabolomics framework, along with an appropriate checklist refined from current guidelines and our previously published assessment, in the attempt to accurately translate achievements in metabolomics into clinical and epidemiological research."
117	Long, N P and Park, S and Anh, N H and Nghi, T D and Yoon, S J and Park, J H and Lim, J and Kwon, S W	High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer	Int J Mol Sci	20	2		2019	Vietnam	http://dx.doi.org/10.3390/ijms20020296	article	"The data set GSE44861 comprised 56 CRC tissues and 55 adjacent noncancerous tissues from the United States. The data set GSE41258 had 183 CRC and 44 adjacent noncancerous tissues from the United States between 1992 and 2004. The data set GSE83889 contained 101 CRC tissues and 35 non-neoplastic mucosal tissues from all patients with stage III of CRC from Korea."	Case-control study	AUC, sensitivity, specificity (5-times repeated 10-fold CV, test set)	cross-validation + test set	"his study employed a novel approach combining multi-platform transcriptomics and cutting-edge algorithms to introduce novel signatures for accurate diagnosis of colorectal cancer (CRC). All models showed satisfactory performance in which RF appeared to be the best. For instance, regarding the RF model, the following were observed: mean accuracy 0.998 (standard deviation (SD) < 0.003), mean specificity 0.999 (SD < 0.003), and mean sensitivity 0.998 (SD < 0.004). Moreover, proposed biomarker signatures were highly associated with multifaceted hallmarks in cancer."	"his study employed a novel approach combining multi-platform transcriptomics and cutting-edge algorithms to introduce novel signatures for accurate diagnosis of colorectal cancer (CRC). All models showed satisfactory performance in which RF appeared to be the best. For instance, regarding the RF model, the following were observed: mean accuracy 0.998 (standard deviation (SD) < 0.003), mean specificity 0.999 (SD < 0.003), and mean sensitivity 0.998 (SD < 0.004). Moreover, proposed biomarker signatures were highly associated with multifaceted hallmarks in cancer."
118	Long, N P and Yoon, S J and Anh, N H and Nghi, T D and Lim, D K and Hong, Y J and Hong, S S and Kwon, S W	A systematic review on metabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer	Metabolomics	14	8	109-109	2018	South Korea	http://dx.doi.org/10.1007/s11306-018-1404-2	article	review (not applicable)	review	AUC, sensitivity, specificity (25 discovery studies + different validation strategies across 9 validation studies)		"In this study, we conducted a systematic review to examine recent advancements in the oncometabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer. The included 25 studies primarily focused on the identification rather than the validation of predictive capacity of potential biomarkers. The sample size ranged from 10 to 8760. External validation of the biomarker panels was observed in nine studies. The diagnostic area under the curve ranged from 0.68 to 1.00 (sensitivity: 0.43–1.00, specificity: 0.73–1.00). The effects of patients' bio-parameters on metabolome alterations in a context-dependent manner have not been thoroughly elucidated."	"In this study, we conducted a systematic review to examine recent advancements in the oncometabolomics-based diagnostic biomarker discovery and validation in pancreatic cancer. The included 25 studies primarily focused on the identification rather than the validation of predictive capacity of potential biomarkers. The sample size ranged from 10 to 8760. External validation of the biomarker panels was observed in nine studies. The diagnostic area under the curve ranged from 0.68 to 1.00 (sensitivity: 0.43–1.00, specificity: 0.73–1.00). The effects of patients' bio-parameters on metabolome alterations in a context-dependent manner have not been thoroughly elucidated."

119	Lopez, C and Tucker, S and Salameh, T and Tucker, C	An unsupervised machine learning method for discovering patient clusters based on genetic signatures	J Biomed Inform	85		30-39	2018	USA	http://dx.doi.org/10.1016/j.jbi.2018.07.004	article	191 multiple sclerosis patient	Cases only (sub-group stratification)	Rand index on benchmark clusters (10-fold CV)	cross-validation	"This work presents an unsupervised machine learning method to cluster patients based on their genomic makeup without providing input parameters a priori. The method implements internal validity metrics to algorithmically identify the number of clusters, as well as statistical analyses to test for the significance of the results. Furthermore, the method takes advantage of the high degree of linkage disequilibrium between single nucleotide polymorphisms. Finally, a gene pathway analysis is performed to identify potential relationships between the clusters in the context of known biological knowledge. Benchmark results indicate that the proposed method provides the greatest performance out of the methods tested."
120	Lotsch, J and Schiffmann, S and Schmitz, K and Brunkhorst, R and Lerch, F and Ferreiros, N and Wicker, S and Tegeder, I and Geisslinger, G and Ultsch, A	Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy	Sci Rep	8	1	14884-14884	2018	Germany	http://dx.doi.org/10.1038/s41598-018-33077-8	article	MS patients (n = 102) and healthy subjects (n = 301)	Case-control study	accuracy, sensitivity, specificity (10-fold nested CV)	cross-validation	"Based on increasing evidence suggesting that MS pathology involves alterations in bioactive lipid metabolism, the present analysis was aimed at generating a complex serum lipid-biomarker. Using unsupervised machine-learning, implemented as emergent self-organizing maps of neuronal networks, swarm intelligence and Minimum Curvilinear Embedding, a cluster structure was found in the input data space comprising serum concentrations of d = 43 different lipid-markers of various classes. This was subsequently assessed using supervised machine-learning, implemented as random forests and computed ABC analysis-based feature selection. Bayesian statistics-based biomarker creation was used to map the diagnostic classes of either MS patients (n = 102) or healthy subjects (n = 301). A complex classifier or biomarker was developed that predicted MS at a sensitivity, specificity and accuracy of approximately 95% in training and test data sets, respectively."
121	Lu, T P and Kuo, K T and Chen, C H and Chang, M C and Lin, H P and Hu, Y H and Chiang, Y C and Cheng, W F and Chen, C A	Developing a Prognostic Gene Panel of Epithelial Ovarian Cancer Patients by a Machine Learning Model	Cancers	11	2	13-13	2019	Switzerland	http://dx.doi.org/10.3390/cancers11020270	article	3 different dataset with > 50 samples per group	Cases only (prognosis study)	accuracy, log-rank test p-value (LOOCV)	cross-validation	"Epithelial ovarian cancer patients usually relapse after primary management. We utilized the support vector machine algorithm to develop a model for the chemo-response using the Cancer Cell Line Encyclopedia (CCLE) and validated the model in The Cancer Genome Atlas (TCGA) and the GSE9891 dataset. The 10-gene predictive model demonstrated that the high response group had a longer recurrence-free survival (RFS) (log-rank test, p = 0.015 for TCGA, p = 0.013 for GSE9891 and p = 0.039 for NTUH) and overall survival (OS) (log-rank test, p = 0.002 for TCGA and p = 0.016 for NTUH). In a multivariate Cox hazard regression model, the predictive model (HR: 0.644, 95% CI: 0.436–0.952, p = 0.027) and residual tumor size < 1 cm (HR: 0.312, 95% CI: 0.170–0.573, p < 0.001) were significant factors for recurrence. The predictive model (HR: 0.511, 95% CI: 0.334–0.783, p = 0.002) and residual tumor size < 1 cm (HR: 0.252, 95% CI: 0.128–0.496, p < 0.001) were still significant factors for death." "Lung adenocarcinoma (LUAD) accounts for a majority of cancer-related deaths worldwide annually. The identification of prognostic biomarkers and prediction of prognosis for LUAD patients is necessary. In this study, LUAD RNA-Seq data and clinical data from the Cancer Genome Atlas (TCGA) were divided into TCGA cohort I (n = 338) and II (n = 168). First, the survival-related seed genes were selected from the cohort I using the machine learning model (random survival forest, RSF), and then in order to improve prediction accuracy, the forward selection model was utilized to identify the prognosis-related key genes among the seed genes using the clinically-integrated RNA-Seq data. Second, the survival risk score system was constructed by using these key genes in the cohort II, the GSE72094 cohort and the GSE11969 cohort, and the evaluation metrics such as HR, p value and C-index were calculated to validate the proposed method. Based on the RSF model and clinically-integrated RNA-Seq data, we identified sixteen key genes that formed the prognostic gene expression signature. These sixteen key genes could achieve a strong power for prognostic prediction of LUAD patients in cohort II (HR = 3.80, p = 1.63e-06, C-index = 0.656), and were further validated in the GSE72094 cohort (HR = 4.12, p = 1.34e-10, C-index = 0.672) and GSE11969 cohort (HR = 3.87, p = 6.81e-07, C-index = 0.670)."
122	Ma, B and Geng, Y and Meng, F and Yan, G and Song, F	Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method	Journal of Cancer	11	5	1288-1298	2020	China	http://dx.doi.org/10.7150/jca.34585	article	TCGA cohort I (n = 338) and II (n = 168)	Cases only (prognosis study)	hazard ratio, p-value and C-index (training / test set split)	training + test set	"We show that by examining how fast ISV performance approaches RCV as the number of studies is increased, one can estimate when "sufficient" diversity has been achieved for learning a molecular signature likely to translate without significant loss of accuracy to new clinical settings."
123	Ma, S and Sung, J and Magis, A T and Wang, Y and Geman, D and Price, N D	Measuring the effect of inter-study variability on estimating prediction error	PLoS One	9	10	e110840-e110840	2014	United States of America	http://dx.doi.org/10.1371/journal.pone.0110840	article	"1,470 microarray samples of 6 lung phenotypes from 26 independent experimental studies and 769 RNA-seq samples of 2 lung phenotypes from 4 independent studies"	Case-control study	"Here we quantify the impact of these combined "study-effects" on a disease signature's predictive performance by comparing two types of validation methods: ordinary randomized cross-validation (RCV), which extracts random subsets of samples for testing, and inter-study validation (ISV), which excludes an entire study for testing."	cross-validation + external cohort validation	"In this chapter, we review common bioinformatics approaches that aim to use sequencing data to predict sample-specific drug susceptibility. First, we explain the importance of customized drug regimens to the future of medical care. Second, we discuss the different public databases and community efforts that can be leveraged to develop new methods for identifying new predictive biomarkers. Third, we cover the basic methods that are currently used to identify markers or signatures of drug response, without any prior knowledge of the drug's mechanism of action. We further discuss how one can integrate knowledge about drug targets, mechanisms, and predictive markers to better estimate drug response in a diverse set of samples. We begin this section with a primer on popular methods to identify targets and mechanism of action for new small molecules. This discussion also includes a set of computational methods that incorporate other drug features, which do not relate to drug-induced genetic changes or sequencing data such as drug structures, side-effects, and efficacy profiles."
124	Madhukar, N S and Elemento, O	Bioinformatics Approaches to Predict Drug Responses from Genomic Sequencing	Methods Mol Biol	1711		277-296	2018	USA	http://dx.doi.org/10.1007/978-1-4939-7493-1_14	article	review (not applicable)	review			"Deep neural networks (DNNs) are efficient algorithms based on the use of compositional layers of neurons, with advantages well matched to the challenges - omics data presents. While achieving state-of-the-art results and even surpassing human accuracy in many challenging tasks, the adoption of deep learning in biomedicine has been comparatively slow. Here, we discuss key features of deep learning that may give this approach an edge over other machine learning methods. We then consider limitations and review a number of applications of deep learning in biomedical studies demonstrating proof of concept and practical utility."
125	Mamoshina, P and Vieira, A and Putin, E and Zhavoronkov, A	Applications of Deep Learning in Biomedicine	Mol Pharm	13	5	1445-1454	2016	China	http://dx.doi.org/10.1021/acs.molpharmaceut.5b00982	article	review (not applicable)	review			"DNA methylation data-based precision cancer diagnostics is emerging as the state of the art for molecular tumor classification. Standards for choosing statistical methods with regard to well-calibrated probability estimates for these typically highly multiclass classification tasks are still lacking. To support this choice, we evaluated well-established machine learning (ML) classifiers including random forests (RFs), elastic net (ELNET), support vector machines (SVMs) and boosted trees in combination with post-processing algorithms and developed ML workflows that allow for unbiased class probability (CP) estimation. ELNET was the top stand-alone classifier with the best calibration profiles. The best overall two-stage workflow was MR-calibrated SVM with linear kernels closely followed by ridge-calibrated tuned RF. For calibration, MR was the most effective regardless of the primary classifier. The protocols developed as a result of these comparisons provide valuable guidance on choosing ML workflows and their tuning to generate well-calibrated CP estimates for precision diagnostics using DNA methylation data."
126	Maros, M E and Capper, D and Jones, D T W and Hovestadt, V and von Deimling, A and Pfister, S M and Benner, A and Zucknick, M and Sill, M	Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data	Nature Protocols	15	2	479-512	2020	Germany	http://dx.doi.org/10.1038/s41596-019-0251-6	article	brain tumor 450k DNA methylation cohort of 2,801 samples with 91 diagnostic categories (multiple groups with > 50 samples per group)	Cases only (sub-group stratification)	accuracy (5 × 5-fold nested cross-validation scheme)	cross-validation	"Traumatic brain injury (TBI) affects 1.7 million people in the United States each year, causing lifelong functional deficits in cognition and behavior. The complex pathophysiology of neural injury is a primary barrier to developing sensitive and specific diagnostic tools, which consequentially has a detrimental effect on treatment regimens. Biomarkers of other diseases (e.g. cancer) have provided critical insight into disease emergence and progression that lend to developing powerful clinical tools for intervention. Therefore, the biomarker discovery field has recently focused on TBI and made substantial advancements to characterize markers with promise of transforming TBI patient diagnostics and care. This review focuses on these key advances in neural injury biomarkers discovery, including novel approaches spanning from omics-based approaches to imaging and machine learning as well as the evolution of established techniques. [...] Several biomarkers of TBI have been identified but they carry the disadvantage of either not being sensitive or specific to TBI, which diminishes their clinical utility. Biomarkers have the potential for improving diagnostic accuracy, predicting the severity of injury progression, and conveying information to clinicians about injury progression for individual patients. Advancements in biomarker discovery range from improving upon already established techniques to applying novel methods to elucidate mechanisms of the neural injury environment."
127	Martinez, B I and Stabenfeldt, S E	Current trends in biomarker discovery and analysis tools for traumatic brain injury	Journal of Biological Engineering	13	1		2019	USA	http://dx.doi.org/10.1186/s13036-019-0146-8	article	review (not applicable)	review			"Non-Alcoholic Fatty Liver Disease encompasses a spectrum of diseases ranging from simple steatosis to NASH and cirrhosis/HCC. The challenge in this field is to recognize the more severe and/or progressive pathology. A reliable non-invasive method based on biomarkers does not exist at the moment. Metabolomics technique has a great potential for this task, because it can non-invasively perform a complete "metabolic fingerprint" of a disease and, in turn, potentially detect all its evolution steps. With this aim, we performed a serum metabolomics characterization of several NAFLD forms and then tested its accuracy confronting it with an independent cohort by mean of machine-learning models' approach. [...] Blind analysis using the described test showed a global accuracy for NAFLD identification of 96.8% ± 2.1, 94.0% ± 4.2 for NASH and 81.2% ± 12.2 for NASH cirrhosis identification."
128	Masarone, M and Troisi, J and Aglitti, A and Calvanese, G and Torre, P and Caruso, R and Colucci, A and Dallio, M and Federico, A and Balsano, C and Persico, M	Accuracy of metabolomics profiles to non-invasively diagnose NAFLD stages and evolution by mean of machine-learning automated algorithms	Digestive and Liver Disease	52		e9-e9	2020		http://dx.doi.org/10.1016/j.dld.2019.12.023	meeting abstract	two cohorts of a total of 319 subjects: "The first cohort, coming from the main recruitment center (University of Salerno),was composed by 69 healthy subjects (CTRL) and 144 NAFLD patients (78 NAFL, 23 NASH, 43 NASH cirrhosis) and the second, coming from the other centers, was composed by 106 subjects(40 CTRL, 34 NAFL, 10 NASH, 18 NASH cirrhosis)"	Case-control study	accuracy (training / test set split)	training + test set	

129	Matlock, K and De Niz, C and Rahman, R and Ghosh, S and Pal, R	Investigation of model stacking for drug sensitivity prediction	BMC Bioinformatics	19		71-71	2018	USA	http://dx.doi.org/10.1186/s12859-018-2060-2	article	"we segregate 50 training samples into our vertical and horizontal groups, build individual predictive model RF with 50 trees, build the stacking model using a set of 150 samples, and obtain the prediction MSEs of candidate models on a set of 50 testing samples. We then add 2 training samples and reestimate the MSEs. We repeat this process until the training set has a total of 150 samples. The entire process is replicated 100 times with randomly selected training, testing, and validation samples in every iteration."	Cases only (drug sensitivity prediction)	normalized AUC (training, testind and validation set)	training + test set	"A significant problem in precision medicine is the prediction of drug sensitivity for individual cancer cell lines. [...] We explore the predictive performance of model stacking and the effect of stacking on the predictive bias and squared error. In addition we discuss the analytical underpinnings supporting the advantages of stacking in reducing squared error and inherent bias of random forests in prediction of outliers. The performance of individual and stacked models are compared. We note that stacking models built on two heterogeneous datasets provide superior performance to stacking different models built on the same dataset. It is also noted that stacking provides a noticeable reduction in the bias of our predictors when the dominant eigenvalue of the principle axis of variation in the residuals is significantly higher than the remaining eigenvalues."	"A significant problem in precision medicine is the prediction of drug sensitivity for individual cancer cell lines. [...] We explore the predictive performance of model stacking and the effect of stacking on the predictive bias and squared error. In addition we discuss the analytical underpinnings supporting the advantages of stacking in reducing squared error and inherent bias of random forests in prediction of outliers. The performance of individual and stacked models are compared. We note that stacking models built on two heterogeneous datasets provide superior performance to stacking different models built on the same dataset. It is also noted that stacking provides a noticeable reduction in the bias of our predictors when the dominant eigenvalue of the principle axis of variation in the residuals is significantly higher than the remaining eigenvalues."
130	McCarthy, C and Shrestha, S and Ibrahim, N E and Van Kimmenade, R and Gaggin, H K and Mukai, R and Magaret, C A and Barnes, G and Rhyne, R and Garasic, J M and Januzzi, J L	Performance of a clinical/proteomic panel to predict obstructive peripheral artery disease in patients with and without diabetes mellitus	European Heart Journal	39		117-117	2018	Netherlands	http://dx.doi.org/10.1093/eurheartj/ehy564.P732	meeting abstract	"354 patients undergoing peripheral and/or coronary angiography, performance of this diagnostic panel was assessed in patients with (N=94) and without DM (N=260) using Monte Carlo cross validation"	Case-control study	sensitivity, specificity, PPV, NPV (Monte Carlo cross-validation)	cross-validation	"Peripheral artery disease (PAD) is a global health problem associated with significant morbidity and mortality. Patients with diabetes mellitus (DM) are at substantial risk of developing PAD, however its diagnosis is often delayed until advanced stages when complications arise. Using proteomics and machine learning, a sub-set of artificial intelligence, we recently described a biomarker clinical/proteomic panel to predict prevalent obstructive PAD (HART PAD) in patients undergoing diagnostic peripheral angiography and/or coronary angiography. In this study, we sought to compare the accuracy of this clinical/proteomic panel for the diagnosis of PAD in patients with and without DM. [...] In patients with DM, the HART PAD panel had excellent performance or prediction of peripheral stenosis >50%. The model had an area under the receiver operating characteristic curve of 0.85 for obstructive PAD; higher scores were associated with greater severity of angiographic stenosis"	"Peripheral artery disease (PAD) is a global health problem associated with significant morbidity and mortality. Patients with diabetes mellitus (DM) are at substantial risk of developing PAD, however its diagnosis is often delayed until advanced stages when complications arise. Using proteomics and machine learning, a sub-set of artificial intelligence, we recently described a biomarker clinical/proteomic panel to predict prevalent obstructive PAD (HART PAD) in patients undergoing diagnostic peripheral angiography and/or coronary angiography. In this study, we sought to compare the accuracy of this clinical/proteomic panel for the diagnosis of PAD in patients with and without DM. [...] In patients with DM, the HART PAD panel had excellent performance or prediction of peripheral stenosis >50%. The model had an area under the receiver operating characteristic curve of 0.85 for obstructive PAD; higher scores were associated with greater severity of angiographic stenosis"
131	McDermott, J E and Wang, J and Mitchell, H and Webb-Robertson, B J and Hafen, R and Ramey, J and Rodland, K D	Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data	Expert Opinion on Medical Diagnostics	7	1	37-51	2013	USA	http://dx.doi.org/10.1517/17530059.2012.718329	article	review (not applicable)	review			"In this review we will present examples of current practices for biomarker discovery from complex omic datasets and the challenges that have been encountered in deriving valid and useful signatures of disease. We will then present a high-level review of data-driven (statistical) and knowledge-based methods applied to biomarker discovery, highlighting some current efforts to combine the two distinct approaches. [...] Effective, reproducible and objective tools for combining data-driven and knowledge-based approaches to identify predictive signatures of disease are key to future success in the biomarker field. We will describe our recommendations for possible approaches to this problem including metrics for the evaluation of biomarkers."	"More effective integration of data-driven and expert-driven strategies for biomarker discovery and identification is seen as key to improved success. Data-driven approaches to biomarker discovery focus on data reduction, data classification, and data visualization. Expert knowledge-driven approaches feature curated knowledge of protein-protein interactions, pathways, gene function, and ontologies."
132	McGeachie, M and Kelly, R S and Litonjua, A A and Weiss, S T and Lasky-Su, J A	Network of year-3 metabolites indicative of early-life asthma	American Journal of Respiratory and Critical Care Medicine	197			2018		http://dx.doi.org/10.3390/nu12051233	meeting abstract	cohort of 411 three-year olds at high-risk for asthm	Case-control study	AUC (five-fold cross-validation)	cross-validation	"While some small children display early-life wheeze, fewer have diagnosable asthma. We aimed to identify a metabolomic signature of early-life asthma, which could be used as a diagnostic test or which would provide insight into the biochemical basis of asthma in young children. [...] The BN methodology identified a Bayesian Network predictive of Year-3 asthma with 21 metabolites achieving an Area Under the Receiver Operator Characteristic Curve (AUC) of 86.5%"	"While some small children display early-life wheeze, fewer have diagnosable asthma. We aimed to identify a metabolomic signature of early-life asthma, which could be used as a diagnostic test or which would provide insight into the biochemical basis of asthma in young children. [...] The BN methodology identified a Bayesian Network predictive of Year-3 asthma with 21 metabolites achieving an Area Under the Receiver Operator Characteristic Curve (AUC) of 86.5%"
133	Meikle, P and Tsorotes, D and Barlow, C and Weir, J and MacIntosh, G and Barber, M and Goudey, B and Bedo, J and Stern, L and Kowalczyk, A and Haviv, I and White, A and Dart, A and Duffy, S and Kingwell, B	Plasma lipidomic analysis of stable and unstable coronary artery disease	Atherosclerosis Supplements	11	2	24-24	2010	Australia	http://dx.doi.org/10.1016/S1567-5688(10)70103-3	meeting abstract	202 participants (control, n = 60, stable CAD, n = 61, unstable CAD, n = 81)	Case-control study	AUC (multiple cross-validation iterations)	cross-validation	"Currently there is no means to predict plaque instability in coronary artery disease (CAD). Modified ceramide and modified phosphatidylcholine species were shown to distinguish stable and unstable CAD. These newly identified biomarkers were measured together with known plasma lipids, including sphingolipids, acylglycerols, and phospholipids, to establish a plasma lipid profile using electrospray-ionisation tandem mass spectrometry. [...] Multivariate analysis using a statistical machine learning approach combined with recursive feature elimination and multiple cross-validation iterations was applied for the creation of prediction models. Comparison of models with varying number of features showed that models with only 8 lipids were sufficient to provide optimal discrimination between stable and unstable cohorts (AUC = 0.75) while 16 lipids were sufficient to discriminate control from CAD patients (AUC = 0.94)."	"Currently there is no means to predict plaque instability in coronary artery disease (CAD). Modified ceramide and modified phosphatidylcholine species were shown to distinguish stable and unstable CAD. These newly identified biomarkers were measured together with known plasma lipids, including sphingolipids, acylglycerols, and phospholipids, to establish a plasma lipid profile using electrospray-ionisation tandem mass spectrometry. [...] Multivariate analysis using a statistical machine learning approach combined with recursive feature elimination and multiple cross-validation iterations was applied for the creation of prediction models. Comparison of models with varying number of features showed that models with only 8 lipids were sufficient to provide optimal discrimination between stable and unstable cohorts (AUC = 0.75) while 16 lipids were sufficient to discriminate control from CAD patients (AUC = 0.94)."
134	Menden, M P and Iorio, F and Garnett, M and McDermott, U and Benes, C H and Ballester, P J and Saez-Rodriguez, J	Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties	Plos One	8	4	7-7	2013	United Kingdom	http://dx.doi.org/10.1371/journal.pone.0061318	article	"We used the data from the Genomics of Drug Sensitivity in Cancer project [3], which contains 639 cancer cell lines, each of them characterised by a set of genomic features (details in the next section). The characterisation is not complete for every cell line, and therefore we filtered out cell lines with more than 15 missing genomic features, which reduced the set of selected cell lines from 639 to 608. The dataset contains 131 drugs."	Cases only (drug sensitivity prediction)	R-squared (8-fold cross-validation, hold-out test set)	cross-validation + test set	"Predicting the response of a specific cancer to a therapy is a major goal in modern oncology that should ultimately lead to a personalised treatment. we developed machine learning models to predict the response of cancer cell lines to drug treatment, quantified through IC50 values, based on both the genomic features of the cell lines and the chemical properties of the considered drugs. Models predicted IC50 values in a 8-fold cross-validation and an independent blind test with coefficient of determination R2 of 0.72 and 0.64 respectively."	"Predicting the response of a specific cancer to a therapy is a major goal in modern oncology that should ultimately lead to a personalised treatment. we developed machine learning models to predict the response of cancer cell lines to drug treatment, quantified through IC50 values, based on both the genomic features of the cell lines and the chemical properties of the considered drugs. Models predicted IC50 values in a 8-fold cross-validation and an independent blind test with coefficient of determination R2 of 0.72 and 0.64 respectively."
135	Midorikawa, Y and Tsuji, S and Takayama, T and Aburatani, H	Genomic approach towards personalized anticancer drug therapy	Pharmacogenomics	13	2	191-199	2012	Japan	http://dx.doi.org/10.2217/pgs.11.157	article	review (not applicable)	review			"Here, we review recent advances in the development of classification algorithms using microarray technology for prediction of anticancer sensitivity, discuss the availability of ensemble methods for prediction models, and present data regarding the identification of potential responders to FOLFOX therapy using random forests algorithm."	"Ensemble methods, such as random forests, are effective when there are more predictors than observations, and in predicting response to multiagent chemotherapy. To reproduce public data, stricter rules for public data availability and transparent data processing and analysis are necessary. It is difficult to apply prediction models based on cell lines to samples from patients because of heterogeneity of cancer cells."
136	Mobadersany, P and Yousefi, S and Amgad, M and Gutman, D A and Barnholtz-Sloan, J S and Velazquez Vega, J E and Brat, D J and Cooper, L A D	Predicting cancer outcomes from histology and genomics using convolutional networks	Proc Natl Acad Sci U S A	115	13	E2970-e2979	2018	USA	http://dx.doi.org/10.1073/pnas.1717139115	article	769 unique patients	Cases only (prognosis study)	15 accuracy measurements, including Harrell's C-index for measuring concordance between predicted risks and actual survival (Monte Carlo cross-validation)	cross-validation	"We developed a computational approach based on deep learning to predict the overall survival of patients diagnosed with brain tumors from microscopic images of tissue biopsies and genomic biomarkers. This method uses adaptive feedback to simultaneously learn the visual patterns and molecular biomarkers associated with patient outcomes. Our approach surpasses the prognostic accuracy of human experts using the current clinical standard for classifying brain tumors and presents an innovative approach for objective, accurate, and integrated prediction of patient outcomes."	"We developed a computational approach based on deep learning to predict the overall survival of patients diagnosed with brain tumors from microscopic images of tissue biopsies and genomic biomarkers. This method uses adaptive feedback to simultaneously learn the visual patterns and molecular biomarkers associated with patient outcomes. Our approach surpasses the prognostic accuracy of human experts using the current clinical standard for classifying brain tumors and presents an innovative approach for objective, accurate, and integrated prediction of patient outcomes."
137	Modlin, I and Kidd, M and Drozdov, I and Bodei, L and Malczewska, A and Matar, S	Automated finger prick blood genomic diagnosis of neuroendocrine tumors	Neuroendocrinology	108		132-132	2019		http://dx.doi.org/10.1159/000498996	meeting abstract	whole blood samples from >6,000 NETs and controls	Case-control study	sensitivity, specificity (training/test set split)	training + test set	"A sensitive, noninvasive technology that is safely repeatable and provides real-time information is required for neuroendocrine tumors (NET) diagnosis/management. "	"NET transcriptomes from GEP NETs were defined (2005-2008) and in 2010, RNA-tissue classifiers based on machine learning developed. From 2010-2015, we defined using AI 51 genes co-expressed in blood and tumor tissue (R>0.8) and established a liquid biopsy (NETest). Diagnostic metrics were >90% sensitivity and specificity. Clinical evaluation (n>6,000 patients) demonstrated the multigene assay to be an effective prognostic biomarker."
138	Mohammed, A and Biegert, G and Adamec, J and Helikar, T	Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers	Oncotarget	8	49	85692-85715	2017	USA	http://dx.doi.org/10.18632/oncotarget.21127	article	"A total of 2,175 tissue samples, both normal and cancerous, were collected from nine distinct tissues: blood (595), breast (171), colon (105), gastric (333), head and neck (82), lung (542), prostate (56), thyroid (224), and tongue (67)"	Case-control study	accuracy, sensitivity, specificity, precision, F1 score (10-fold cross-validation)	cross-validation	"Machine learning techniques for cancer prediction and biomarker discovery can hasten cancer detection and significantly improve prognosis. Recent “OMICS” studies which include a variety of cancer and normal tissue samples along with machine learning approaches have the potential to further accelerate such discovery. To demonstrate this potential, 2,175 gene expression samples from nine tissue types were obtained to identify gene sets whose expression is characteristic of each cancer class. Using random forests classification and ten-fold cross-validation, we developed nine single-tissue classifiers, two multi-tissue cancer-versus-normal classifiers, and one multi-tissue normal classifier. Given a sample of a specified tissue type, the single-tissue models classified samples as cancer or normal with a testing accuracy between 85.29% and 100%. Given a sample of non-specific tissue type, the multi-tissue bi-class model classified the sample as cancer versus normal with a testing accuracy of 97.89%. Given a sample of non-specific tissue type, the multi-tissue multi-class model classified the sample as cancer versus normal and as a specific tissue type with a testing accuracy of 97.43%. Given a normal sample of any of the nine tissue types, the multi-tissue normal model classified the sample as a particular tissue type with a testing accuracy of 97.35%."	"Machine learning techniques for cancer prediction and biomarker discovery can hasten cancer detection and significantly improve prognosis. Recent “OMICS” studies which include a variety of cancer and normal tissue samples along with machine learning approaches have the potential to further accelerate such discovery. To demonstrate this potential, 2,175 gene expression samples from nine tissue types were obtained to identify gene sets whose expression is characteristic of each cancer class. Using random forests classification and ten-fold cross-validation, we developed nine single-tissue classifiers, two multi-tissue cancer-versus-normal classifiers, and one multi-tissue normal classifier. Given a sample of a specified tissue type, the single-tissue models classified samples as cancer or normal with a testing accuracy between 85.29% and 100%. Given a sample of non-specific tissue type, the multi-tissue bi-class model classified the sample as cancer versus normal with a testing accuracy of 97.89%. Given a sample of non-specific tissue type, the multi-tissue multi-class model classified the sample as cancer versus normal and as a specific tissue type with a testing accuracy of 97.43%. Given a normal sample of any of the nine tissue types, the multi-tissue normal model classified the sample as a particular tissue type with a testing accuracy of 97.35%."
139	Mourikis, T P and Benedetti, L and Foxall, E and Temelkovski, D and Nulsen, J and Perner, J and Cereda, M and Lagergren, J and Howell, M and Yau, C and Fitzgerald, R C and Scaffidi, P and Ciccarelli, F D	Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma	Nat Commun	10	1	3101-3101	2019	Italy	http://dx.doi.org/10.1038/s41467-019-10898-3	article	261 esophageal adenocarcinomas (EACs) + 107 additional EACs for validation	Cases only (survival prediction)	log-rank test p-value (cross-validation)	cross-validation	"The identification of cancer-promoting genetic alterations is challenging particularly in highly unstable and heterogeneous cancers, such as esophageal adenocarcinoma (EAC). Here we describe a machine learning algorithm to identify cancer genes in individual patients considering all types of damaging alterations simultaneously. [...] Experimentally mimicking the alterations of predicted helper genes in cancer and pre-cancer cells validates their contribution to disease progression"	"The identification of cancer-promoting genetic alterations is challenging particularly in highly unstable and heterogeneous cancers, such as esophageal adenocarcinoma (EAC). Here we describe a machine learning algorithm to identify cancer genes in individual patients considering all types of damaging alterations simultaneously. [...] Experimentally mimicking the alterations of predicted helper genes in cancer and pre-cancer cells validates their contribution to disease progression"
140	Murugesan, K and Javle, M and Schrock, A B and Ngo, N and Frampton, G M and Alexander, B M and Miller, V A and Bekaii-Saab, T and Albacker, L A and Ross, J S and Ali, S M	Comprehensive genomic profiling (CGP) of mixed hepatocellular cholangiocarcinomas (cHCC-CCA)	Annals of Oncology	30		v256-v257	2019		http://dx.doi.org/10.1093/annonc/mdz247.005	article	"1269 HCC, 3965 CCA and 44 cHCC-CCA" (> 50 samples for the main conditions)	Case-control study	error rate, AUC (Random Forest out-of-bag error)	outofbag	"cHCC-CCA is a rare primary liver carcinoma, with histologic features of both hepatocellular carcinomas (HCC) and liver cholangiocarcinomas (CCA). In order to elucidate their shared and distinctive biology, we used Comprehensive genomic profiling (CGP) to compare the genomic alterations (GA) of cHCC-CCA with those of HCC and CCA. [...] Utilizing the significant differences in GAs, biomarkers, and demographics between HCC and CCAs (Fischer exact test, FDR < 0.05, only GAs with freq > 10% shown, Table), we built a random forest-based machine learning model to rank a cHCC-CCA specimen in the CCA-HCC spectrum (Out of Bag error rate = 12.6%, AUC= 0.94). Biomarkers exclusively associated with one disease type included IDH1 (15% in CCA vs 1% in HCC, p=1e-57), TERT (4% in CCA vs 47% in HCC, p=2e-273) and Hepatitis B virus (HBV, 0.8% in CCA vs 8.5% in HCC, p=9e-42)."	"cHCC-CCA is a rare primary liver carcinoma, with histologic features of both hepatocellular carcinomas (HCC) and liver cholangiocarcinomas (CCA). In order to elucidate their shared and distinctive biology, we used Comprehensive genomic profiling (CGP) to compare the genomic alterations (GA) of cHCC-CCA with those of HCC and CCA. [...] Utilizing the significant differences in GAs, biomarkers, and demographics between HCC and CCAs (Fischer exact test, FDR < 0.05, only GAs with freq > 10% shown, Table), we built a random forest-based machine learning model to rank a cHCC-CCA specimen in the CCA-HCC spectrum (Out of Bag error rate = 12.6%, AUC= 0.94). Biomarkers exclusively associated with one disease type included IDH1 (15% in CCA vs 1% in HCC, p=1e-57), TERT (4% in CCA vs 47% in HCC, p=2e-273) and Hepatitis B virus (HBV, 0.8% in CCA vs 8.5% in HCC, p=9e-42)."
141	Nakamura, M and Bax, H J and Scotto, D and Souri, E A and Sollie, S and Harris, R J and Hammar, N and Walldius, G and Winship, A and Ghosh, S and Montes, A and Spicer, J F and Van Hemelrijck, M and Josephs, D H and Lacy, K E and Tsoka, S and Karagiannis, S N	Immune mediator expression signatures are associated with improved outcome in ovarian carcinoma	OncoImmunology	8	6		2019	Sweden	http://dx.doi.org/10.1080/2162402X.2019.1593811	article	1,656 ovarian carcinoma patient tumors	Cases only (survival prediction)	accuracy, recall, sensitivity, Matthew's correlation coefficient, F1 score (5 times 10 fold cross-validation)	cross-validation	"Immune and inflammatory cascades may play multiple roles in ovarian cancer. We aimed to identify relationships between expression of immune and inflammatory mediators and patient outcomes. We interrogated differential gene expression of 44 markers and marker combinations (n = 1,978) in 1,656 ovarian carcinoma patient tumors, alongside matched 5-year overall survival (OS) data in silico. Expression of the 44 mediators could discriminate between malignant and non-malignant tissues with at least 96% accuracy."	"Immune and inflammatory cascades may play multiple roles in ovarian cancer. We aimed to identify relationships between expression of immune and inflammatory mediators and patient outcomes. We interrogated differential gene expression of 44 markers and marker combinations (n = 1,978) in 1,656 ovarian carcinoma patient tumors, alongside matched 5-year overall survival (OS) data in silico. Expression of the 44 mediators could discriminate between malignant and non-malignant tissues with at least 96% accuracy."

142	Nakariyakul, S	A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification	PLoS One	14	2	e0212333-e0212333	2019	Thailand	http://dx.doi.org/10.1371/journal.pone.0212333	article	ten microarray data sets with > 50 samples per group for multiple datasets	Case-control study	accuracy, precision, recall, F-score (nested cross-validation)	cross-validation	"We address gene selection and machine learning methods for cancer classification using microarray gene expression data. Due to the high dimensionality of microarray data, traditional gene selection algorithms are filter-based, focusing on intrinsic properties of the data such as distance, dependency, and correlation. These methods are fast but select far too many genes to use for the classification task. In this work, we present a new hybrid filter-wrapper gene subset selection algorithm that is an improved modification of our prior algorithm. Our proposed method employs interaction information to rank candidate genes to add into a gene subset. [...] Experimental results on ten public cancer microarray data sets show that our method consistently outperforms prior gene selection algorithms in terms of classification accuracy, while requiring a small number of selected genes."	"We address gene selection and machine learning methods for cancer classification using microarray gene expression data. Due to the high dimensionality of microarray data, traditional gene selection algorithms are filter-based, focusing on intrinsic properties of the data such as distance, dependency, and correlation. These methods are fast but select far too many genes to use for the classification task. In this work, we present a new hybrid filter-wrapper gene subset selection algorithm that is an improved modification of our prior algorithm. Our proposed method employs interaction information to rank candidate genes to add into a gene subset. [...] Experimental results on ten public cancer microarray data sets show that our method consistently outperforms prior gene selection algorithms in terms of classification accuracy, while requiring a small number of selected genes."
143	Naorem, L D and Muthaiyan, M and Venkatesan, A	Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer	Journal of Cellular Biochemistry	120	4	6154-6167	2019	India	http://dx.doi.org/10.1002/jcb.27903	article	Six microarray data sets consisting of 463 non-TNBC and 405 TNBC samples	Case-control study	AUC (training / test set split)	training + test set	"Triple-negative breast cancer (TNBC) has attracted more attention compared with other breast cancer subtypes due to its aggressive nature, poor prognosis, and chemotherapy remains the mainstay of treatment with no other approved targeted therapy. Therefore, the study aimed to discover more promising therapeutic targets and investigating new insights of biological mechanism of TNBC. Six microarray data sets consisting of 463 non-TNBC and 405 TNBC samples were mined from Gene Expression Omnibus. [...] A naïve Bayes based classifier built using the expression profiles of 16 features (hub genes) accurately and reliably classify TNBC from non-TNBC samples in the validation test data set with a receiver operating curve of 0.93 to 0.98."	"Triple-negative breast cancer (TNBC) has attracted more attention compared with other breast cancer subtypes due to its aggressive nature, poor prognosis, and chemotherapy remains the mainstay of treatment with no other approved targeted therapy. Therefore, the study aimed to discover more promising therapeutic targets and investigating new insights of biological mechanism of TNBC. Six microarray data sets consisting of 463 non-TNBC and 405 TNBC samples were mined from Gene Expression Omnibus. [...] A naïve Bayes based classifier built using the expression profiles of 16 features (hub genes) accurately and reliably classify TNBC from non-TNBC samples in the validation test data set with a receiver operating curve of 0.93 to 0.98."
144	Nazha, A and Komrokji, R S and Barnard, J and Al-Issa, K and Padron, E and Madanat, Y F and Kuzmanovic, T and Abuhadra, N and Steensma, D P and DeZern, A E and Roboz, G J and Garcia-Manero, G and List, A F and Maciejewski, J P and Sekeres, M A	A Personalized Prediction Model to Risk Stratify Patients with Myelodysplastic Syndromes (MDS)	Blood	130			2017		https://ashpublications.org/blood/article/132/Supplement%201/793/266156/A-Personalized-Prediction-Model-to-Risk-Stratify	article	"Of 2302 pts, 1471 were included in the training cohort and 831 in the validation cohort"	Cases only (survival prediction)	C-index (training and validation cohort)	external cohort validation	"We built a personalized prediction model based on clinical and genomic data that outperformed IPSS and IPSS-R in predicting OS and AML transformation. The new model gives survival probabilities at different time points that are unique for a given pt. Incorporating clinical and mutational data outperformed a mutations only model even when cytogenetics and age were added"	"We built a personalized prediction model based on clinical and genomic data that outperformed IPSS and IPSS-R in predicting OS and AML transformation. The new model gives survival probabilities at different time points that are unique for a given pt. Incorporating clinical and mutational data outperformed a mutations only model even when cytogenetics and age were added"
145	Nazha, A and Sekeres, M A and Bejar, R and Komrokji, R S and Barnard, J and Al-Issa, K and Przychodzen, B P and Hirsch, C M and Steensma, D P and DeZern, A E and Roboz, G J and Garcia-Manero, G and Ebert, B L and Maciejewski, J P	Genomic Biomarkers to Predict Response to Hypomethylating Agents in Patients with Myelodysplastic Syndromes (MDS)	Blood	130			2017		https://ashpublications.org/blood/article/130/Supplement%201/157/79546/Genomic-Biomarkers-to-Predict-Response-to	article	433 pts with MDS (per 2008 WHO criteria) who received HMAs (230 at our institution [training cohort], and 203 at multiple other academic institutions [validation cohort])	Cases only (treatment response prediction)	accuracy, sensitivity, specificity (training and validation cohort)	external cohort validation	"While treatment with the hypomethylating agents (HMAs) azacitidine (AZA) and decitabine (DAC) improves cytopenias and prolongs survival in MDS patients (pts), response is not guaranteed. Identification of non-responders could prevent prolonged exposure to ineffective therapy, avoid toxicities and decrease unnecessary costs. We developed an unbiased framework to study the association of several mutations in predicting response to HMAs, analogous to Netflix or Amazon's recommender system in which customers who bought products A and B is likely to buy C: pts who have a mutation in gene A, and B are likely to respond or not respond to HMAs. [...] When applying these rules to the validation cohort, the sensitivity of these genomic biomarkers for no response to HMA was 1, specificity for response to HMA was 1, and accuracy was .85."	"While treatment with the hypomethylating agents (HMAs) azacitidine (AZA) and decitabine (DAC) improves cytopenias and prolongs survival in MDS patients (pts), response is not guaranteed. Identification of non-responders could prevent prolonged exposure to ineffective therapy, avoid toxicities and decrease unnecessary costs. We developed an unbiased framework to study the association of several mutations in predicting response to HMAs, analogous to Netflix or Amazon's recommender system in which customers who bought products A and B is likely to buy C: pts who have a mutation in gene A, and B are likely to respond or not respond to HMAs. [...] When applying these rules to the validation cohort, the sensitivity of these genomic biomarkers for no response to HMA was 1, specificity for response to HMA was 1, and accuracy was .85."
146	Nowak, C and Carlsson, A C and Östgren, C J and Nyström, F H and Alam, M and Feldreich, T R and Sundström, J and Carrero Roig, J J and Leppert, J and Hedberg, P O and Cordeiro, A C and Lind, L and Ingelsson, E and Fall, T and Ärnlöv, J	Multiplex proteomics for prediction of major cardiovascular events in type 2 diabetes	Diabetologia	61		S65-S65	2018	Brazil	http://dx.doi.org/10.1007/s00125-018-4693-0	meeting abstract	1,211 adults with type 2 diabetes	Case-control study	accuracy with 95%-confidence intervals (training / test set split)	training + test set	"Multiplex proteomics could improve understanding and risk prediction of major adverse cardiovascular events (MACE) in type 2 diabetes. This study assessed 80 cardiovascular and inflammatory proteins for biomarker discovery and prediction of MACE in type 2 diabetes. [...] Addition of the 80-protein assay to the established risk model improved discrimination in the separate validation sample from 68.6% (95%-CI, 68.2%-68.9%) to 74.8% (95%-CI, 74.6%-75.1%)."	"Multiplex proteomics could improve understanding and risk prediction of major adverse cardiovascular events (MACE) in type 2 diabetes. This study assessed 80 cardiovascular and inflammatory proteins for biomarker discovery and prediction of MACE in type 2 diabetes. [...] Addition of the 80-protein assay to the established risk model improved discrimination in the separate validation sample from 68.6% (95%-CI, 68.2%-68.9%) to 74.8% (95%-CI, 74.6%-75.1%)."
147	O'Reilly, P and Ortutay, C and Gernon, G and O'Connell, E and Seoighe, C and Boyce, S and Serrano, L and Szegezdi, E	Co-acting gene networks predict TRAIL responsiveness of tumour cells with high accuracy	BMC Genomics	15		1144-1144	2014	Ireland	http://dx.doi.org/10.1186/1471-2164-15-1144	article	Gene expression microarray data for 109 tumor cell lines with known sensitivity to the death ligand cytokine tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)	Case-control study	AUC, sensitivity, specificity (training and validation cohort)	external cohort validation	"Gene expression microarray data for 109 tumor cell lines with known sensitivity to the death ligand cytokine tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) was used to identify genes with potential functional relationships determining responsiveness to TRAIL-induced apoptosis. The machine learning technique Random Forest in the statistical environment "R" with backward elimination was used to identify the key predictors of TRAIL. [...] Prediction accuracy was assessed by calculating the area under the receiver operator curve using an independent dataset. We show that the gene panel identified could predict TRAIL-sensitivity with a very high degree of sensitivity and specificity (AUC = 0 · 84). The genes in the panel are co-regulated and at least 40% of them functionally interact in signal transduction pathways that regulate cell death and cell survival, cellular differentiation and morphogenesis. Importantly, only 12% of the TRAIL-predictor genes were differentially expressed highlighting the importance of functional interactions in predicting the biological response."	"Gene expression microarray data for 109 tumor cell lines with known sensitivity to the death ligand cytokine tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) was used to identify genes with potential functional relationships determining responsiveness to TRAIL-induced apoptosis. The machine learning technique Random Forest in the statistical environment "R" with backward elimination was used to identify the key predictors of TRAIL. [...] Prediction accuracy was assessed by calculating the area under the receiver operator curve using an independent dataset. We show that the gene panel identified could predict TRAIL-sensitivity with a very high degree of sensitivity and specificity (AUC = 0 · 84). The genes in the panel are co-regulated and at least 40% of them functionally interact in signal transduction pathways that regulate cell death and cell survival, cellular differentiation and morphogenesis. Importantly, only 12% of the TRAIL-predictor genes were differentially expressed highlighting the importance of functional interactions in predicting the biological response."
148	Oh, J H and Lotan, Y and Gurnani, P and Rosenblatt, K P and Gao, J	Prostate cancer biomarker discovery using high performance mass spectral serum profiling	Comput Methods Programs Biomed	96	1	33-41	2009	USA	http://dx.doi.org/10.1016/j.cmpb.2009.04.003	article	Serum samples from 179 prostate cancer patients and 74 benign patients	Case-control study	accuracy, sensitivity, specificity, NPV, PPV (20 times 10-fold CV)	cross-validation	"Prostate-specific antigen (PSA) is the most widely used serum biomarker for early detection of prostate cancer (PCA). Nevertheless, PSA level can be falsely elevated due to prostatic enlargement, inflammation or infection, which limits the PSA test specificity. The objective of this study is to use a machine learning approach for the analysis of mass spectrometry data to discover more reliable biomarkers that distinguish PCA from benign specimens. [...] From the new marker selection algorithm, a panel of 26 peaks achieved an accuracy of 80.7%, a sensitivity of 83.5%, a specificity of 74.4%, a positive predic- tive value (PPV) of 87.9%, and a negative predictive value (NPV) of 68.2%. On the other hand, when PSA alone was used (with a cutoff of 4.0 ng/ml), a sensitivity of 66.7%, a specificity of 53.6%, a PPV of 73.5%, and a NPV of 45.4% were obtained."	"Prostate-specific antigen (PSA) is the most widely used serum biomarker for early detection of prostate cancer (PCA). Nevertheless, PSA level can be falsely elevated due to prostatic enlargement, inflammation or infection, which limits the PSA test specificity. The objective of this study is to use a machine learning approach for the analysis of mass spectrometry data to discover more reliable biomarkers that distinguish PCA from benign specimens. [...] From the new marker selection algorithm, a panel of 26 peaks achieved an accuracy of 80.7%, a sensitivity of 83.5%, a specificity of 74.4%, a positive predic- tive value (PPV) of 87.9%, and a negative predictive value (NPV) of 68.2%. On the other hand, when PSA alone was used (with a cutoff of 4.0 ng/ml), a sensitivity of 66.7%, a specificity of 53.6%, a PPV of 73.5%, and a NPV of 45.4% were obtained."
149	Okser, S and Pahikkala, T and Aittokallio, T	Genetic variants and their interactions in disease risk prediction - Machine learning and network perspectives	BioData Mining	6	1		2013	Finland	http://dx.doi.org/10.1186/1756-0381-6-5	article	review (not applicable)	review			"A central challenge in systems biology and medical genetics is to understand how interactions among genetic loci contribute to complex phenotypic traits and human diseases. While most studies have so far relied on statistical modeling and association testing procedures, machine learning and predictive modeling approaches are increasingly being applied to mining genotype-phenotype relationships, also among those associations that do not necessarily meet statistical significance at the level of individual variants, yet still contributing to the combined predictive power at the level of variant panels. Network-based analysis of genetic variants and their interaction partners is another emerging trend by which to explore how sub-network level features contribute to complex disease processes and related phenotypes. In this review, we describe the basic concepts and algorithms behind machine learning-based genetic feature selection approaches, their potential benefits and limitations in genome-wide setting, and how physical or genetic interaction networks could be used as a priori information for providing improved predictive power and mechanistic insights into the disease networks."	"A central challenge in systems biology and medical genetics is to understand how interactions among genetic loci contribute to complex phenotypic traits and human diseases. While most studies have so far relied on statistical modeling and association testing procedures, machine learning and predictive modeling approaches are increasingly being applied to mining genotype-phenotype relationships, also among those associations that do not necessarily meet statistical significance at the level of individual variants, yet still contributing to the combined predictive power at the level of variant panels. Network-based analysis of genetic variants and their interaction partners is another emerging trend by which to explore how sub-network level features contribute to complex disease processes and related phenotypes. In this review, we describe the basic concepts and algorithms behind machine learning-based genetic feature selection approaches, their potential benefits and limitations in genome-wide setting, and how physical or genetic interaction networks could be used as a priori information for providing improved predictive power and mechanistic insights into the disease networks."
150	Orgueira, A M and Rodríguez, B A and Vence, N A and López Á, B and Arias, J A D and Varela, N D and Pérez, M S G and Encinas, M M P and López, J L B	Time to treatment prediction in chronic lymphocytic leukemia based on new transcriptional patterns	Frontiers in Oncology	9			2019	Spain	http://dx.doi.org/10.3389/fonc.2019.00079	article	"The first cohort (EGAD00001001443, hereafter study cohort) contains RNAseq data and from CLL-purified cells of 196 individuals along with clinical data. The cohort was composed of 169 CLL, 22 monoclonal B cell lymphocytosis (MBL), and five small lymphocytic lymphoma (SLL) samples. There were 132 IGHV mutated cases and 64 IGHV unmutated cases in 119 males and 77 females. By staging at diagnosis, there were 22 MBL cases, 151 Binet Stage A cases, 14 Binet Stage B cases, and 8 Binet C stage cases. The second cohort (EGAD00001000258, hereafter validation cohort) is composed of RNAseq data of CLL-purified cells from 98 individuals, of which 79 (55 males and 24 females) have publicly available phenotypic information. In this cohort there were 72 CLL, 4 SLL, and 3 MBL samples. 45 of the patients had mutated IGHV and 34 had unmutated IGHV. By staging at diagnosis, there were 3 MBL, 72 Binet Stage A, 3 Binet Stage B, and 1 Binet Stage C cases."	Case-control study	accuracy, precision, recall, ROC (training and validation cohort)	external cohort validation	"Chronic lymphocytic leukemia (CLL) is the most frequent lymphoproliferative syndrome in western countries. CLL evolution is frequently indolent, and treatment is mostly reserved for those patients with signs or symptoms of disease progression. In this work, we used RNA sequencing data from the International Cancer Genome Consortium CLL cohort to determine new gene expression patterns that correlate with clinical evolution.We determined that a 290-gene expression signature, in addition to immunoglobulin heavy chain variable region (IGHV) mutation status, stratifies patients into four groups with notably different time to first treatment. This finding was confirmed in an independent cohort. Similarly, we present a machine learning algorithm that predicts the need for treatment within the first 5 years following diagnosis using expression data from 2,198 genes. This predictor achieved 90% precision and 89% accuracy when classifying independent CLL cases."	"Chronic lymphocytic leukemia (CLL) is the most frequent lymphoproliferative syndrome in western countries. CLL evolution is frequently indolent, and treatment is mostly reserved for those patients with signs or symptoms of disease progression. In this work, we used RNA sequencing data from the International Cancer Genome Consortium CLL cohort to determine new gene expression patterns that correlate with clinical evolution.We determined that a 290-gene expression signature, in addition to immunoglobulin heavy chain variable region (IGHV) mutation status, stratifies patients into four groups with notably different time to first treatment. This finding was confirmed in an independent cohort. Similarly, we present a machine learning algorithm that predicts the need for treatment within the first 5 years following diagnosis using expression data from 2,198 genes. This predictor achieved 90% precision and 89% accuracy when classifying independent CLL cases."
151	Oriol, J D and Vallejo, E E and Estrada, K and Pena, J G T and Alzheimers Dis Neuroimaging, Initia	Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data	Bmc Bioinformatics	20	1	17-17	2019	Mexico	http://dx.doi.org/10.1186/s12859-019-3158-x	article	more than 50 samples per group for both discovery and validation cohort	Case-control study	balanced error, accuracy, sensitivity, specificity, AUC (cross-validation, training + validation cohort)	cross-validation + external cohort validation	"Late-Onset Alzheimer's Disease (LOAD) is a leading form of dementia. There is no effective cure for LOAD, leaving the treatment efforts to depend on preventive cognitive therapies, which stand to benefit from the timely estimation of the risk of developing the disease. We conducted systematic comparisons of representative Machine Learning models for predicting LOAD from genetic variation data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Our experimental results demonstrate that the classification performance of the best models tested yielded ~72% of area under the ROC curve"	"Late-Onset Alzheimer's Disease (LOAD) is a leading form of dementia. There is no effective cure for LOAD, leaving the treatment efforts to depend on preventive cognitive therapies, which stand to benefit from the timely estimation of the risk of developing the disease. We conducted systematic comparisons of representative Machine Learning models for predicting LOAD from genetic variation data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Our experimental results demonstrate that the classification performance of the best models tested yielded ~72% of area under the ROC curve"

152	Orlenko, A and Moore, J H and Orzechowski, P and Olson, R S and Cairns, J and Caraballo, P J and Weinshilboum, R M and Wang, L W and Breitenstein, M K	Considerations for automated machine learning in clinical metabolic profiling: Altered homocysteine plasma concentration associated with metformin exposure	Pacific Symposium on Biocomputing 2018			460-471	2018	USA	https://pubmed.ncbi.nlm.nih.gov/29218905/	article	546 unique patients (Cases (n=273) included patients exposed to metformin monotherapy with type 2 diabetes having glycemic control; controls consisted of healthy normal patients with no known metformin exposure)	Case-control study	accuracy (training / test set split)	training + test set	"With the maturation of metabolomics science and proliferation of biobanks, clinical metabolic profiling is an increasingly opportunistic frontier for advancing translational clinical research. Automated Machine Learning (AutoML) approaches provide exciting opportunity to guide feature selection in agnostic metabolic profiling endeavors, where potentially thousands of independent data points must be evaluated. In previous research, AutoML using high-dimensional data of varying types has been demonstrably robust, outperforming traditional approaches. However, considerations for application in clinical metabolic profiling remain to be evaluated. Particularly, regarding the robustness of AutoML to identify and adjust for common clinical confounders. In this study, we present a focused case study regarding AutoML considerations for using the Tree-Based Optimization Tool (TPOT) in metabolic profiling of exposure to metformin in a biobank cohort. [...] First, we propose a tandem rank-accuracy measure to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors. Second, while AutoML, using default parameters, demonstrated potential to lack sensitivity to low-effect confounding clinical covariates, we demonstrated residual training and adjustment of metabolite features as an easily applicable approach to ensure AutoML adjustment for potential confounding characteristics. Finally, we present increased homocysteine with long-term exposure to metformin as a potentially novel, non-replicated metabolite association suggested by TPOT; an association not identified in parallel clinical metabolic profiling endeavors."	"With the maturation of metabolomics science and proliferation of biobanks, clinical metabolic profiling is an increasingly opportunistic frontier for advancing translational clinical research. Automated Machine Learning (AutoML) approaches provide exciting opportunity to guide feature selection in agnostic metabolic profiling endeavors, where potentially thousands of independent data points must be evaluated. In previous research, AutoML using high-dimensional data of varying types has been demonstrably robust, outperforming traditional approaches. However, considerations for application in clinical metabolic profiling remain to be evaluated. Particularly, regarding the robustness of AutoML to identify and adjust for common clinical confounders. In this study, we present a focused case study regarding AutoML considerations for using the Tree-Based Optimization Tool (TPOT) in metabolic profiling of exposure to metformin in a biobank cohort. [...] First, we propose a tandem rank-accuracy measure to guide agnostic feature selection and corresponding threshold determination in clinical metabolic profiling endeavors. Second, while AutoML, using default parameters, demonstrated potential to lack sensitivity to low-effect confounding clinical covariates, we demonstrated residual training and adjustment of metabolite features as an easily applicable approach to ensure AutoML adjustment for potential confounding characteristics. Finally, we present increased homocysteine with long-term exposure to metformin as a potentially novel, non-replicated metabolite association suggested by TPOT; an association not identified in parallel clinical metabolic profiling endeavors."
153	Pandey, G and Pandey, O P and Rogers, A J and Ahsen, M E and Hoffman, G E and Raby, B A and Weiss, S T and Schadt, E E and Bunyavanich, S	A Nasal Brush-based Classifier of Asthma Identified by Machine Learning Analysis of Nasal RNA Sequence Data	Scientific Reports	8		15-15	2018	USA	http://dx.doi.org/10.1038/s41598-018-27189-4	article	190 subjects with mild/moderate asthma and controls	Case-control study	AUC (5x5-fold CV)	cross-validation	"Asthma is a common, under-diagnosed disease affecting all ages. We sought to identify a nasal brush-based classifier of mild/moderate asthma. 190 subjects with mild/moderate asthma and controls underwent nasal brushing and RNA sequencing of nasal samples. A machine learning-based pipeline identified an asthma classifier consisting of 90 genes interpreted via an L2-regularized logistic regression classification model. This classifier performed with strong predictive value and sensitivity across eight test sets"	"Asthma is a common, under-diagnosed disease affecting all ages. We sought to identify a nasal brush-based classifier of mild/moderate asthma. 190 subjects with mild/moderate asthma and controls underwent nasal brushing and RNA sequencing of nasal samples. A machine learning-based pipeline identified an asthma classifier consisting of 90 genes interpreted via an L2-regularized logistic regression classification model. This classifier performed with strong predictive value and sensitivity across eight test sets"
154	Parker, B J and Gunter, S and Bedo, J	Stratification bias in low signal microarray studies	BMC Bioinformatics	8		326-326	2007	Australia	http://dx.doi.org/10.1186/1471-2105-8-326	article	review (not applicable)	review			"When analysing microarray and other small sample size biological datasets, care is needed to avoid various biases. We analyse a form of bias, stratification bias, that can substantially affect analyses using sample-reuse validation techniques and lead to inaccurate results. This bias is due to imperfect stratification of samples in the training and test sets and the dependency between these stratification errors, i.e. the variations in class proportions in the training and test sets are negatively correlated. We show that when estimating the performance of classifiers on low signal datasets (i.e. those which are difficult to classify), which are typical of many prognostic microarray studies, commonly used performance measures can suffer from a substantial negative bias. For error rate this bias is only severe in quite restricted situations, but can be much larger and more frequent when using ranking measures such as the receiver operating characteristic (ROC) curve and area under the ROC (AUC). [...] The classification error rate can have large negative biases for balanced datasets, whereas the AUC shows substantial pessimistic biases even for imbalanced datasets. [...] Stratification bias can substantially affect several performance measures. In computing the AUC, the strategy of pooling the test samples from the various folds of cross-validation can lead to large biases; computing it as the average of per-fold estimates avoids this bias and is thus the recommended approach. As a more general solution applicable to other performance measures, we show that stratified repeated holdout and a modified version of k-fold cross-validation, balanced, stratified cross-validation and balanced leave-one-out cross-validation, avoids the bias. Therefore for model selection and evaluation of microarray and other small biological datasets, these methods should be used and unstratified versions avoided. In particular, the commonly used (unbalanced) leave-one-out cross-validation should not be used to estimate AUC for small datasets."	"When analysing microarray and other small sample size biological datasets, care is needed to avoid various biases. We analyse a form of bias, stratification bias, that can substantially affect analyses using sample-reuse validation techniques and lead to inaccurate results. This bias is due to imperfect stratification of samples in the training and test sets and the dependency between these stratification errors, i.e. the variations in class proportions in the training and test sets are negatively correlated. We show that when estimating the performance of classifiers on low signal datasets (i.e. those which are difficult to classify), which are typical of many prognostic microarray studies, commonly used performance measures can suffer from a substantial negative bias. For error rate this bias is only severe in quite restricted situations, but can be much larger and more frequent when using ranking measures such as the receiver operating characteristic (ROC) curve and area under the ROC (AUC). [...] The classification error rate can have large negative biases for balanced datasets, whereas the AUC shows substantial pessimistic biases even for imbalanced datasets. [...] Stratification bias can substantially affect several performance measures. In computing the AUC, the strategy of pooling the test samples from the various folds of cross-validation can lead to large biases; computing it as the average of per-fold estimates avoids this bias and is thus the recommended approach. As a more general solution applicable to other performance measures, we show that stratified repeated holdout and a modified version of k-fold cross-validation, balanced, stratified cross-validation and balanced leave-one-out cross-validation, avoids the bias. Therefore for model selection and evaluation of microarray and other small biological datasets, these methods should be used and unstratified versions avoided. In particular, the commonly used (unbalanced) leave-one-out cross-validation should not be used to estimate AUC for small datasets."
155	Patil, S and Awan, K H and Arakeri, G and Seneviratne, C J and Muddur, N and Malik, S and Ferrari, M and Rahimi, S and Brennan, P A	Machine learning and its potential applications to the genomic study of head and neck cancer-A systematic review	Journal of Oral Pathology & Medicine	48	9	773-779	2019	India	http://dx.doi.org/10.1111/jop.12854	article	review (not applicable)	review			"The aim of this systematic review was to evaluate the existing literature and assess the application of machine learning of genomic data in head and neck cancer (HNC). [...] Two studies each evaluated oral cancer and laryngeal cancer, while other one study each evaluated nasopharyngeal cancer and oropharyngeal cancer. The majority of studies employed support vector machine (SVM) as a ML technique. Among the included studies, the accuracy rates for ML techniques ranged from 56.7% to 99.4%. Our findings showed that ML techniques for the analysis of genomic data can play a role in the prognostic prediction of HNC."	"The aim of this systematic review was to evaluate the existing literature and assess the application of machine learning of genomic data in head and neck cancer (HNC). [...] Two studies each evaluated oral cancer and laryngeal cancer, while other one study each evaluated nasopharyngeal cancer and oropharyngeal cancer. The majority of studies employed support vector machine (SVM) as a ML technique. Among the included studies, the accuracy rates for ML techniques ranged from 56.7% to 99.4%. Our findings showed that ML techniques for the analysis of genomic data can play a role in the prognostic prediction of HNC."
156	Porter, D and Goodyear, C S and Nijjar, J S and Messow, M and Siebert, S and Mudaliar, M and McInnes, I B	Predicting the response to TNF inhibition or B cell depletion therapy from peripheral whole blood gene expression profiles in patients with rheumatoid arthritis	Arthritis and Rheumatology	68		4130-4131	2016		http://dx.doi.org/10.1002/art.39977	meeting abstract	70% (n=169) of samples were used to develop response prediction models, and 30% (n=72) were reserved for validation	Cases only (drug response study)	sensitivity, specificity, PPV and NPV (10-fold CV)	cross-validation	"The ORBIT study demonstrated that rituximab is non-inferior to a TNFi-first strategy in biologic naive, sero-positive patients with active rheumatoid arthritis (RA) over 12 months (Lancet doi.org/10.1016/S0140-6736(16)00380-9). However, a significant proportion of patients failed to respond to their first biologic drug and switched to an alternative. The ability to identify and stratify these patients prior to treatment would improve patient care and optimise the use of scarce financial resources. The aim of this study was to identify peripheral blood transcriptional biomarkers in the ORBIT cohort that can predict subsequent response/non-response to biologic therapy. Three gene sets were identified using support vector machine (SVM) recursive feature elimination. These predicted: general responsiveness to both TNFi and rituximab therapy (8 genes), response to TNFi therapy (23 genes) or rituximab (23 genes) respectively. When tested on the validation set, these models resulted in ROC plots with an AUC of 91.6% for general responsiveness, 89.7% for TNFi response, and 85.7% for rituximab response."	"The ORBIT study demonstrated that rituximab is non-inferior to a TNFi-first strategy in biologic naive, sero-positive patients with active rheumatoid arthritis (RA) over 12 months (Lancet doi.org/10.1016/S0140-6736(16)00380-9). However, a significant proportion of patients failed to respond to their first biologic drug and switched to an alternative. The ability to identify and stratify these patients prior to treatment would improve patient care and optimise the use of scarce financial resources. The aim of this study was to identify peripheral blood transcriptional biomarkers in the ORBIT cohort that can predict subsequent response/non-response to biologic therapy. Three gene sets were identified using support vector machine (SVM) recursive feature elimination. These predicted: general responsiveness to both TNFi and rituximab therapy (8 genes), response to TNFi therapy (23 genes) or rituximab (23 genes) respectively. When tested on the validation set, these models resulted in ROC plots with an AUC of 91.6% for general responsiveness, 89.7% for TNFi response, and 85.7% for rituximab response."
157	Pratt, A G and Swan, D and Richardson, S and Wilson, G and Hilkens, C and Young, D and Isaacs, J D	A transcriptional profile present in CD4 T-cells of patients with undifferentiated arthritis predicts the future development of seronegative rheumatoid arthritis and implicates IL-6 in disease evolution	Arthritis and Rheumatism	63	10		2011		https://insights.ovid.com/arthritis-rheumatism/arhe/2011/11/001/transcriptional-profile-present-cd4-cells-patients/383/00000889	article	"Microarray analysis of 111 RNA samples was performed [...] Machine learning approaches were used to test the utility of a classification model amongst an independent validation cohort of 62 patients presenting with UA."	Case-control study	sensitivity, specificity (training and validation cohort)	external cohort validation	"The diagnosis of seronegative rheumatoid arthritis (RA) remains challenging in the early arthritis clinic. Recent GWAS data strongly implicate CD4+ T-cells in the pathogenesis of seropositive RA. Our objectives were to identify biomarker(s) present in CD4+ T-cells, or in serum, that identified patients with undifferentiated arthritis (UA) destined to develop seronegative RA. [...] Machine learning approaches were used to test the utility of a classification model amongst an independent validation cohort of 62 patients presenting with UA. [...] A 12-gene expression "signature" predicted the subsequent development of RA amongst ACPA-negative UA patients in the validation cohort (sensitivity 85%, specificity 75%). The signature had a predictive value equivalent to the Leiden score in these patients and provided enhanced predictive power in combination with the Leiden score. The 12-gene signature contained an over-representation of STAT3-target genes, and pathway analysis confirmed that genes functionally involved with CD4+ T-cell survival, including STAT pathway components, were deregulated in early RA."	"The diagnosis of seronegative rheumatoid arthritis (RA) remains challenging in the early arthritis clinic. Recent GWAS data strongly implicate CD4+ T-cells in the pathogenesis of seropositive RA. Our objectives were to identify biomarker(s) present in CD4+ T-cells, or in serum, that identified patients with undifferentiated arthritis (UA) destined to develop seronegative RA. [...] Machine learning approaches were used to test the utility of a classification model amongst an independent validation cohort of 62 patients presenting with UA. [...] A 12-gene expression "signature" predicted the subsequent development of RA amongst ACPA-negative UA patients in the validation cohort (sensitivity 85%, specificity 75%). The signature had a predictive value equivalent to the Leiden score in these patients and provided enhanced predictive power in combination with the Leiden score. The 12-gene signature contained an over-representation of STAT3-target genes, and pathway analysis confirmed that genes functionally involved with CD4+ T-cell survival, including STAT pathway components, were deregulated in early RA."
158	Pu, W and Wang, C and Chen, S and Zhao, D and Zhou, Y and Ma, Y and Wang, Y and Li, C and Huang, Z and Jin, L and Guo, S and Wang, J and Wang, M	Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC)	Clin Epigenetics	9		129-129	2017	China	http://dx.doi.org/10.1186/s13148-017-0430-7	article	"A high-throughput DNA methylation dataset (100 samples) of ESCC from The Cancer Genome Atlas (TCGA) project was analyzed and validated along with another independent dataset (12 samples) from the Gene Expression Omnibus (GEO) database. [...] The candidate CpG sites as well as their adjacent regions were further validated in 94 pairs of ESCC tumor and adjacent normal tissues from the Chinese Han population using the targeted bisulfite sequencing method."	Case-control study	AUC, accuracy, sensitivity, specificity (5-fold cross-validation + test set)	cross-validation + test set	"DNA methylation has been implicated as a promising biomarker for precise cancer diagnosis. However, limited DNA methylation-based biomarkers have been described in esophageal squamous cell carcinoma (ESCC). [...] A high-throughput DNA methylation dataset (100 samples) of ESCC from The Cancer Genome Atlas (TCGA) project was analyzed and validated along with another independent dataset (12 samples) from the Gene Expression Omnibus (GEO) database. The methylation status of peripheral blood mononuclear cells and peripheral blood leukocytes from healthy controls was also utilized for biomarker selection. The candidate CpG sites as well as their adjacent regions were further validated in 94 pairs of ESCC tumor and adjacent normal tissues from the Chinese Han population using the targeted bisulfite sequencing method. [...] Eight statistical models along with five-fold cross-validation were further applied, in which the SVM model reached the best accuracy in both training and test dataset (accuracy = 0.82 and 0.80, respectively)."	"DNA methylation has been implicated as a promising biomarker for precise cancer diagnosis. However, limited DNA methylation-based biomarkers have been described in esophageal squamous cell carcinoma (ESCC). [...] A high-throughput DNA methylation dataset (100 samples) of ESCC from The Cancer Genome Atlas (TCGA) project was analyzed and validated along with another independent dataset (12 samples) from the Gene Expression Omnibus (GEO) database. The methylation status of peripheral blood mononuclear cells and peripheral blood leukocytes from healthy controls was also utilized for biomarker selection. The candidate CpG sites as well as their adjacent regions were further validated in 94 pairs of ESCC tumor and adjacent normal tissues from the Chinese Han population using the targeted bisulfite sequencing method. [...] Eight statistical models along with five-fold cross-validation were further applied, in which the SVM model reached the best accuracy in both training and test dataset (accuracy = 0.82 and 0.80, respectively)."
159	Pujos-Guillot, E and Bertrand, J and Rambeau, M and Pétéra, M and Brandolini, M and Fernandes, A and Matta, J and Lévy-Marchal, C and Czernichow, S and Comte, B	Contribution of an integrative multi-omic approach in the metabolic syndrome prediction: A nested case-control study	Drug Metabolism and Personalized Therapy	31	4	eA33-eA34	2016		http://dx.doi.org/10.1515/dmpt-2016-0030	meeting abstract	n=92 born small vs n=76 born adequate for gestational age	Case-control study	error rate (training/test split + validation set)	cross-validation + test set	"The rising worldwide prevalence of metabolic syndrome (MetS), a cluster of cardiometabolic risk factors of predictive of type 2 diabetes, relates largely to increasing obesity and sedentary but also to early metabolic life events [1]. Objective: The objective of the study was to identify predictive biomarkers of evolution toward MetS 8 years later, and to bring new knowledge about this pathological state using a multidisciplinary approach in an at-risk population (subjects with small birth weight). [...] Individual predictive models were first built using linear logistic regressions from the omics datasets. Metabolomic and proteomic data were finally integrated using random forest to determine whether multidimensional models improve prediction. The resulting models based on either 4 metabolites or 4 proteins showed good performances: 22% misclassification on training set, 25% on validation set vs 11% misclassification on training set, 33% on validation set, respectively. Multi-omic data integration improved performance and robustness of the prediction (11% misclassification on training set, 8% on validation set)."	"The rising worldwide prevalence of metabolic syndrome (MetS), a cluster of cardiometabolic risk factors of predictive of type 2 diabetes, relates largely to increasing obesity and sedentary but also to early metabolic life events [1]. Objective: The objective of the study was to identify predictive biomarkers of evolution toward MetS 8 years later, and to bring new knowledge about this pathological state using a multidisciplinary approach in an at-risk population (subjects with small birth weight). [...] Individual predictive models were first built using linear logistic regressions from the omics datasets. Metabolomic and proteomic data were finally integrated using random forest to determine whether multidimensional models improve prediction. The resulting models based on either 4 metabolites or 4 proteins showed good performances: 22% misclassification on training set, 25% on validation set vs 11% misclassification on training set, 33% on validation set, respectively. Multi-omic data integration improved performance and robustness of the prediction (11% misclassification on training set, 8% on validation set)."

160	Pusztai, L and Hess, K R	Clinical trial design for microarray predictive marker discovery and assessment	Ann Oncol	15	12	1731-1737	2004	USA	http://dx.doi.org/10.1093/annonc/mdh466	article	review (not applicable)	review			"This manuscript reviews methodological and statistical issues relevant to clinical trial design to discover and validate multigene predictors of response to therapy."	"To discover a predictive marker for a given treatment, a single-arm study design may be sufficient. A simple strategy is to base sample-size calculations on the number needed to ensure adequate power for the univariate screening of discriminating genes; in other words, how many training samples are needed to identify reliably an individually predictive gene? If we can assume that the array data are approximately normally distributed on some scale, then we can use standard two-sample testing methods to perform sample-size calculations [...] Once a candidate predictor has been identified and its predictive accuracy was estimated, the goal of an independent validation study is to: (i) define the sensitivity, specificity and the positive (PPV) and negative predictive values (NPV) with greater precision; and (ii) to prove clinical utility of the test. Different trial designs may be needed for different clinical situations, but there may not be a single best design for any particular clinical scenario. Several designs could yield complementary information (Figure 2). An important question for a predictive marker validation study is to determine whether the response rate is higher (and how much higher) in the group that is predicted to respond compared with unselected patients that may represent the current standard of care (in the case of chemotherapy for example)."
161	Rao, R and Dean, K and Misganaw, B and Somvanshi, P and Doyle, F	Validation of a Multi-Omic Biomarker Panel and Analysis of Disease Progression Trajectories in a Novel Longitudinal PTSD Cohort	Biological Psychiatry	85	10	S96-S96	2019		http://dx.doi.org/10.1016/j.biopsych.2019.03.245	meeting abstract	83 PTSD positive cases and 83 PTSD negative matched controls, and subsequently refined and validated in a cohort of 29 PTSD cases and 40 controls	Case-control study	AUC, accuracy, sensitivity, specificity (training / test set + external validation)	cross-validation + external cohort validation	"Signals associated with PTSD development might emerge across multiple levels of physiological regulation. Diagnostic classifiers synthesizing signals from several single-layer molecular signatures into a multi-omic panel can improve diagnostic performance compared to any individual molecular signature. [...] Single and multi-omic classifiers were initially identified in a cohort of 83 PTSD positive cases and 83 PTSD negative matched controls, and subsequently refined and validated in a cohort of 29 PTSD cases and 40 controls. A novel longitudinal cohort of 1800 active duty soldiers is used for external validation [...] We previously found that the multi-omic panel results in a small improvement in diagnostic performance in comparison to individual single-omic panels in the initial training and validation cohorts (AUC=0.80, 77% accuracy, 81% sensitivity, 73% specificity). Preliminary external validation in the longitudinal cohort suggests that single-omic metabolic panels constituting the multi-omic panel are significantly associated with PTSD status."	
162	Rappoport, N and Shamir, R	Multi-omic and multi-view clustering algorithms: Review and cancer benchmark	Nucleic Acids Research	46	20	10546-10562	2018	Israel	https://academic.oup.com/nar/article/46/20/10546/5123392	article	review (not applicable)	review			"Here, we review algorithms for multi-omics clustering, and discuss key issues in applying these algorithms. Our review covers methods developed specifically for omic data as well as generic multi-view methods developed in the machine learning community for joint clustering of multiple data types. In addition, using cancer data from TCGA, we perform an extensive benchmark spanning ten different cancer types, providing the first systematic comparison of leading multi-omics and multi-view clustering algorithms. The results highlight key issues regarding the use of single- versus multi-omics, the choice of clustering strategy, the power of generic multi-view methods and the use of approximated p-values for gauging solution quality."	"In our benchmark, single-omic data alone sometimes gave better results than multi-omics data. This was intensified when for each algorithm the 'best' single-omic for each cancer type was chosen. These results question the current assumptions underlying multi-omics analysis in general and multi-omics clustering in particular. [...] We detected large differences between the p-values derived from the χ2 approximation compared to the P-values derived from the permutation tests in the statistical tests we used. The differences were especially large due to the small sample size, small cluster sizes (in solutions with a high number of clusters) and due to a low number of events (high survival) for the logrank test. These p-values are used by single and multi-omic methods to assess their performance, and the logrank p-value is often the main argument for an algorithm's merit. The large differences between the P-values question the validity of analyses that are based on the χ2 approximation, at least for TCGA data. [...] The benchmark we performed is not without limitations. Gauging performance using patient survival is somewhat biased to known cancer subtypes, which may have been used in treatment decisions. Additionally, cancer subtypes that are biologically different may have similar survival. This is also true for enrichment of clinical parameters, although we attempted to choose parameters that would not lead to this bias. However, these measures are widely used for clustering assessment, including in the papers describing some of the benchmarked methods."
163	Reeve, J and Madill-Thomsen, K S and Halloran, P F	Using ensembles of machine learning classifers to maximize the accuracy and stability of molecular biopsy interpretation	American Journal of Transplantation	19		452-453	2019		http://dx.doi.org/10.1111/ajt.15405	meeting abstract	"1679 kidney transplant biopsies were repeatedly split at random into two training sets (N=600 each) and a test set (N=479)"	Case-control study	accuracy (training and validation set)	training + test set	"The Molecular Microscope diagnostic system (MMDx), based on microarray gene expression, uses ensembles of machine learning classifers rather than single genes, gene sets, or classifers, to maximize the accuracy of rejection diagnoses and injury assessment. We tested its accuracy and stability, and developed an automated system for generating molecular reports on kidney transplant biopsies. [...] Twelve separate machine learning methods and their median were evaluated. In a separate analysis, a random forest classifer was used to predict the report sign-outs of an expert clinician. Results: There was considerable variation between the 12 classifer methods for any given biopsy. The median had a higher accuracy than any of the individual classifers, and was among the most stable (highest correlation between predictions from separate random training sets-Figure 1C and D). A random forest classifer was used to predict the sign-out of an expert evaluator. Accuracies for the expert's molecular TCMR and ABMR diagnoses were ~98 and 97% respectively. Most disagreements were in biopsies near diagnostic thresholds."	"The Molecular Microscope diagnostic system (MMDx), based on microarray gene expression, uses ensembles of machine learning classifers rather than single genes, gene sets, or classifers, to maximize the accuracy of rejection diagnoses and injury assessment. We tested its accuracy and stability, and developed an automated system for generating molecular reports on kidney transplant biopsies. [...] Twelve separate machine learning methods and their median were evaluated. In a separate analysis, a random forest classifer was used to predict the report sign-outs of an expert clinician. Results: There was considerable variation between the 12 classifer methods for any given biopsy. The median had a higher accuracy than any of the individual classifers, and was among the most stable (highest correlation between predictions from separate random training sets-Figure 1C and D). A random forest classifer was used to predict the sign-out of an expert evaluator. Accuracies for the expert's molecular TCMR and ABMR diagnoses were ~98 and 97% respectively. Most disagreements were in biopsies near diagnostic thresholds."
164	Reeve, J and Sellares, J and De Freitas, D and Einecke, G and Bromberg, J and Matas, A and Halloran, P	Predicting graft failure in kidney transplant patients: A combined clinical/molecular approach to analyzing biopsies	American Journal of Transplantation	13		109-109	2013		http://dx.doi.org/10.1111/ajt.12265	meeting abstract	562 patients (1 biopsy per patient)	Case-control study	accuracy (training / test set split)	training + test set	"We combined an existing dataset with new data from a multi-centre international collaboration (INTERCOM), to predict graft loss in a patient population undergoing kidney transplant biopsies for cause. We used a machine learning method - Random Survival Forests - combining histologic, clinical, and microarray data from 562 patients (1 biopsy per patient). [...] In addition, three previously published molecular scores, IRRAT (injury response and repair transcripts), RS (molecular risk score) and ABMR score (probability of ABMR) were used. [...] Inclusion of molecular variables increased accuracy (All vs All except molecular), but the best model used only IRRATs and TxBx."	"We combined an existing dataset with new data from a multi-centre international collaboration (INTERCOM), to predict graft loss in a patient population undergoing kidney transplant biopsies for cause. We used a machine learning method - Random Survival Forests - combining histologic, clinical, and microarray data from 562 patients (1 biopsy per patient). [...] In addition, three previously published molecular scores, IRRAT (injury response and repair transcripts), RS (molecular risk score) and ABMR score (probability of ABMR) were used. [...] Inclusion of molecular variables increased accuracy (All vs All except molecular), but the best model used only IRRATs and TxBx."
165	Reinhold, W C and Varma, S and Rajapakse, V N and Luna, A and Sousa, F G and Kohn, K W and Pommier, Y G	Using drug response data to identify molecular effectors, and molecular "omic" data to identify candidate drugs in cancer	Hum Genet	134	1	3-11	2015	USA	https://pubmed.ncbi.nlm.nih.gov/25213708/	article	review (not applicable)	review			"The current convergence of molecular and pharmacological data provides unprecedented opportunities to gain insights into the relationships between the two types of data. Multiple forms of large-scale molecular data, including but not limited to gene and microRNA transcript expression, DNA somatic and germline variations from next-generation DNA and RNA sequencing, and DNA copy number from array comparative genomic hybridization are all potentially informative when one attempts to recognize the panoply of potentially influential events both for cancer progression and therapeutic outcome. [...] For cancer cell lines, the National Cancer Institute cell line panel (NCI-60), the Cancer Cell Line Encyclopedia (CCLE), and the collaborative Genomics of Drug Sensitivity in Cancer (GDSC) databases all provide subsets of these forms of data. For the patient-derived data, The Cancer Genome Atlas (TCGA) provides analogous forms of genomic information along with treatment histories. Integration of these data in turn relies on the fields of statistics and statistical learning. Multiple algorithmic approaches may be chosen, depending on the data being considered, and the nature of the question being asked. [...] A promising new direction for enhancing all these techniques is to leverage prior biological knowledge, such as molecular interactions derived from biological pathways using literature-curated resources, or computationally inferred gene regulatory networks."	"The current convergence of molecular and pharmacological data provides unprecedented opportunities to gain insights into the relationships between the two types of data. Multiple forms of large-scale molecular data, including but not limited to gene and microRNA transcript expression, DNA somatic and germline variations from next-generation DNA and RNA sequencing, and DNA copy number from array comparative genomic hybridization are all potentially informative when one attempts to recognize the panoply of potentially influential events both for cancer progression and therapeutic outcome. [...] For cancer cell lines, the National Cancer Institute cell line panel (NCI-60), the Cancer Cell Line Encyclopedia (CCLE), and the collaborative Genomics of Drug Sensitivity in Cancer (GDSC) databases all provide subsets of these forms of data. For the patient-derived data, The Cancer Genome Atlas (TCGA) provides analogous forms of genomic information along with treatment histories. Integration of these data in turn relies on the fields of statistics and statistical learning. Multiple algorithmic approaches may be chosen, depending on the data being considered, and the nature of the question being asked. [...] A promising new direction for enhancing all these techniques is to leverage prior biological knowledge, such as molecular interactions derived from biological pathways using literature-curated resources, or computationally inferred gene regulatory networks."
166	Ressom, H W and Varghese, R S and Abdel-Hamid, M and Eissa, S A and Saha, D and Goldman, L and Petricoin, E F and Conrads, T P and Veenstra, T D and Loffredo, C A and Goldman, R	Analysis of mass spectral serum profiles for biomarker selection	Bioinformatics	21	21	4039-4045	2005	USA	http://dx.doi.org/10.1093/bioinformatics/bti670	article	411 sera samples (199 from HCC patients and 212 from matched healthy individuals)	Case-control study	accuracy, sensitivity, specificity (k-fold cross-validation and bootstrapping methods)	cross-validation	"Mass spectrometric profiles of peptides and proteins obtained by current technologies are characterized by complex spectra, high dimensionality and substantial noise. These characteristics generate challenges in the discovery of proteins and protein-profiles that distinguish disease states, e.g. cancer patients from healthy individuals. We present low-level methods for the processing of mass spectral data and a machine learning method that combines support vector machines, with particle swarm optimization for biomarker selection. The proposed method identified mass points that achieved high prediction accuracy in distinguishing liver cancer patients from healthy individuals in SELDI-QqTOF profiles of serum."	"Mass spectrometric profiles of peptides and proteins obtained by current technologies are characterized by complex spectra, high dimensionality and substantial noise. These characteristics generate challenges in the discovery of proteins and protein-profiles that distinguish disease states, e.g. cancer patients from healthy individuals. We present low-level methods for the processing of mass spectral data and a machine learning method that combines support vector machines, with particle swarm optimization for biomarker selection. The proposed method identified mass points that achieved high prediction accuracy in distinguishing liver cancer patients from healthy individuals in SELDI-QqTOF profiles of serum."
167	Ritari, J and Hyvarinen, K and Koskela, S and Itala-Remes, M and Niittyvuopio, R and Nihtinen, A and Salmenniemi, U and Putkonen, M and Volin, L and Kwan, T and Pastinen, T and Partanen, J	Genomic prediction of relapse in recipients of allogeneic haematopoietic stem cell transplantation	Leukemia	33	1	240-248	2019	Canada	http://dx.doi.org/10.1038/s41375-018-0229-3	article	"We studied 151 graft recipients with HLA-matched sibling donors by sequencing the whole-exome, active immunoregulatory regions, and the full MHC region"	Cases only (relapse prediction)	AUC with confidence intervalls (LOOCV)	cross-validation	"Allogeneic haematopoietic stem cell transplantation currently represents the primary potentially curative treatment for cancers of the blood and bone marrow. While relapse occurs in approximately 30% of patients, few risk-modifying genetic variants have been identified. The present study evaluates the predictive potential of patient genetics on relapse risk in a genome-wide manner. [...] Our results show that germline genetic polymorphisms in patients entail a significant contribution to relapse risk, as judged by the predictive performance of the model (AUC = 0.72 [95% CI: 0.63–0.81])."	"Allogeneic haematopoietic stem cell transplantation currently represents the primary potentially curative treatment for cancers of the blood and bone marrow. While relapse occurs in approximately 30% of patients, few risk-modifying genetic variants have been identified. The present study evaluates the predictive potential of patient genetics on relapse risk in a genome-wide manner. [...] Our results show that germline genetic polymorphisms in patients entail a significant contribution to relapse risk, as judged by the predictive performance of the model (AUC = 0.72 [95% CI: 0.63–0.81])."
168	Roder, J and Oliveira, C and Net, L and Tsypin, M and Linstid, B and Roder, H	A dropout-regularized classifier development approach optimized for precision medicine test discovery from omics data	Bmc Bioinformatics	20		14-14	2019	USA	http://dx.doi.org/10.1186/s12859-019-2922-2	article	". For the GSE50081 cohort expression profiling was performed on RNA from frozen, resected tumor tissue from 181 subjects with stage I or II NSCLC [...] Expression profiling for the GSE42127 cohort was performed for 176 subjects with stage I-IV NSCLC "	Case-control study	AUC (training and validation cohort)	external cohort validation	"We describe a novel approach to classifier development designed to create clinically useful tests together with reliable estimates of their performance. The method incorporates elements of traditional and modern machine learning to facilitate the use of cohorts where the number of samples is less than the number of measured patient attributes. It is based on a hierarchy of classification and information abstraction and combines boosting, bagging, and strong dropout regularization. [...] We apply this dropout-regularized combination approach to two clinical problems in oncology using mRNA expression and associated clinical data and compare performance with other methods of classifier generation, including Random Forest. Performance of the new method is similar to or better than the Random Forest in the two classification tasks used for comparison. The dropout-regularized combination method also generates an effective classifier in a classification task with a known confounding variable. Most importantly, it provides a reliable estimate of test performance from a relatively small development set of samples."#	"We describe a novel approach to classifier development designed to create clinically useful tests together with reliable estimates of their performance. The method incorporates elements of traditional and modern machine learning to facilitate the use of cohorts where the number of samples is less than the number of measured patient attributes. It is based on a hierarchy of classification and information abstraction and combines boosting, bagging, and strong dropout regularization. [...] We apply this dropout-regularized combination approach to two clinical problems in oncology using mRNA expression and associated clinical data and compare performance with other methods of classifier generation, including Random Forest. Performance of the new method is similar to or better than the Random Forest in the two classification tasks used for comparison. The dropout-regularized combination method also generates an effective classifier in a classification task with a known confounding variable. Most importantly, it provides a reliable estimate of test performance from a relatively small development set of samples."#

169	Rodriguez-Ortiz, M E and Pontillo, C and Rodriguez, M and Zurbig, P and Mischak, H and Ortiz, A	Novel Urinary Biomarkers For Improved Prediction Of Progressive Egfr Loss In Early Chronic Kidney Disease Stages And In High Risk Individuals Without Chronic Kidney Disease	Sci Rep	8	1	15940-15940	2018	Germany	http://dx.doi.org/10.1038/s41598-018-34386-8	article	Rapid progressors (n = 342) Non-rapid progressors (n = 1140)	Case-control study	AUC, sensitivity, specificity, PPV, NPV (LOOCV + external validation cohort)	cross-validation + external cohort validation	"Chronic kidney disease is associated with increased risk of CKD progression and death. Therapeutic approaches to limit progression are limited. Developing tools for the early identification of those individuals most likely to progress will allow enriching clinical trials in high risk early CKD patients. The CKD273 classifier is a panel of 273 urinary peptides that enables early detection of CKD and prognosis of progression. We have generated urine capillary electrophoresis-mass spectrometry-based peptidomics CKD273 subclassifiers specific for CKD stages to allow the early identification of patients at high risk of CKD progression. [...] In individuals with eGFR > 60 ml/min/1.73 m2 and albuminuria <30 mg/day, the CKD273 subclassifiers predicted rapid eGFR loss with AUC ranging from 0.797 (0.743–0.844) to 0.736 (0.689–0.780). The association between CKD273 subclassifiers and rapid progression remained significant after adjustment for age, sex, albuminuria, DM, baseline eGFR, and systolic blood pressure."	"Chronic kidney disease is associated with increased risk of CKD progression and death. Therapeutic approaches to limit progression are limited. Developing tools for the early identification of those individuals most likely to progress will allow enriching clinical trials in high risk early CKD patients. The CKD273 classifier is a panel of 273 urinary peptides that enables early detection of CKD and prognosis of progression. We have generated urine capillary electrophoresis-mass spectrometry-based peptidomics CKD273 subclassifiers specific for CKD stages to allow the early identification of patients at high risk of CKD progression. [...] In individuals with eGFR > 60 ml/min/1.73 m2 and albuminuria <30 mg/day, the CKD273 subclassifiers predicted rapid eGFR loss with AUC ranging from 0.797 (0.743–0.844) to 0.736 (0.689–0.780). The association between CKD273 subclassifiers and rapid progression remained significant after adjustment for age, sex, albuminuria, DM, baseline eGFR, and systolic blood pressure."
170	Rosenberg, S and Ducray, F and Alentorn, A and Dehais, C and Elarouci, N and Kamoun, A and Marie, Y and Tanguy, M L and De Reynies, A and Mokhtari, K and Figarella-Branger, D and Delattre, J Y and Idbaih, A and Adam, C and Andraud, M and Aubriot-Lorton, M H and Bauchet, L and Beauchesne, P and Bekaert, L and Blechet, C and Campone, M and Carpentier, A and Carpiuc, I and Cazals-Hatem, D and Lhermitte, B and Chiforeanu, D and Chinot, O and Cohen-Moyal, E and Colin, P and Cruel, T and Dam-Hieu, P and Desenclos, C and Desse, N and Dhermain, F and Diebold, M D and Eimer, S and Faillot, T and Fesneau, M and Fontaine, D and Gaillard, S and Forest, F and Gauchotte, G and Gaultier, C and Ghiringhelli, F and Godfraind, C and Gueye, E M and Elouadhani-Hamdi, S and Honnorat, J and Khallil, T and Labrousse, F and Lahiani, W and Langlois, O and Laquerriere, A and Larrieu-Ciron, D and Lechapt-Zalcman, E and Loiseau, H and Lopez, S and Loussouarn, D and Maurage, C A and Menei, P and Mihai, M I and Milin, S and Fotso, M J M and Noel, G and Parker, F and Petit, A and Quintin-Roue, I and Ramirez, C and Rousseau, A and Rousselot-Denis, C and Ricard, D and Richard, P and Rigau, V and Runavot, G and Sevestre, H and Tortel, M C and Vandenbos, F and Vauleon, E and Villa, C and Zemmoura, I and Desenclos, C and Sevestre, H and Menei, P and Rousseau, A and Cruel, T and Lopez, S and Mihai, M I and Petit, A and Adam, C and Parker, F and Carpentier, A and Dam-Hieu, P and Quintin-Roue, I and Eimer, S and Loiseau, H and Bekaert, L and Lechapt-Zalcman, E and Godfraind, C and Khallil, T and Cazals-Hatem, D and Faillot, T and Gaultier, C and Tortel, M C and Carpiuc, I and Richard, P and Lahiani, W and Aubriot-Lorton, H and Ghiringhelli, F and Maurage, C A and Ramirez, C and Gueye, E M and Labrousse, F and Ducray, F and Jouvet, A and Figarella-Branger, D and Chinot, O and Bauchet, L and Rigau, V and Beauchesne, P and Gauchotte, G and Campone, M and Loussouarn, D and Fontaine, D and Vandenbos-Burel, F and Blechet, C and Fesneau, M and Dehais, C and Delattre, J Y and Elouadhani-Hamdi, S and Ricard, D and Larrieu-Ciron, D and Milin, S and Colin, P and Diebold, M D and Chiforeanu, D and Vauleon, E and Langlois, O and Laquerriere, A and Forest, F and Motso-Fotso, M J and Andraud, M and Runavot, G and Lhermitte, B and Noel, G and Gaillard, S and Villa, C and Desse, N and Cohen-Moyal, E and Uro-Coste, E and Dhermain, F and Network, Pola	Machine Learning for Better Prognostic Stratification and Driver Gene Identification Using Somatic Copy Number Variations in Anaplastic Oligodendroglioma	Oncologist	23	12	1500-1510	2018	France	http://dx.doi.org/10.1634/theoncologist.2017-0495	article	97 patients with anaplastic oligodendroglioma	Cases only (prognosis study)	error rate, C-index (cross-validation, external validation)	cross-validation + external cohort validation	"1p/19q-codeleted anaplastic gliomas have variable clinical behavior. We have recently shown that the common 9p21.3 allelic loss is an independent prognostic factor in this tumor type. The aim of this study is to identify less frequent genomic copy number variations (CNVs) with clinical importance that may shed light on molecular oncogenesis of this tumor type. [...] Computational biology and feature selection based on the random forests method were used to identify CNV events associated with overall survival and other clinical-pathological variables. [...] Several recurrent CNV events, detected in anaplastic oligodendroglioma, enable better survival prediction. More importantly, they help in identifying potential genes for understanding oncogenesis and for personalized therapy."	"1p/19q-codeleted anaplastic gliomas have variable clinical behavior. We have recently shown that the common 9p21.3 allelic loss is an independent prognostic factor in this tumor type. The aim of this study is to identify less frequent genomic copy number variations (CNVs) with clinical importance that may shed light on molecular oncogenesis of this tumor type. [...] Computational biology and feature selection based on the random forests method were used to identify CNV events associated with overall survival and other clinical-pathological variables. [...] Several recurrent CNV events, detected in anaplastic oligodendroglioma, enable better survival prediction. More importantly, they help in identifying potential genes for understanding oncogenesis and for personalized therapy."
171	Rychkov, D and Sirota, M and Lin, C	Leveraging publicly available gene expression data and applying machine learning to identify novel biomarkers for rheumatoid arthritis	Arthritis and Rheumatology	70		2206-2206	2018		http://dx.doi.org/10.1002/art.40700	meeting abstract	"For the synovium, we collected 13 datasets with 312 biopsy samples. Among them, there were 276 RA samples and 36 healthy tissue biopsies. For whole blood data, we collected 11 datasets with 2,153 samples: 1,394 RA and 759 healthy controls" (sufficient samples for whole blood)	Case-control study	Cohen's kappa, sensitivity, specificity (5-fold CV)	cross-validation	"Diagnosis and monitoring the disease progression of RA is challenging requiring a combination of imaging techniques and blood tests. There is currently no biochemical test for detection of early-stage disease. In this study, we aimed to define a Rheumatoid Arthritis meta-profile and identify biomarkers by leveraging publicly available gene expression data with machine learning approaches. [...] Finally, we built a Random Forest classification model on the synovium data with these 5 genes. We applied 5-fold cross-validation with 10 repeats technique and used Cohen's Kappa statistic as a metric. We obtained Kappa equals 0.61 with sensitivity 0.86 and specificity 0.9 on the testing set. In the final step, we validated the prediction model on the whole blood data, resulting kappa of 0.57 with sensitivity 0.54 and specificity 0.98."	"Diagnosis and monitoring the disease progression of RA is challenging requiring a combination of imaging techniques and blood tests. There is currently no biochemical test for detection of early-stage disease. In this study, we aimed to define a Rheumatoid Arthritis meta-profile and identify biomarkers by leveraging publicly available gene expression data with machine learning approaches. [...] Finally, we built a Random Forest classification model on the synovium data with these 5 genes. We applied 5-fold cross-validation with 10 repeats technique and used Cohen's Kappa statistic as a metric. We obtained Kappa equals 0.61 with sensitivity 0.86 and specificity 0.9 on the testing set. In the final step, we validated the prediction model on the whole blood data, resulting kappa of 0.57 with sensitivity 0.54 and specificity 0.98."
172	Saini, G and Mittal, K and Rida, P and Janssen, E A M and Gogineni, K and Aneja, R	Panoptic view of prognostic models for personalized breast cancer management	Cancers	11	9		2019	USA	http://dx.doi.org/10.3390/cancers11091325	article	review (not applicable)	review			"The efforts to personalize treatment for patients with breast cancer have led to a focus on the deeper characterization of genotypic and phenotypic heterogeneity among breast cancers. [...] This review summarizes the prognostic and predictive insights provided by commercially available gene expression-based tests and other multivariate or clinical -omics-based prognostic/predictive models currently under development, and proposes a more inclusive multiplatform approach to tackling the challenging heterogeneity of breast cancer to individualize its management."	"The efforts to personalize treatment for patients with breast cancer have led to a focus on the deeper characterization of genotypic and phenotypic heterogeneity among breast cancers. [...] This review summarizes the prognostic and predictive insights provided by commercially available gene expression-based tests and other multivariate or clinical -omics-based prognostic/predictive models currently under development, and proposes a more inclusive multiplatform approach to tackling the challenging heterogeneity of breast cancer to individualize its management."
173	Scheubert, L and Lustrek, M and Schmidt, R and Repsilber, D and Fuellen, G	Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets	BMC Bioinformatics	13		266-266	2012	Germany	http://dx.doi.org/10.1186/1471-2105-13-266	article	PLURI and AD dataset contain > 50 samples per group	Case-control study	accuracy (3-fold CV)	cross-validation	"Alzheimer's disease has been known for more than 100 years and the underlying molecular mechanisms are not yet completely understood. The identification of genes involved in the processes in Alzheimer affected brain is an important step towards such an understanding. [...] Based on microarray data we identify potential biomarkers as well as biomarker combinations using three feature selection methods: information gain, mean decrease accuracy of random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM). [...] Compared to the other methods, GA/SVM has the advantage of finding small, less redundant sets of genes that, in combination, show superior classification characteristics."	"Alzheimer's disease has been known for more than 100 years and the underlying molecular mechanisms are not yet completely understood. The identification of genes involved in the processes in Alzheimer affected brain is an important step towards such an understanding. [...] Based on microarray data we identify potential biomarkers as well as biomarker combinations using three feature selection methods: information gain, mean decrease accuracy of random forest and a wrapper of genetic algorithm and support vector machine (GA/SVM). [...] Compared to the other methods, GA/SVM has the advantage of finding small, less redundant sets of genes that, in combination, show superior classification characteristics."
174	Shafi, A and Nguyen, T and Peyvandipour, A and Nguyen, H and Draghici, S	A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures	Frontiers in Genetics	10			2019	United States	http://dx.doi.org/10.3389/fgene.2019.00159	article	622 samples: 533 samples from GBM patients and 89 from healthy (non-tumor) individuals	Cases only (prognosis study)	Cox p-value (training + validation cohort)	external cohort validation	"Although massive amounts of condition-specific molecular profiles are being accumulated in public repositories every day, meaningful interpretation of these data remains a major challenge. In an effort to identify the biomarkers that describe the key biological phenomena for a given condition, several approaches have been developed over the past few years. However, the majority of these approaches either (i) do not consider the known intermolecular interactions, or (ii) do not integrate molecular data of multiple types (e.g., genomics, transcriptomics, proteomics, epigenomics, etc.), and thus potentially fail to capture the true biological changes responsible for complex diseases (e.g., cancer). In addition, these approaches often ignore the heterogeneity and study bias present in independent molecular cohorts. In this manuscript, we propose a novel multi-cohort and multi-omics meta-analysis framework that overcomes all three limitations mentioned above in order to identify robust molecular subnetworks that capture the key dynamic nature of a given biological condition. [...] We demonstrate the proposed framework by constructing subnetworks related to two complex diseases: glioblastoma and low-grade gliomas. We validate the identified subnetworks by showing their ability to predict patients' clinical outcome on multiple independent validation cohorts."	"Although massive amounts of condition-specific molecular profiles are being accumulated in public repositories every day, meaningful interpretation of these data remains a major challenge. In an effort to identify the biomarkers that describe the key biological phenomena for a given condition, several approaches have been developed over the past few years. However, the majority of these approaches either (i) do not consider the known intermolecular interactions, or (ii) do not integrate molecular data of multiple types (e.g., genomics, transcriptomics, proteomics, epigenomics, etc.), and thus potentially fail to capture the true biological changes responsible for complex diseases (e.g., cancer). In addition, these approaches often ignore the heterogeneity and study bias present in independent molecular cohorts. In this manuscript, we propose a novel multi-cohort and multi-omics meta-analysis framework that overcomes all three limitations mentioned above in order to identify robust molecular subnetworks that capture the key dynamic nature of a given biological condition. [...] We demonstrate the proposed framework by constructing subnetworks related to two complex diseases: glioblastoma and low-grade gliomas. We validate the identified subnetworks by showing their ability to predict patients' clinical outcome on multiple independent validation cohorts."
175	Shaia, K L and Acharya, C R and Smeltzer, S and Lyerly, H K and Acharya, K S	Non-invasive diagnosis of endometriosis: using machine learning instead of the operating room	Fertility and Sterility	112	3	e80-e80	2019		http://dx.doi.org/10.1016/j.fertnstert.2019.07.331	article	"We trained Random Forest classifiers on ten gene-expression based modules, derived from spectral decomposition of the discovery dataset (n 1⁄4 148) to predict the presence of endometriosis"	Case-control study	AUC, accuracy, NPV, PPV (10-fold CV + external test set)	cross-validation + external cohort validation	"Endometriosis affects an estimated 1 in 10 women during their reproductive years, and up to 30% to 50% of women with endometriosis may experience infertility. [...] A previous study developed classifiers for prediction of endometriosis in a cycle-phase specific manner by using margin tree classification within one dataset. Our aim was to build on this research by utilizing machine learning to predict and independently validate the presence or absence of endometriosis, regardless of cycle phase and other uterine pathology, through endometrial biopsy (EMB) samples. [...] We identified a 280-gene predictor of endometriosis using Random Forests that was found to predict the presence of endometriosis, regardless of the endometrial phase and other pathology, with an accuracy of 84% (area under ROC 1⁄4 0.84; p-value: 6.14e-05), with a negative predic- tive value of 86% and a positive predictive value of 81%. We reduced model over-fitting by performing 10-fold cross-validation of our discovery data."	"Endometriosis affects an estimated 1 in 10 women during their reproductive years, and up to 30% to 50% of women with endometriosis may experience infertility. [...] A previous study developed classifiers for prediction of endometriosis in a cycle-phase specific manner by using margin tree classification within one dataset. Our aim was to build on this research by utilizing machine learning to predict and independently validate the presence or absence of endometriosis, regardless of cycle phase and other uterine pathology, through endometrial biopsy (EMB) samples. [...] We identified a 280-gene predictor of endometriosis using Random Forests that was found to predict the presence of endometriosis, regardless of the endometrial phase and other pathology, with an accuracy of 84% (area under ROC 1⁄4 0.84; p-value: 6.14e-05), with a negative predic- tive value of 86% and a positive predictive value of 81%. We reduced model over-fitting by performing 10-fold cross-validation of our discovery data."
176	Shan, L and Chen, Y A and Davis, L and Han, G and Zhu, W and Molina, A D and Arango, H and LaPolla, J P and Hoffman, M S and Sellers, T and Kirby, T and Nicosia, S V and Sutphen, R	Measurement of phospholipids may improve diagnostic accuracy in ovarian cancer	PLoS One	7	10	e46846-e46846	2012	USA	http://dx.doi.org/10.1371/journal.pone.0046846	article	"a total of 1057 women with suspected ovarian cancer were enrolled [...] Only patients who underwent surgery based on clinical suspicion of ovarian cancer were eligible and if a patient was diagnosed with EOC, surgical staging was documented (including 233 in whom EOC was confirmed [...] A total of 211 cases and 212 benigns was included in the analysis."	Case-control study	error rate, sensitivity, specificity (5-fold CV)	cross-validation	"More than two-thirds of women who undergo surgery for suspected ovarian neoplasm do not have cancer. Our previous results suggest phospholipids as potential biomarkers of ovarian cancer. In this study, we measured the serum levels of multiple phospholipids among women undergoing surgery for suspected ovarian cancer to identify biomarkers that better predict whether an ovarian mass is malignant. [...] The HH-SVM model using the measurements of specific combinations of phospholipids supplements clinical CA125 measurement and improves diagnostic accuracy. Specifically, the measurement of phospholipids improved sensitivity (identification of cases with preoperative CA125 levels below 35) among two types of cases in which CA125 performance is historically poor - early stage cases and those of mucinous histology. Measurement of phospholipids improved the identification of early stage cases from 65% (based on CA125) to 82%, and mucinous cases from 44% to 88%."	"More than two-thirds of women who undergo surgery for suspected ovarian neoplasm do not have cancer. Our previous results suggest phospholipids as potential biomarkers of ovarian cancer. In this study, we measured the serum levels of multiple phospholipids among women undergoing surgery for suspected ovarian cancer to identify biomarkers that better predict whether an ovarian mass is malignant. [...] he HH-SVM model using the measurements of specific combinations of phospholipids supplements clinical CA125 measurement and improves diagnostic accuracy. Specifically, the measurement of phospholipids improved sensitivity (identification of cases with preoperative CA125 levels below 35) among two types of cases in which CA125 performance is historically poor - early stage cases and those of mucinous histology. Measurement of phospholipids improved the identification of early stage cases from 65% (based on CA125) to 82%, and mucinous cases from 44% to 88%."
177	Shao, C H and Chen, C L and Lin, J Y and Chen, C J and Fu, S H and Chen, Y T and Chang, Y S and Yu, J S and Tsui, K H and Juo, C G and Wu, K P	Metabolite marker discovery for the detection of bladder cancer by comparative metabolomics	Oncotarget	8	24	38802-38810	2017	China	http://dx.doi.org/10.18632/oncotarget.16393	article	" metabolite profiles of 87 samples from bladder cancer patients and 65 samples from hernia patients"	Case-control study	AUC, accuracy, sensitivity, specificity (5-fold cross-validation + test set)	cross-validation + test set	"In this study, we applied ultra-performance liquid chromatography time-of-flight mass spectrometry to profile metabolite profiles of 87 samples from bladder cancer patients and 65 samples from hernia patients. An OPLS-DA classification revealed that bladder cancer samples can be discriminated from hernia samples based on the profiles. A marker discovery pipeline selected six putative markers from the metabolomic profiles. [...] A machine learning model, decision trees, was built based on the metabolomic profiles and the six marker candidates. The decision tree obtained an accuracy of 76.60%, a sensitivity of 71.88%, and a specificity of 86.67% from an independent test."	"In this study, we applied ultra-performance liquid chromatography time-of-flight mass spectrometry to profile metabolite profiles of 87 samples from bladder cancer patients and 65 samples from hernia patients. An OPLS-DA classification revealed that bladder cancer samples can be discriminated from hernia samples based on the profiles. A marker discovery pipeline selected six putative markers from the metabolomic profiles. [...] A machine learning model, decision trees, was built based on the metabolomic profiles and the six marker candidates. The decision tree obtained an accuracy of 76.60%, a sensitivity of 71.88%, and a specificity of 86.67% from an independent test."

178	Sharma, A and Rani, R	C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods	Comput Methods Programs Biomed	178		219-235	2019	India	http://dx.doi.org/10.1016/j.cmpb.2019.06.029	article	the proposed machine learning approachwas tested on 7 microarray datasets, including a datasets with > 50 samples per group	Case-control study	accuracy (LOOCV + test set)	cross-validation + test set	"we have proposed framework (C-HMOSHSSA) for gene selection using multi-objective spotted hyena optimizer (MOSHO) and salp swarm algorithm (SSA). The real-life optimization problems with more than one objective usually face the challenge to maintain convergence and diversity. Salp Swarm Algorithm (SSA) maintains diversity but, suffers from the overhead of main- taining the necessary information. On the other hand, the calculation of MOSHO requires low computa- tional efforts hence is used for maintaining the necessary information. Therefore, the proposed algorithm is a hybrid algorithm that utilizes the features of both SSA and MOSHO to facilitate its exploration and exploitation capability. [...] Four different classifiers are trained on seven high-dimensional datasets using a subset of features (genes), which are obtained after applying the proposed hybrid gene selection algorithm. The results show that the proposed technique significantly outperforms existing state-of-the-art techniques."	"we have proposed framework (C-HMOSHSSA) for gene selection using multi-objective spotted hyena optimizer (MOSHO) and salp swarm algorithm (SSA). The real-life optimization problems with more than one objective usually face the challenge to maintain convergence and diversity. Salp Swarm Algorithm (SSA) maintains diversity but, suffers from the overhead of main- taining the necessary information. On the other hand, the calculation of MOSHO requires low computa- tional efforts hence is used for maintaining the necessary information. Therefore, the proposed algorithm is a hybrid algorithm that utilizes the features of both SSA and MOSHO to facilitate its exploration and exploitation capability. [...] Four different classifiers are trained on seven high-dimensional datasets using a subset of features (genes), which are obtained after applying the proposed hybrid gene selection algorithm. The results show that the proposed technique significantly outperforms existing state-of-the-art techniques."
179	Shen, L and Tan, E C	Dimension reduction-based penalized logistic regression for cancer classification using microarray data	IEEE/ACM Trans Comput Biol Bioinform	2	2	166-175	2005	Singapore	http://dx.doi.org/10.1109/tcbb.2005.22	article	the proposed machine learning approachwas tested on 7 microarray datasets, including a datasets with > 50 samples per group	Case-control study	mean error + standard deviation (LOOCV, test set)	cross-validation + test set	"The use of penalized logistic regression for cancer classification using microarray expression data is presented. Two dimension reduction methods are respectively combined with the penalized logistic regression so that both the classification accuracy and computational speed are enhanced. Two other machine-learning methods, support vector machines and least-squares regression, have been chosen for comparison. It is shown that our methods have achieved at least equal or better results. They also have the advantage that the output probability can be explicitly given and the regression coefficients are easier to interpret."	"The use of penalized logistic regression for cancer classification using microarray expression data is presented. Two dimension reduction methods are respectively combined with the penalized logistic regression so that both the classification accuracy and computational speed are enhanced. Two other machine-learning methods, support vector machines and least-squares regression, have been chosen for comparison. It is shown that our methods have achieved at least equal or better results. They also have the advantage that the output probability can be explicitly given and the regression coefficients are easier to interpret."
180	Sherman, S I and Pagan, M and Huang, J and Lin, B and Diggans, J and Tom, E and Haugen, B and Tuttle, R M and Kennedy, G	Augmenting pre-operative risk of recurrence stratification in differentiated thyroid carcinoma using machine learning and high dimensional transcriptional data from thyroid FNA	Journal of Clinical Oncology	33	15		2015		http://meeting.ascopubs.org/cgi/content/abstract/33/15_suppl/6044?sid=0dbce579-6935-4171-971e-51df4209b3af	meeting abstract	"81 samples preoperatively collected in a previous study and post-surgically diagnosed as PTC [...] Each patient was categorized as either ATA low risk or ATA intermediate/high risk using established guidelines for recurrence risk stratification." (< 50 samples per group)	Cases only (risk of recurrence prediction)	AUC (cross-validation)	cross-validation	"Transcriptional data from FNA of thyroid nodules may improve the pre-operative prediction of risk for post-operative recurrence. If independently validated in a sufficiently large number of patients, such molecular classifiers may augment initial risk stratification and individualization of patient care"	"Transcriptional data from FNA of thyroid nodules may improve the pre-operative prediction of risk for post-operative recurrence. If independently validated in a sufficiently large number of patients, such molecular classifiers may augment initial risk stratification and individualization of patient care"
181	Shi, P and Ray, S and Zhu, Q F and Kon, M A	Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction	Bmc Bioinformatics	12		15-15	2011	USA	http://dx.doi.org/10.1186/1471-2105-12-375	article	4 cancer prognosis microarray datasets, including data with > 50 samples per group	Case-control study	error rate (LOOCV, test set)	cross-validation + test set	"The widely used k top scoring pair (k-TSP) algorithm is a simple yet powerful parameter-free classifier. It owes its success in many cancer microarray datasets to an effective feature selection algorithm that is based on relative expression ordering of gene pairs. However, its general robustness does not extend to some difficult datasets, such as those involving cancer outcome prediction, which may be due to the relatively simple voting scheme used by the classifier. We believe that the performance can be enhanced by separating its effective feature selection component and combining it with a powerful classifier such as the support vector machine (SVM). [...] We developed an approach integrating the k-TSP ranking algorithm (TSP) with other machine learning methods, allowing combination of the computationally efficient, multivariate feature ranking of k-TSP with multivariate classifiers such as SVM. We evaluated this hybrid scheme (k-TSP+SVM) in a range of simulated datasets with known data structures. As compared with other feature selection methods, such as a univariate method similar to Fisher's discriminant criterion (Fisher), or a recursive feature elimination embedded in SVM (RFE), TSP is increasingly more effective than the other two methods as the informative genes become progressively more correlated, which is demonstrated both in terms of the classification performance and the ability to recover true informative genes."	"The widely used k top scoring pair (k-TSP) algorithm is a simple yet powerful parameter-free classifier. It owes its success in many cancer microarray datasets to an effective feature selection algorithm that is based on relative expression ordering of gene pairs. However, its general robustness does not extend to some difficult datasets, such as those involving cancer outcome prediction, which may be due to the relatively simple voting scheme used by the classifier. We believe that the performance can be enhanced by separating its effective feature selection component and combining it with a powerful classifier such as the support vector machine (SVM). [...] We developed an approach integrating the k-TSP ranking algorithm (TSP) with other machine learning methods, allowing combination of the computationally efficient, multivariate feature ranking of k-TSP with multivariate classifiers such as SVM. We evaluated this hybrid scheme (k-TSP+SVM) in a range of simulated datasets with known data structures. As compared with other feature selection methods, such as a univariate method similar to Fisher's discriminant criterion (Fisher), or a recursive feature elimination embedded in SVM (RFE), TSP is increasingly more effective than the other two methods as the informative genes become progressively more correlated, which is demonstrated both in terms of the classification performance and the ability to recover true informative genes."
182	Sinnott, J A and Cai, T	Omnibus risk assessment via accelerated failure time kernel machine modeling	Biometrics	69	4	861-873	2013	United States	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869038/	article	"training set of 454 lymph node negative breast cancer patients [...] A total of 119 deaths or recurrences were observed"	Case-control study	C-statistic (training + validation data)	cross-validation + test set	"Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and possible complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. [...] In recent years, much work has been done to group genes into pathways and networks. Integrating such biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which can capture nonlinear effects if nonlinear kernels are used [...] In this article, we derive testing and prediction methods for KM regression under the accelerated failure time (AFT) model, a useful alternative to the PH model. We approximate the null distribution of our test statistic using resampling procedures. When multiple kernels are of potential interest, it may be unclear in advance which kernel to use for testing and estimation. We propose a robust Omnibus Test that combines information across kernels, and an approach for selecting the best kernel for estimation. The methods are illustrated with an application in breast cancer."	"Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and possible complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. [...] In recent years, much work has been done to group genes into pathways and networks. Integrating such biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which can capture nonlinear effects if nonlinear kernels are used [...] In this article, we derive testing and prediction methods for KM regression under the accelerated failure time (AFT) model, a useful alternative to the PH model. We approximate the null distribution of our test statistic using resampling procedures. When multiple kernels are of potential interest, it may be unclear in advance which kernel to use for testing and estimation. We propose a robust Omnibus Test that combines information across kernels, and an approach for selecting the best kernel for estimation. The methods are illustrated with an application in breast cancer."
183	Sosa, J and Angell, T E and Babiarz, J and Barth, N and Blevins, T and Duh, Q and Ghossein, R A and Harrell, R M and Huang, J and Imtiaz, U and Kennedy, G and Kim, S and Kloos, R T and LiVolsi, V A and Patel, K N and Randolph, G and Sadow, P M and Shanik, M H and Traweek, S T and Walsh, P S and Whitney, D and Yeh, M and Ladenson, P W	Clinical validation of the AFIRMA genomic sequencing parathyroid classifier	Thyroid	27		A50-A51	2017		http://dx.doi.org/10.1089/thy.2017.29046.abstracts	meeting abstract	476 FNAs-6 parathyroid and 470 thyroid FNAs	Case-control study	sensitivity, specificity (training and validation cohort)	cross-validation + external cohort validation	"The parathyroid glands are located adjacent to the thyroid and occasionally within it. Enlarged and imbedded parathyroid glands can be mistaken as thyroid nodules or suspicious lymph nodes. On fine needle aspiration biopsy (FNAB) of such lesions, cytology is often indeterminate, failing to identify its parathyroid origin and potentially resulting in an unnecessary thyroid surgery. The Afirma Genomic Sequencing Classifier (GSC) identifies genomically benign thyroid nodules among those with indeterminate FNAB to prevent unnecessary diagnostic surgery using RNA sequencing and machine learning algorithms. [...] The final classifier was blindly tested on an independent test set of 195 FNAs (118 Bethesda III, 77 Bethesda IV). The classifier had 100% sensitivity [4/4 parathyroid correctly called positive; CI 39.8-100%] and 100% specificity [191/191 thyroid correctly called negative; CI 98.1-100%]."	
184	Srivastava, S and Wang, W and Manyam, G and Ordonez, C and Baladandayuthapani, V	Integrating multi-platform genomic data using hierarchical Bayesian relevance vector machines	Eurasip Journal on Bioinformatics and Systems Biology	2013	1		2013		http://dx.doi.org/10.1186/1687-4153-2013-9	article	"GBM data have multiple molecular measurements on over 500 samples that include gene expression, copy number, methylation and microRNA expression."	Case-control study	mean square prediction error ("We randomly split the GBM survival data into a training data and a test data with 223 (90%) and 25 (10%) patients, respectively")	training + test set	"For predicting relevant clinical outcomes, we propose a flexible statistical machine learning approach that acknowledges and models the interaction between platform-specific measurements through nonlinear kernel machines and borrows information within and between platforms through a hierarchical Bayesian framework. Our model has parameters with direct interpretations in terms of the effects of platforms and data interactions within and across platforms. The parameter estimation algorithm in our model uses a computationally efficient variational Bayes approach that scales well to large high-throughput datasets. [...] We apply our methods of integrating gene/mRNA expression and microRNA profiles for predicting patient survival times to The Cancer Genome Atlas (TCGA) based glioblastoma multiforme (GBM) dataset. In terms of prediction accuracy, we show that our non-linear and interaction-based integrative methods perform better than linear alternatives and non-integrative methods that do not account for interactions between the platforms. We also find several prognostic mRNAs and microRNAs that are related to tumor invasion and are known to drive tumor metastasis and severe inflammatory response in GBM. [...] Our approach gains its flexibility and power by modeling the non-linear interaction structures between and within the platforms."	"For predicting relevant clinical outcomes, we propose a flexible statistical machine learning approach that acknowledges and models the interaction between platform-specific measurements through nonlinear kernel machines and borrows information within and between platforms through a hierarchical Bayesian framework. Our model has parameters with direct interpretations in terms of the effects of platforms and data interactions within and across platforms. The parameter estimation algorithm in our model uses a computationally efficient variational Bayes approach that scales well to large high-throughput datasets. [...] We apply our methods of integrating gene/mRNA expression and microRNA profiles for predicting patient survival times to The Cancer Genome Atlas (TCGA) based glioblastoma multiforme (GBM) dataset. In terms of prediction accuracy, we show that our non-linear and interaction-based integrative methods perform better than linear alternatives and non-integrative methods that do not account for interactions between the platforms. We also find several prognostic mRNAs and microRNAs that are related to tumor invasion and are known to drive tumor metastasis and severe inflammatory response in GBM. [...] Our approach gains its flexibility and power by modeling the non-linear interaction structures between and within the platforms."
185	Stamate, D and Kim, M and Proitsi, P and Westwood, S and Baird, A and Nevado-Holgado, A and Hye, A and Bos, I and Vos, S J B and Vandenberghe, R and Teunissen, C E and Kate, M T and Scheltens, P and Gabel, S and Meersmans, K and Blin, O and Richardson, J and De Roeck, E and Engelborghs, S and Sleegers, K and Bordet, R and Ramit, L and Kettunen, P and Tsolaki, M and Verhey, F and Alcolea, D and Lléo, A and Peyratout, G and Tainta, M and Johannsen, P and Freund-Levi, Y and Frölich, L and Dobricic, V and Frisoni, G B and Molinuevo, J L and Wallin, A and Popp, J and Martinez-Lage, P and Bertram, L and Blennow, K and Zetterberg, H and Streffer, J and Visser, P J and Lovestone, S and Legido-Quigley, C	A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort	Alzheimer's and Dementia: Translational Research and Clinical Interventions	5		933-938	2019	Belgium	http://dx.doi.org/10.1016/j.trci.2019.11.001	article	242 cognitively normal (CN) people and 115 with AD-type dementia utilizing plasma metabolites	Case-control study	AUC (nested cross-validation, external test set)	cross-validation + external cohort validation	"Machine learning (ML) may harbor the potential to capture the metabolic complexity in Alzheimer Disease (AD). Here we set out to test the performance of metabolites in blood to categorize AD when compared to CSF biomarkers. [...] Deep Learning (DL), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) were used to differentiate AD from CN. These models were internally validated using Nested Cross Validation (NCV). [...] On the test data, DL produced the AUC of 0.85 (0.80–0.89), XGBoost produced 0.88 (0.86–0.89) and RF produced 0.85 (0.83–0.87). By comparison, CSF measures of amyloid, p-tau and t-tau (together with age and gender) produced with XGBoost the AUC values of 0.78, 0.83 and 0.87, respectively. [...] This study showed that plasma metabolites have the potential to match the AUC of well-established AD CSF biomarkers in a relatively small cohort."	"Machine learning (ML) may harbor the potential to capture the metabolic complexity in Alzheimer Disease (AD). Here we set out to test the performance of metabolites in blood to categorize AD when compared to CSF biomarkers. [...] Deep Learning (DL), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) were used to differentiate AD from CN. These models were internally validated using Nested Cross Validation (NCV). [...] On the test data, DL produced the AUC of 0.85 (0.80–0.89), XGBoost produced 0.88 (0.86–0.89) and RF produced 0.85 (0.83–0.87). By comparison, CSF measures of amyloid, p-tau and t-tau (together with age and gender) produced with XGBoost the AUC values of 0.78, 0.83 and 0.87, respectively. [...] This study showed that plasma metabolites have the potential to match the AUC of well-established AD CSF biomarkers in a relatively small cohort."
186	Statnikov, A and Aliferis, C F and Tsamardinos, I and Hardin, D and Levy, S	A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis	Bioinformatics	21	5	631-643	2005	USA	http://dx.doi.org/10.1093/bioinformatics/bti033	article	11 datasets spanning 74 diagnostic categories and 41 cancer types and 12 normal tissue types	Case-control study	accuracy, relative classifier information (Design I: nested stratified 10-fold CV outer loop, 9-fold CV inner loop, Design 2: nested LOOCV outer loop, 10-fold CV inner loop)	cross-validation	"[...] we performed a systematic and comprehensive evaluation of several major algorithms for multicategory classification, several gene selection methods, multiple ensemble classifier methods and two cross-validation designs using 11 datasets spanning 74 diagnostic categories and 41 cancer types and 12 normal tissue types. [...] Multicategory support vector machines (MC-SVMs) are the most effective classifiers in performing accurate cancer diagnosis from gene expression data. The MC-SVM techniques by Crammer and Singer, Weston and Watkins and one-versus-rest were found to be the best methods in this domain. MC-SVMs outperform other popular machine learning algorithms, such as k-nearest neighbors, backpropagation and probabilistic neural networks, often to a remarkable degree. Gene selection techniques can significantly improve the classification performance of both MC-SVMs and other non-SVM learning algorithms. Ensemble classifiers do not generally improve performance of the best non-ensemble models. These results guided the construction of a software system GEMS (Gene Expression Model Selector) that automates high-quality model construction and enforces sound optimization and performance estimation procedures."	"[...] we performed a systematic and comprehensive evaluation of several major algorithms for multicategory classification, several gene selection methods, multiple ensemble classifier methods and two cross-validation designs using 11 datasets spanning 74 diagnostic categories and 41 cancer types and 12 normal tissue types. [...] Multicategory support vector machines (MC-SVMs) are the most effective classifiers in performing accurate cancer diagnosis from gene expression data. The MC-SVM techniques by Crammer and Singer, Weston and Watkins and one-versus-rest were found to be the best methods in this domain. MC-SVMs outperform other popular machine learning algorithms, such as k-nearest neighbors, backpropagation and probabilistic neural networks, often to a remarkable degree. Gene selection techniques can significantly improve the classification performance of both MC-SVMs and other non-SVM learning algorithms. Ensemble classifiers do not generally improve performance of the best non-ensemble models. These results guided the construction of a software system GEMS (Gene Expression Model Selector) that automates high-quality model construction and enforces sound optimization and performance estimation procedures."

187	Stetson, L C and Pearl, T and Chen, Y W and Barnholtz-Sloan, J S	Computational identification of multi-omic correlates of anticancer therapeutic response	Bmc Genomics	15		8-8	2014	USA	https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-S7-S2	article	"To develop multi-omic predictors of anticancer therapeutic response we curated data from the CCLE, CGP, and NCI60 databases. The resulting datasets consisted of the gene expression (Affymetrix U133A and Affymetrix U133A plus 2.0), copy number variation (Affymetrix SNP6.0), and mutational status (targeted and whole exome sequencing) of 1299 distinct human cancer cell lines representing 35 cancer types."	Cases only (drug response study)	precision + standard deviation of precision (10-fold CV, external validation data)	cross-validation	"Three large-scale pharmacogenomic studies have screened anticancer compounds in greater than 1000 distinct human cancer cell lines. We combined these datasets to generate and validate multi-omic predictors of drug response. We compared drug response signatures built using a penalized linear regression model and two non-linear machine learning techniques, random forest and support vector machine. [...] Multi-omic predictors of drug response can be generated and validated for many drugs. Specifically, the random forest algorithm generated more precise and robust prediction signatures when compared to support vector machines and the more commonly used elastic net regression. The resulting drug response signatures can be used to stratify patients into treatment groups based on their individual tumor biology [...]"	"Three large-scale pharmacogenomic studies have screened anticancer compounds in greater than 1000 distinct human cancer cell lines. We combined these datasets to generate and validate multi-omic predictors of drug response. We compared drug response signatures built using a penalized linear regression model and two non-linear machine learning techniques, random forest and support vector machine. [...] Multi-omic predictors of drug response can be generated and validated for many drugs. Specifically, the random forest algorithm generated more precise and robust prediction signatures when compared to support vector machines and the more commonly used elastic net regression. The resulting drug response signatures can be used to stratify patients into treatment groups based on their individual tumor biology [...]"
188	Sweatt, A J and Hedlin, H K and Balasubramanian, V and Hsi, A and Blum, L K and Robinson, W H and Haddad, F and Hickey, P M and Condliffe, R and Lawrie, A and Nicolls, M R and Rabinovitch, M and Khatri, P and Zamanian, R T	Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension	Circulation Research	124	6	904-919	2019	United Kingdom	http://dx.doi.org/10.1161/circresaha.118.313911	article	"In a prospective observational study of group 1 PAH patients evaluated at Stanford University (discovery cohort; n=281) and University of Sheffield (validation cohort; n=104) between 2008 and 2014, we measured a circulating proteomic panel of 48 cytokines, chemokines, and factors using multiplex immunoassay."	Case-control study	log-rank test p-value (discovery + validation cohort)	external cohort validation	"In a prospective observational study of group 1 PAH patients evaluated at Stanford University (discovery cohort; n=281) and University of Sheffield (validation cohort; n=104) between 2008 and 2014, we measured a circulating proteomic panel of 48 cytokines, chemokines, and factors using multiplex immunoassay. Unsupervised machine learning (consensus clustering) was applied in both cohorts independently to classify patients into proteomic immune clusters, without guidance from clinical features. [...] Findings were replicated in the validation cohort, where machine learning classified 4 immune clusters with comparable proteomic, clinical, and prognostic features."	"In a prospective observational study of group 1 PAH patients evaluated at Stanford University (discovery cohort; n=281) and University of Sheffield (validation cohort; n=104) between 2008 and 2014, we measured a circulating proteomic panel of 48 cytokines, chemokines, and factors using multiplex immunoassay. Unsupervised machine learning (consensus clustering) was applied in both cohorts independently to classify patients into proteomic immune clusters, without guidance from clinical features. [...] Findings were replicated in the validation cohort, where machine learning classified 4 immune clusters with comparable proteomic, clinical, and prognostic features."
189	Tabari, E and Lovejoy, A F and Lin, H and Bolen, C R and Saelee, S L and Lefkowitz, J P and Kurtz, D M and Vitazka, P and Venstrom, J M and Nielsen, T G and Parreira, J M and Klass, D M and Luong, K T	Molecular characteristics and disease burden metrics determined by next-generation sequencing on circulating tumor DNA correlate with progression free survival in previously untreated diffuse large B-cell lymphoma	Blood	134			2019		http://dx.doi.org/10.1182/blood-2019-123633	meeting abstract	"targeted NGS on plasma samples from 310 previously untreated DLBCL pts enrolled in the GOYA study"	Cases only (predicting progression-free surivial)	Correlation with progression-free survival (training/test set split)	training + test set	"Due to the range of biological and molecular heterogeneity in diffuse large B-cell lymphoma (DLBCL), personalized risk stratification and treatment is a promising avenue to improving outcomes. [...] We performed targeted NGS on plasma samples from 310 previously untreated DLBCL pts enrolled in the GOYA study (NCT01287741) with a custom DLBCL-specific panel using a workflow optimized for ctDNA. [...] We describe a single NGS-based method, which calls variants, determines COO, and assesses tumor burden from plasma. Using these results, we show that pre-treatment plasma-based molecular and tumor burden measurements in previously untreated DLBCL pts correlate with PFS."	"Due to the range of biological and molecular heterogeneity in diffuse large B-cell lymphoma (DLBCL), personalized risk stratification and treatment is a promising avenue to improving outcomes. [...] We performed targeted NGS on plasma samples from 310 previously untreated DLBCL pts enrolled in the GOYA study (NCT01287741) with a custom DLBCL-specific panel using a workflow optimized for ctDNA. [...] We describe a single NGS-based method, which calls variants, determines COO, and assesses tumor burden from plasma. Using these results, we show that pre-treatment plasma-based molecular and tumor burden measurements in previously untreated DLBCL pts correlate with PFS."
190	Tan, A C and Gilbert, D	Ensemble machine learning on gene expression data for cancer classification	Appl Bioinformatics	2	3	S75-83	2003	UK	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.9189&rep=rep1&type=pdf	article	Seven microarray datasets were used, including data with > 50 samples group	Case-control study	accuracy, sensitivity, specificity, PPV (10-fold CV)	cross-validation	"In this paper, we focus on three different supervised machine learning techniques in cancer classification, namely C4.5 decision tree, and bagged and boosted decision trees. We have performed classification tasks on seven publicly available cancerous microarray data and compared the classification/prediction performance of these methods. We have observed that ensemble learning (bagged and boosted decision trees) often performs better than single decision trees in this classification task."	"In this paper, we focus on three different supervised machine learning techniques in cancer classification, namely C4.5 decision tree, and bagged and boosted decision trees. We have performed classification tasks on seven publicly available cancerous microarray data and compared the classification/prediction performance of these methods. We have observed that ensemble learning (bagged and boosted decision trees) often performs better than single decision trees in this classification task."
191	Tan, A C and Naiman, D Q and Xu, L and Winslow, R L and Geman, D	Simple decision rules for classifying human cancers from gene expression profiles	Bioinformatics	21	20	3896-3904	2005	USA	http://dx.doi.org/10.1093/bioinformatics/bti631	article	"19 publicly available microarray datasets, with sample sizes ranging from 33 to 327" (more than 50 samples per group for multiple datasets)	Case-control study	accuracy (LOOCV, test set)	cross-validation + test set	"Various studies have shown that cancer tissue samples can be successfully detected and classified by their gene expression patterns using machine learning approaches. One of the challenges in applying these techniques for classifying gene expression data is to extract accurate, readily interpretable rules providing biological insight as to how classification is performed. [...] In this study, we have compared our approach to other machine learning techniques for class prediction in 19 binary and multi-class gene expression datasets involving human cancers. The k-TSP classifier performs as efficiently as Prediction Analysis of Microarray and support vector machine, and outperforms other learning methods (decision trees, k-nearest neighbour and naïve Bayes). Our approach is easy to interpret as the classifier involves only a small number of informative genes."	"Various studies have shown that cancer tissue samples can be successfully detected and classified by their gene expression patterns using machine learning approaches. One of the challenges in applying these techniques for classifying gene expression data is to extract accurate, readily interpretable rules providing biological insight as to how classification is performed. [...] In this study, we have compared our approach to other machine learning techniques for class prediction in 19 binary and multi-class gene expression datasets involving human cancers. The k-TSP classifier performs as efficiently as Prediction Analysis of Microarray and support vector machine, and outperforms other learning methods (decision trees, k-nearest neighbour and naïve Bayes). Our approach is easy to interpret as the classifier involves only a small number of informative genes."
192	Tang, K L and Li, T H and Xiong, W W and Chen, K	Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data	BMC Bioinformatics	11		109-109	2010	China	http://dx.doi.org/10.1186/1471-2105-11-109	article	"high-resolution SELDI-TOF ovarian data set for 95 control samples and 121 cancer samples"	Case-control study	accuracy, sensitivity, specificity (cross-validation)	cross-validation	"Recent advances in proteomics technologies such as SELDI-TOF mass spectrometry has shown promise in the detection of early stage cancers. However, dimensionality reduction and classification are considerable challenges in statistical machine learning. We therefore propose a novel approach for dimensionality reduction and tested it using published high-resolution SELDI-TOF data for ovarian cancer. [...] The method achieved average sensitivity of 0.9950, specificity of 0.9916, accuracy of 0.9935 and a correlation coefficient of 0.9869 for 100 five-fold cross validations. Furthermore, only one control was misclassified in leave-one-out cross validation."	"Recent advances in proteomics technologies such as SELDI-TOF mass spectrometry has shown promise in the detection of early stage cancers. However, dimensionality reduction and classification are considerable challenges in statistical machine learning. We therefore propose a novel approach for dimensionality reduction and tested it using published high-resolution SELDI-TOF data for ovarian cancer. [...] The method achieved average sensitivity of 0.9950, specificity of 0.9916, accuracy of 0.9935 and a correlation coefficient of 0.9869 for 100 five-fold cross validations. Furthermore, only one control was misclassified in leave-one-out cross validation."
193	Tao, M and Song, T and Du, W and Han, S and Zuo, C and Li, Y and Wang, Y and Yang, Z	Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data	Genes (Basel)	10	3		2019	China	http://dx.doi.org/10.3390/genes10030200	article	"Our dataset contained 606 distinct patient samples of breast cancer, which was divided into five subtypes: 277 luminal A, 40 luminal B, 70 Triple Negative Breast Cancer (TNBC), 11 HER2 (+), and 208 unclear"	Case-control study	accuracy, AUC (10-fold CV)	cross-validation	"It is very significant to explore the intrinsic differences in breast cancer subtypes. These intrinsic differences are closely related to clinical diagnosis and designation of treatment plans. [...] In this article, we use estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) to define breast cancer subtypes and classify any two breast cancer subtypes using SMO-MKL algorithm. We collected mRNA data, methylation data and copy number variation (CNV) data from TCGA to classify breast cancer subtypes. Multiple Kernel Learning (MKL) is employed to use these omics data distinctly. The result of using three omics data with multiple kernels is better than that of using single omics data with multiple kernels."	"It is very significant to explore the intrinsic differences in breast cancer subtypes. These intrinsic differences are closely related to clinical diagnosis and designation of treatment plans. [...] In this article, we use estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) to define breast cancer subtypes and classify any two breast cancer subtypes using SMO-MKL algorithm. We collected mRNA data, methylation data and copy number variation (CNV) data from TCGA to classify breast cancer subtypes. Multiple Kernel Learning (MKL) is employed to use these omics data distinctly. The result of using three omics data with multiple kernels is better than that of using single omics data with multiple kernels."
194	Tebani, A and Afonso, C and Marret, S and Bekri, S	Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations	Int J Mol Sci	17	9		2016	France	http://dx.doi.org/10.3390/ijms17091555	article	review (not applicable)	review	"In this review, we present state-of-the-art multi-omics data analysis strategies in a clinical context. The challenges of omics-based biomarker translation are discussed. Perspectives regarding the use of multi-omics approaches for inborn errors of metabolism (IEM) are presented by introducing a new paradigm shift in addressing IEM investigations in the post-genomic era."		"The small number of multi-omics datasets in the field of [Inborn Errors of Metabolism] (IEM) and the lack of standardized and harmonized protocols affect the wide dissemination of these approaches. To overcome these drawbacks, attention should be given to validation strategies at all stages. Moreover, the development of new analytical and machine learning methods will facilitate analysis of multi-tissue and multi-organ data, thus enabling a real investigation of systemic effects [95,141,163]. Extended and effective resources for biobanking are also essential to ensure consistency. Addressing these challenges will improve healthcare management of IEM by moving from a reactive, targeted, and reductionist approach to a more proactive, global, and integrative one."	
195	Theofilatos, K and Korfiati, A and Mavroudi, S and Cowperthwaite, M C and Shpak, M	Discovery of stroke-related blood biomarkers from gene expression network models	BMC Med Genomics	12	1	118-118	2019	Greece	http://dx.doi.org/10.1186/s12920-019-0566-8	article	blood samples from 82 stroke patients and 68 controls	Case-control study	accuracy (5-fold cross validation)	cross-validation	"Identifying molecular biomarkers characteristic of ischemic stroke has the potential to aid in distinguishing stroke cases from stroke mimicking symptoms, as well as advancing the understanding of the physiological changes that underlie the body's response to stroke. This study uses machine learning-based analysis of gene co-expression to identify transcription patterns characteristic of patients with acute ischemic stroke. [...] A predictive model with 89.6% accuracy was identified using 6 network-central and differentially expressed genes (ID3, MBTPS1, NOG, SFXN2, BMX, SLC22A1), characterized by large differences in association network connectivity between stroke and control samples. In contrast, classification models based solely on individual genes identified by significant fold-changes in expression level provided lower predictive accuracies: < 71% for any single gene, and even models with larger (10–25) numbers of gene transcript biomarkers gave lower predictive accuracies (≤ 82%) than the 6 network-based gene signature classification. [...] Network-based models have the potential to identify a more statistically robust pattern of gene expression typical of acute ischemic stroke and to generate hypotheses about possible interactions among functionally relevant genes, leading to the identification of more informative biomarkers."	
196	Toh, T S and Dondelinger, F and Wang, D	Looking beyond the hype: Applied AI and machine learning in translational medicine	Ebiomedicine	47		607-615	2019	UK	http://dx.doi.org/10.1016/j.ebiom.2019.08.027	article	review (not applicable)	review			"Both new and old techniques of artificial intelligence (AI) and machine learning (ML) can now help increase the success of translational studies in three areas: drug discovery, imaging, and genomic medicine. However, ML technologies do not come without their limitations and shortcomings. Current technical limitations and other limitations including governance, reproducibility, and interpretation will be discussed in this article.	"There are general non-technical issues that are required to be addressed before mainstream application of ML within translational medicine takes place. Oftentimes, sensitive data is required to train ML algorithms. Access to data should be carefully regulated to ensure privacy without stifling innovation and technological advancement to improve outcomes [57]. A proposed scheme called privacy-preserving clinical decision with cloud support (PPCD) is an encouraging step in this direction [58]. Biases within the training datasets of ML algorithms need to be avoided to reduce the risk of failure of ML methods to generalise. Rethinking responsibility and accountability of individuals or organisations selecting datasets used to train ML algorithms are key to address this. Ethical frameworks should be developed by scientific committees and regulatory bodies to recognise and minimise the effect of biased models while guiding design choices to introduce systems that build trust, understanding, and maintaining individual privacy [57,59]. Reproducibility is another aspect that needs to be managed to ensure widespread adoption of ML in translational medicine. Caution should be exercised when drawing conclusions solely from large reservoirs of clinical data as it is often fraught with heterogeneity in quality [60]. Responsibly sharing data and code should be made requirements for authors alongside their publications to ensure trust in research findings. Reproducibility of results requires the software environment, source code, and raw data used during experiments [61]. Lastly, there is also a need to explain and easily interpret predictions of ML systems to implement them in clinical settings. Black-box algorithms may have good prediction accuracy but their predictions are difficult to interpret and are not actionable, hence limiting their clinical application [62]. Fortunately, new methods have been devised to allow researchers to interpret black-box algorithms to ensure their predictions are sensible [63, 64, 65]."
197	Tong, D L and Boocock, D J and Coveney, C and Saif, J and Gomez, S G and Querol, S and Rees, R and Ball, G R	A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis: Stem cell and melanoma cancer studies	Clinical Proteomics	8	1		2011	UK	http://dx.doi.org/10.1186/1559-0275-8-14	article	Melanoma data set: 101 patients analyzed (yielding mass spectral data for 99 samples), Cord blood data set: 158 samples, 70 samples were categorised as containing a "High" number of stem cells and the remaining 88 samples with a "Low" number of stem cells	Case-control study	AUC, accuracy (Monte Carlo cross-validation + external validation set)	cross-validation + external cohort validation	"Raw spectral data from matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) with MS profiling techniques usually contains complex information not readily providing biological insight into disease. The association of identified features within raw data to a known peptide is extremely difficult. Data preprocessing to remove uncertainty characteristics in the data is normally required before performing any further analysis. This study proposes an alternative yet simple solution to preprocess raw MALDI-TOF-MS data for identification of candidate marker ions. Two in-house MALDI-TOF-MS data sets from two different sample sources (melanoma serum and cord blood plasma) are used in our study. [...] Our model identified 10 candidate marker ions for both data sets. These ion panels achieved over 90% classification accuracy on blind validation data. Receiver operating characteristics analysis was performed and the area under the curve for melanoma and cord blood classifiers was 0.991 and 0.986, respectively."	"Raw spectral data from matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) with MS profiling techniques usually contains complex information not readily providing biological insight into disease. The association of identified features within raw data to a known peptide is extremely difficult. Data preprocessing to remove uncertainty characteristics in the data is normally required before performing any further analysis. This study proposes an alternative yet simple solution to preprocess raw MALDI-TOF-MS data for identification of candidate marker ions. Two in-house MALDI-TOF-MS data sets from two different sample sources (melanoma serum and cord blood plasma) are used in our study. [...] Our model identified 10 candidate marker ions for both data sets. These ion panels achieved over 90% classification accuracy on blind validation data. Receiver operating characteristics analysis was performed and the area under the curve for melanoma and cord blood classifiers was 0.991 and 0.986, respectively."

198	Trakadis, Y J and Sardaar, S and Chen, A and Fulginiti, V and Krishnan, A	Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes	American Journal of Medical Genetics Part B-Neuropsychiatric Genetics	180	2	103-112	2019	Canada	http://dx.doi.org/10.1002/ajmg.b.32638	article	"This study applies ML to WES data from 2,545 individuals with SCZ and 2,545 unaffected individuals"	Case-control study	AUC, accuracy, sensitivity, specificity, precision, recall, F1-measure (training + test set)	training + test set	"Our hypothesis is that machine learning (ML) analysis of whole exome sequencing (WES) data can be used to identify individuals at high risk for schizophrenia (SCZ). This study applies ML to WES data from 2,545 individuals with SCZ and 2,545 unaffected individuals, accessed via the database of genotypes and phenotypes (dbGaP). [...] The supervised ML algorithm, gradient boosted trees with regularization (eXtreme Gradient Boosting implementation) was the best performing algorithm yielding promising results (accuracy: 85.7%, specificity: 86.6%, sensitivity: 84.9%, area under the receiver-operator characteristic curve: 0.95). The top 50 features (genes) of the algorithm were analyzed using bioinformatics resources for new insights about the pathophysiology of SCZ."	"Our hypothesis is that machine learning (ML) analysis of whole exome sequencing (WES) data can be used to identify individuals at high risk for schizophrenia (SCZ). This study applies ML to WES data from 2,545 individuals with SCZ and 2,545 unaffected individuals, accessed via the database of genotypes and phenotypes (dbGaP). [...] The supervised ML algorithm, gradient boosted trees with regularization (eXtreme Gradient Boosting implementation) was the best performing algorithm yielding promising results (accuracy: 85.7%, specificity: 86.6%, sensitivity: 84.9%, area under the receiver-operator characteristic curve: 0.95). The top 50 features (genes) of the algorithm were analyzed using bioinformatics resources for new insights about the pathophysiology of SCZ."
199	Troisi, J and Sarno, L and Martinelli, P and Di Carlo, C and Landolfi, A and Scala, G and Rinaldi, M and D'Alessandro, P and Ciccone, C and Guida, M	A metabolomics-based approach for non-invasive diagnosis of chromosomal anomalies	Metabolomics	13	11		2017		http://dx.doi.org/10.1007/s11306-017-1274-z	article	"Metabolomic profiles have been obtained on serum of 328 mothers (220 controls and 108 cases)"	Case-control study	AUC, accuracy, sensitivity, specificity, PPV, NPV, F-measure, G-mean (Leave k-out cross-validation, external test set)	cross-validation + external cohort validation	"[To] evaluate the diagnostic performance of a machine learning ensemble model based on the maternal serum metabolomic fingerprint of fetal aneuploidies during the second trimester [...] Eight machines learning and classification models were built and optimized. An ensemble model was built using a voting scheme. All samples were randomly divided into two sets. One was used as training set, the other one for diagnostic performance assessment. Ensemble machine learning model correctly classified all cases and controls. The accuracy was the same for trisomy 21 and 18; also, the other [chromosomal anomalies] CA were correctly detected."	"[To] evaluate the diagnostic performance of a machine learning ensemble model based on the maternal serum metabolomic fingerprint of fetal aneuploidies during the second trimester [...] Eight machines learning and classification models were built and optimized. An ensemble model was built using a voting scheme. All samples were randomly divided into two sets. One was used as training set, the other one for diagnostic performance assessment. Ensemble machine learning model correctly classified all cases and controls. The accuracy was the same for trisomy 21 and 18; also, the other [chromosomal anomalies] CA were correctly detected."
200	Troisi, J and Sarno, L and Richards, S M and Symes, S J and Adair, D C and Scala, G and Taylor, R S and McCowan, L M and Fasano, A and Martinelli, P and Guida, M	Noninvasive screening of fetal anomalies: The serum metabolomic way	Birth Defects Research	110	9	757-757	2018		http://dx.doi.org/10.1002/bdr2.1355	meeting abstract	Metabolomic profiles were obtained from serum of 654 mothers (320 controls, with a normal fetus and 334 cases with a malformed fetus)	Case-control study	accuracy, sensitivity, specificity (training and validation set)	cross-validation + test set	"Fetal malformations (FM) are structural or functional anomalies that occur during intrauterine development. [...] We performed a characterization of maternal serum to build a metabolomic fingerprint resulting from FM. We then tested its' accuracy with an independent population from the early pregnancy biobank from the New Zealand SCOPE Cohort. [...] The ensemble model correctly classified all cases and controls. Performance evaluation was assessed using 1,935 second trimester maternal serum samples from the SCOPE biobank as an independent test set. Blind analysis using the described test showed a global accuracy for FM identification of 99.4±0.1% (sensitivity=78±6% [32/41 FM correctly identified]; specificity=100±0% [1892/1894 controls correctly identified]). Maternal serum metabolomics is therefore a promising tool for the accurate and sensitive, noninvasive screening of FMs."	"Fetal malformations (FM) are structural or functional anomalies that occur during intrauterine development. [...] We performed a characterization of maternal serum to build a metabolomic fingerprint resulting from FM. We then tested its' accuracy with an independent population from the early pregnancy biobank from the New Zealand SCOPE Cohort. [...] The ensemble model correctly classified all cases and controls. Performance evaluation was assessed using 1,935 second trimester maternal serum samples from the SCOPE biobank as an independent test set. Blind analysis using the described test showed a global accuracy for FM identification of 99.4±0.1% (sensitivity=78±6% [32/41 FM correctly identified]; specificity=100±0% [1892/1894 controls correctly identified]). Maternal serum metabolomics is therefore a promising tool for the accurate and sensitive, noninvasive screening of FMs."
201	Tylee, D S and Hess, J L and Quinn, T P and Barve, R and Huang, H and Zhang-James, Y and Chang, J and Stamova, B S and Sharp, F R and Hertz-Picciotto, I and Faraone, S V and Kong, S W and Glatt, S J	Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis	Am J Med Genet B Neuropsychiatr Genet	174	3	181-201	2017	Norway	http://dx.doi.org/10.1002/ajmg.b.32511	article	"Raw microarray data and clinical meta-data were obtained from seven studies, totaling 626 affected and 447 comparison subjects"	Case-control study	AUC, sensitivity, specificity (nested 10-fold CV within 5-times bootstrapped (boot67) samples + test set)	cross-validation + test set	"Blood-based microarray studies comparing individuals affected with autism spectrum disorder (ASD) and typically developing individuals help characterize differences in circulating immune cell functions and offer potential biomarker signal. We sought to combine the subject-level data from previously published studies by mega-analysis to increase the statistical power. We identified studies that compared ex vivo blood or lymphocytes from ASD-affected individuals and unrelated comparison subjects using Affymetrix or Illumina array platforms. [...] We also demonstrated that machine-learning classifiers using blood transcriptome data perform with moderate accuracy when data are combined across studies. Comparing our results with those from blood-based studies of protein biomarkers (e.g., cytokines and trophic factors), we propose that ASD may feature decoupling between certain circulating signaling proteins (higher in ASD samples) and the transcriptional cascades which they typically elicit within circulating immune cells (lower in ASD samples)."	"Blood-based microarray studies comparing individuals affected with autism spectrum disorder (ASD) and typically developing individuals help characterize differences in circulating immune cell functions and offer potential biomarker signal. We sought to combine the subject-level data from previously published studies by mega-analysis to increase the statistical power. We identified studies that compared ex vivo blood or lymphocytes from ASD-affected individuals and unrelated comparison subjects using Affymetrix or Illumina array platforms. [...] We also demonstrated that machine-learning classifiers using blood transcriptome data perform with moderate accuracy when data are combined across studies. Comparing our results with those from blood-based studies of protein biomarkers (e.g., cytokines and trophic factors), we propose that ASD may feature decoupling between certain circulating signaling proteins (higher in ASD samples) and the transcriptional cascades which they typically elicit within circulating immune cells (lower in ASD samples)."
202	Urda, D and Aragon, F and Bautista, R and Franco, L and Veredas, F J and Claros, M G and Jerez, J M	BLASSO: integration of biological knowledge into a regularized linear model	BMC Syst Biol	12		94-94	2018	Spain	http://dx.doi.org/10.1186/s12918-018-0612-8	article	"Out of the 1212 samples, 1013 corresponds to controls (or alive patients) and 199 to cases (or patients who died from the disease)"	Case-control study	AUC (100 repetitions of 10-fold nested CV, witih 9-fold CV nested for hyper-parameter tuning)	cross-validation	"[...] this paper proposes BLASSO, a simple and highly interpretable linear model with l1-regularization that incorporates prior biological knowledge to the prediction of breast cancer outcomes. Two different approaches to integrate biological knowledge in BLASSO, Gene-specific and Gene-disease, are proposed to test their predictive performance and biomarker stability on a public RNA-Seq gene expression dataset for breast cancer. [...] BLASSO has been compared with a baseline LASSO model. Using 10-fold cross-validation with 100 repetitions for models' assessment, average AUC values of 0.7 and 0.69 were obtained for the Gene-specific and the Gene-disease approaches, respectively. These efficacy rates outperform the average AUC of 0.65 obtained with the LASSO. With respect to the stability of the genetic signatures found, BLASSO outperformed the baseline model in terms of the robustness index (RI). The Gene-specific approach gave RI of 0.15±0.03, compared to RI of 0.09±0.03 given by LASSO, thus being 66% times more robust. The functional analysis performed to the genetic signature obtained with the Gene-disease approach showed a significant presence of genes related with cancer, as well as one gene (IFNK) and one pseudogene (PCNAP1) which a priori had not been described to be related with cancer."	"[...] this paper proposes BLASSO, a simple and highly interpretable linear model with l1-regularization that incorporates prior biological knowledge to the prediction of breast cancer outcomes. Two different approaches to integrate biological knowledge in BLASSO, Gene-specific and Gene-disease, are proposed to test their predictive performance and biomarker stability on a public RNA-Seq gene expression dataset for breast cancer. [...] BLASSO has been compared with a baseline LASSO model. Using 10-fold cross-validation with 100 repetitions for models' assessment, average AUC values of 0.7 and 0.69 were obtained for the Gene-specific and the Gene-disease approaches, respectively. These efficacy rates outperform the average AUC of 0.65 obtained with the LASSO. With respect to the stability of the genetic signatures found, BLASSO outperformed the baseline model in terms of the robustness index (RI). The Gene-specific approach gave RI of 0.15±0.03, compared to RI of 0.09±0.03 given by LASSO, thus being 66% times more robust. The functional analysis performed to the genetic signature obtained with the Gene-disease approach showed a significant presence of genes related with cancer, as well as one gene (IFNK) and one pseudogene (PCNAP1) which a priori had not been described to be related with cancer."
203	van Vliet, M H and Klijn, C N and Wessels, L F and Reinders, M J	Module-based outcome prediction using breast cancer compendia	PLoS One	2	10	e1047-e1047	2007	Netherlands	http://dx.doi.org/10.1371/journal.pone.0001047	article	"This compendium contains data from various cancer types and has a total of 1973 arrays" (more than 50 samples per group for combined datasets)	Case-control study	AUC (double-loop cross-validation + external validation)	cross-validation + external cohort validation	"The availability of large collections of microarray datasets (compendia), or knowledge about grouping of genes into pathways (gene sets), is typically not exploited when training predictors of disease outcome. These can be useful since a compendium increases the number of samples, while gene sets reduce the size of the feature space. This should be favorable from a machine learning perspective and result in more robust predictors. [...] We extracted modules of regulated genes from gene sets, and compendia. Through supervised analysis, we constructed predictors which employ modules predictive of breast cancer outcome. [...] We show that modules derived from single breast cancer datasets achieve better performance on the validation data compared to gene-based predictors. We also show that there is a trend in compendium specificity and predictive performance: modules derived from a single breast cancer dataset, and a breast cancer specific compendium perform better compared to those derived from a human cancer compendium. Additionally, the module-based predictor provides a much richer insight into the underlying biology. Frequently selected gene sets are associated with processes such as cell cycle, E2F regulation, DNA damage response, proteasome and glycolysis."	"The availability of large collections of microarray datasets (compendia), or knowledge about grouping of genes into pathways (gene sets), is typically not exploited when training predictors of disease outcome. These can be useful since a compendium increases the number of samples, while gene sets reduce the size of the feature space. This should be favorable from a machine learning perspective and result in more robust predictors. [...] We extracted modules of regulated genes from gene sets, and compendia. Through supervised analysis, we constructed predictors which employ modules predictive of breast cancer outcome. [...] We show that modules derived from single breast cancer datasets achieve better performance on the validation data compared to gene-based predictors. We also show that there is a trend in compendium specificity and predictive performance: modules derived from a single breast cancer dataset, and a breast cancer specific compendium perform better compared to those derived from a human cancer compendium. Additionally, the module-based predictor provides a much richer insight into the underlying biology. Frequently selected gene sets are associated with processes such as cell cycle, E2F regulation, DNA damage response, proteasome and glycolysis."
204	Wan, N and Weinberg, D and Liu, T Y and Niehaus, K and Ariazi, E A and Delubac, D and Kannan, A and White, B and Bailey, M and Bertin, M and Boley, N and Bowen, D and Cregg, J and Drake, A M and Ennis, R and Fransen, S and Gafni, E and Hansen, L and Liu, Y and Otte, G L and Pecson, J and Rice, B and Sanderson, G E and Sharma, A and St John, J and Tang, C and Tzou, A and Young, L and Putcha, G and Haque, I S	Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DANN	BMC Cancer	19	1	832-832	2019	USA	http://dx.doi.org/10.1186/s12885-019-6003-8	article	N = 546 colorectal cancer and 271 non-cancer controls	Case-control study	AUC, sensitivity, specificity (5-fold CV + confounder-based cross-validations)	cross-validation	"[...] early-stage detection of cancer using tumor-derived cfDNA has proven challenging because of the small proportion of cfDNA derived from tumor tissue in early-stage disease. A machine learning approach to discover signatures in cfDNA, potentially reflective of both tumor and non-tumor contributions, may represent a promising direction for the early detection of cancer. [...] Whole-genome sequencing was performed on cfDNA extracted from plasma samples (N = 546 colorectal cancer and 271 non-cancer controls). Reads aligning to protein-coding gene bodies were extracted, and read counts were normalized. cfDNA tumor fraction was estimated using IchorCNA. Machine learning models were trained using k-fold cross-validation and confounder-based cross-validations to assess generalization performance. [...] In a colorectal cancer cohort heavily weighted towards early-stage cancer (80% stage I/II), we achieved a mean AUC of 0.92 (95% CI 0.91–0.93) with a mean sensitivity of 85% (95% CI 83–86%) at 85% specificity. Sensitivity generally increased with tumor stage and increasing tumor fraction. Stratification by age, sequencing batch, and institution demonstrated the impact of these confounders and provided a more accurate assessment of generalization performance."	"[...] early-stage detection of cancer using tumor-derived cfDNA has proven challenging because of the small proportion of cfDNA derived from tumor tissue in early-stage disease. A machine learning approach to discover signatures in cfDNA, potentially reflective of both tumor and non-tumor contributions, may represent a promising direction for the early detection of cancer. [...] Whole-genome sequencing was performed on cfDNA extracted from plasma samples (N = 546 colorectal cancer and 271 non-cancer controls). Reads aligning to protein-coding gene bodies were extracted, and read counts were normalized. cfDNA tumor fraction was estimated using IchorCNA. Machine learning models were trained using k-fold cross-validation and confounder-based cross-validations to assess generalization performance. [...] In a colorectal cancer cohort heavily weighted towards early-stage cancer (80% stage I/II), we achieved a mean AUC of 0.92 (95% CI 0.91–0.93) with a mean sensitivity of 85% (95% CI 83–86%) at 85% specificity. Sensitivity generally increased with tumor stage and increasing tumor fraction. Stratification by age, sequencing batch, and institution demonstrated the impact of these confounders and provided a more accurate assessment of generalization performance."
205	Wang, J and Yan, D and Zhao, A and Hou, X and Zheng, X and Chen, P and Bao, Y and Jia, W and Hu, C and Zhang, Z L and Jia, W	Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods	Osteoporos Int	30	7	1491-1499	2019	China	http://dx.doi.org/10.1007/s00198-019-04892-0	article	"Our study recruited 320 participants, including 138 males and 182 postmenopausal females"	Case-control study	AUC, accuracy, sensitivity, specificity ("The data sets were randomly split into the recommended ratio of 70% for model training and the other 30% for validation")	training + test set	"Our study aimed to evaluate the association of metabolites with bone, trying to find new metabolic markers that are distinguishing for low bone mineral density (BMD). [...] We compared metabolites in three groups with different BMD levels in males and postmenopausal females separately and further filtering these metabolites via random forestbased feature selection, a commonly applied machine learning algorithm which could select the features with the greatest impact on osteoporosis, then metabolites with the highest importance (≥ 5%) (5 inmales and 9 in postmenopausal females) were selected to construct better models for osteoporosis classification. After adding these selected metabolites to the model, the area under the curve (AUC) of receiver operating characteristic (ROC) curves increased significantly"	"Our study aimed to evaluate the association of metabolites with bone, trying to find new metabolic markers that are distinguishing for low bone mineral density (BMD). [...] We compared metabolites in three groups with different BMD levels in males and postmenopausal females separately and further filtering these metabolites via random forestbased feature selection, a commonly applied machine learning algorithm which could select the features with the greatest impact on osteoporosis, then metabolites with the highest importance (≥ 5%) (5 inmales and 9 in postmenopausal females) were selected to construct better models for osteoporosis classification. After adding these selected metabolites to the model, the area under the curve (AUC) of receiver operating characteristic (ROC) curves increased significantly"
206	Wang, J and Zuo, Y and Man, Y G and Avital, I and Stojadinovic, A and Liu, M and Yang, X and Varghese, R S and Tadesse, M G and Ressom, H W	Pathway and network approaches for identification of cancer signature markers from omics data	Journal of Cancer	6	1	54-65	2015	USA	http://dx.doi.org/10.7150/jca.10631	article	review (not applicable)	review			"The advancement of high throughput omic technologies during the past few years has made it possible to perform many complex assays in a much shorter time than the traditional approaches. The rapid accumulation and wide availability of omic data generated by these technologies offer great opportunities to unravel disease mechanisms, but also presents significant challenges to extract knowledge from such massive data and to evaluate the findings. To address these challenges, a number of pathway and network based approaches have been introduced. This review article evaluates these methods and discusses their application in cancer biomarker discovery using hepatocellular carcinoma (HCC) as an example."	The review categorizes computational biomarker discovery methods into statistical methods, machine learning approaches, graph/network theory based methods, Bayesian methods and derivatives, text mining approaches, and integrative methods. Successful applications of these approaches for hepatocellular carcinoma are described
207	Wang, L and Liu, Z P	Detecting diagnostic biomarkers of Alzheimer's disease by integrating gene expression data in six brain regions	Frontiers in Genetics	10			2019	China	http://dx.doi.org/10.3389/fgene.2019.00157	article	gene expression profiles of 161 samples in six brain regions	Case-control study	AUC, sensitivity, specificity (LOOCV)	cross-validation	"In this paper, we propose a machine-learning-based method of identifying potential diagnostic biomarkers of [Alzheimer's disease] AD based on gene coexpression network by integrating gene expression profiles in six brain regions. After building an integrated gene coexpression network of multiple brain regions, we decompose the differential network into some subnetwork modules. The module candidates from these coexpressed gene communities are then identified by screening their discriminative powers in control from disease samples. The potential biomarkers are then validated by multiple cross-validations and functional enrichment analyses."	"In this paper, we propose a machine-learning-based method of identifying potential diagnostic biomarkers of [Alzheimer's disease] AD based on gene coexpression network by integrating gene expression profiles in six brain regions. After building an integrated gene coexpression network of multiple brain regions, we decompose the differential network into some subnetwork modules. The module candidates from these coexpressed gene communities are then identified by screening their discriminative powers in control from disease samples. The potential biomarkers are then validated by multiple cross-validations and functional enrichment analyses."
208	Wang, M and Yu, G and Ressom, H W	Integrative Analysis of Proteomic, Glycomic, and Metabolomic Data for Biomarker Discovery	IEEE J Biomed Health Inform	20	5	1225-1231	2016	United States	http://dx.doi.org/10.1109/jbhi.2016.2574201	article	The TU cohort consists of a total of 89 subjects (40 HCC cases and 49 patients with liver cirrhosis), and the GU cohort comprises of 116 subjects (57 HCC cases and 59 patients with liver cirrhosis) (more than 50 samples per group for the 2nd cohort)	Case-control study	AUC, accuracy, sensitivity, specificity (10-fold cross-validation + test set)	cross-validation + test set	"In this study, we investigate integrative analysis of proteins, N-glycans, and metabolites to take advantage of complementary information to improve the ability to distinguish cancer cases from controls. Specifically, SVM-RFE algorithm is utilized to select a panel of proteins, N-glycans, and metabolites based on LC-MS and GC-MS data previously acquired by analysis of blood samples from two cohorts in a liver cancer study. Improved performances are observed by integrative analysis compared to separate proteomic, glycomic, and metabolomic studies in distinguishing liver cancer cases from patients with liver cirrhosis."	"In this study, we investigate integrative analysis of proteins, N-glycans, and metabolites to take advantage of complementary information to improve the ability to distinguish cancer cases from controls. Specifically, SVM-RFE algorithm is utilized to select a panel of proteins, N-glycans, and metabolites based on LC-MS and GC-MS data previously acquired by analysis of blood samples from two cohorts in a liver cancer study. Improved performances are observed by integrative analysis compared to separate proteomic, glycomic, and metabolomic studies in distinguishing liver cancer cases from patients with liver cirrhosis."

209	Wang, N and Cao, Y and Song, W and He, K and Li, T and Wang, J and Xu, B and Si, H Y and Hu, C J and Li, A L	Serum peptide pattern that differentially diagnoses hepatitis B virus-related hepatocellular carcinoma from liver cirrhosis	J Gastroenterol Hepatol	29	7	1544-1550	2014	China	http://dx.doi.org/10.1111/jgh.12545	article	80 HCC and 67 LC patients	Case-control study	AUC, accuracy, sensitivity, specificity (10-fold cross-validation + test set)	cross-validation + test set	"Although alpha-fetoprotein (AFP) is a useful serologic marker of hepatocellular carcinoma (HCC), it is not sufficiently sensitive to differentiate HCC and liver cirrhosis (LC) caused by hepatitis B virus (HBV) infection. [...] With a highly optimized peptide extraction and matrix-assisted laser desorption/ionization time of flight/time of flight mass spectrometric approach, we investigated serum peptide profiles of 80 HCC and 67 LC patients. Three supervised machine learning methods were employed to construct classifiers. [...] We proposed a novel method for distinguishing HCC from cirrhosis, based on a multilayer perceptron (MLP) method.We obtained a sensitivity of 90.0%, specificity of 79.4%, and overall accuracy of 85.1% on an independent test set. The combination of the MLP model and serum AFP level outperformed serum AFP marker alone in distinguishing HCC patients from LC patients."	"Although alpha-fetoprotein (AFP) is a useful serologic marker of hepatocellular carcinoma (HCC), it is not sufficiently sensitive to differentiate HCC and liver cirrhosis (LC) caused by hepatitis B virus (HBV) infection. [...] With a highly optimized peptide extraction and matrix-assisted laser desorption/ionization time of flight/time of flight mass spectrometric approach, we investigated serum peptide profiles of 80 HCC and 67 LC patients. Three supervised machine learning methods were employed to construct classifiers. [...| We proposed a novel method for distinguishing HCC from cirrhosis, based on a multilayer perceptron (MLP) method.We obtained a sensitivity of 90.0%, specificity of 79.4%, and overall accuracy of 85.1% on an independent test set. The combination of the MLP model and serum AFP level outperformed serum AFP marker alone in distinguishing HCC patients from LC patients."
210	Wang, S J and Li, M C	Impacts of Predictive Genomic Classifier Performance on Subpopulation-Specific Treatment Effects Assessment	Statistics in Biosciences	8	1	129-158	2016		http://dx.doi.org/10.1007/s12561-013-9092-y	article	"frozen tissue was collected only from 169 patients, of which only 166 contained more than 20 % tumor cellularity, and gene expression profiling was completed in 133 patients using the Affymetrix 133A microarray [8]. Among them, 62 NSCLC patients received OBS alone and 71 NSCLC patients received ACT."	Case-control study	accuracy, sensitivity, specificity, PPV, NPV, permutation p-values (cross-validation + external validation)	cross-validation + external cohort validation	"We investigate the classification performance characteristics of a binary genomic composite biomarker (expected to be predictive of treatment effects) including sensitivity, specificity, accuracy, positive predictive value and negative predictive value as a function of true sensitive prevalence. In doing so, we report the finding based on three representative tuning parameter sets with varying degree of rigor in their choices of the parameters ranging from highly rigorous, moderately rigorous to mildly rigorous. We articulate the rationales on the choices of tuning parameter sets. We also study the impacts of misclassification of genomic biomarker classifiers on their assessment of treatment effects in the positive and negative patient subpopulations, and all-comer patients."	"We investigate the classification performance characteristics of a binary genomic composite biomarker (expected to be predictive of treatment effects) including sensitivity, specificity, accuracy, positive predictive value and negative predictive value as a function of true sensitive prevalence. In doing so, we report the finding based on three representative tuning parameter sets with varying degree of rigor in their choices of the parameters ranging from highly rigorous, moderately rigorous to mildly rigorous. We articulate the rationales on the choices of tuning parameter sets. We also study the impacts of misclassification of genomic biomarker classifiers on their assessment of treatment effects in the positive and negative patient subpopulations, and all-comer patients."
211	Wei, Z and Wang, K and Qu, H Q and Zhang, H and Bradfield, J and Kim, C and Frackleton, E and Hou, C and Glessner, J T and Chiavacci, R and Stanley, C and Monos, D and Grant, S F and Polychronakos, C and Hakonarson, H	From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes	PLoS Genet	5	10	e1000678-e1000678	2009	United States of America	http://dx.doi.org/10.1371/journal.pgen.1000678	article	"We accessed the 500K Affymetrix chip genotype data from WTCCC on ~1,500 samples from the 1958 British Birth Cohort, ~1,500 samples from the UK Blood Service Control Group, as well as ~2,000 samples each from the following disease collections: type 1 diabetes (T1D), type 2 diabetes (T2D), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), bipolar disorder (BD), hypertension (HT), coronary artery disease (CAD)"	Case-control study	AUC, accuracy, sensitivity, specificity (5-fold cross-validation + test set)	cross-validation + test set	"Genome-wide association studies (GWAS) have been fruitful in identifying disease susceptibility loci for common and complex diseases. A remaining question is whether we can quantify individual disease risk based on genotype data, in order to facilitate personalized prevention and treatment for complex diseases. Previous studies have typically failed to achieve satisfactory performance, primarily due to the use of only a limited number of confirmed susceptibility loci. Here we propose that sophisticated machine-learning approaches with a large ensemble of markers may improve the performance of disease risk assessment. We applied a Support Vector Machine (SVM) algorithm on a GWAS dataset generated on the Affymetrix genotyping platform for type 1 diabetes (T1D) and optimized a risk assessment model with hundreds of markers. We subsequently tested this model on an independent Illumina-genotyped dataset with imputed genotypes (1,008 cases and 1,000 controls), as well as a separate Affymetrix-genotyped dataset (1,529 cases and 1,458 controls), resulting in area under ROC curve (AUC) of ~0.84 in both datasets."	"Genome-wide association studies (GWAS) have been fruitful in identifying disease susceptibility loci for common and complex diseases. A remaining question is whether we can quantify individual disease risk based on genotype data, in order to facilitate personalized prevention and treatment for complex diseases. Previous studies have typically failed to achieve satisfactory performance, primarily due to the use of only a limited number of confirmed susceptibility loci. Here we propose that sophisticated machine-learning approaches with a large ensemble of markers may improve the performance of disease risk assessment. We applied a Support Vector Machine (SVM) algorithm on a GWAS dataset generated on the Affymetrix genotyping platform for type 1 diabetes (T1D) and optimized a risk assessment model with hundreds of markers. We subsequently tested this model on an independent Illumina-genotyped dataset with imputed genotypes (1,008 cases and 1,000 controls), as well as a separate Affymetrix-genotyped dataset (1,529 cases and 1,458 controls), resulting in area under ROC curve (AUC) of ~0.84 in both datasets."
212	White, B S and Khan, S A and Ammad-Ud-Din, M and Potdar, S and Mason, M J and Tognon, C E and Druker, B J and Heckman, C A and Kallioniemi, O P and Kurtz, S E and Porkka, K and Tyner, J W and Aittokallio, T and Wennerberg, K and Guinney, J	Gene expression predicts ex vivo drug sensitivity in acute myeloid leukemia	Cancer Research	78	13		2018		http://dx.doi.org/10.1158/1538-7445.AM2018-3883	meeting abstract	"We harmonized two large-scale AML ex vivo studies screened for drug response and profiled transcriptomically—OHSU (303 AML patient samples and 160 drugs) and FIMM (48 AML samples and 480 drugs)"	Case-control study	correlation, p-value (10-fold CV)	cross-validation	"Ex vivo drug sensitivity studies of samples derived from acute myeloid leukemia (AML) patients have been shown to be predictive of in vivo response. These findings are based on a limited number of well-characterized agents for which in vivo patient response data and ex vivo drug sensitivity data—on that same patient—are available. To show the feasibility of scaling such ex vivo studies to large drug screens, we characterized the reproducibility of expression-based models of drug response across two independent data sets. [...] For each of the 94 drugs in common between the two data sets, we trained a Ridge regression model on the OHSU data set, used the model to predict response in the FIMM data set, and calculated the Pearson correlation between the predicted and observed FIMM responses. 41 of the 94 drug models had a positive and statistically significant correlation [false discovery rate (FDR) < 20%; mean ρ = 0.43; 95% CI = 0.29 – 0.77]. Drugs corresponding to the top decile of these significant models (mean ρ = 0.54; 95% CI = 0.48 – 0.77) clustered into four primary classes: MEK inhibitors (PD184352, Selumetinib, and Trametinib), EGFR/VEGFR inhibitors (Cabozantinib, Erlotinib, Foretinib, and Sorafenib), and singletons Venetoclax and Sirolimus"	"Ex vivo drug sensitivity studies of samples derived from acute myeloid leukemia (AML) patients have been shown to be predictive of in vivo response. These findings are based on a limited number of well-characterized agents for which in vivo patient response data and ex vivo drug sensitivity data—on that same patient—are available. To show the feasibility of scaling such ex vivo studies to large drug screens, we characterized the reproducibility of expression-based models of drug response across two independent data sets. [...] For each of the 94 drugs in common between the two data sets, we trained a Ridge regression model on the OHSU data set, used the model to predict response in the FIMM data set, and calculated the Pearson correlation between the predicted and observed FIMM responses. 41 of the 94 drug models had a positive and statistically significant correlation [false discovery rate (FDR) < 20%; mean ρ = 0.43; 95% CI = 0.29 – 0.77]. Drugs corresponding to the top decile of these significant models (mean ρ = 0.54; 95% CI = 0.48 – 0.77) clustered into four primary classes: MEK inhibitors (PD184352, Selumetinib, and Trametinib), EGFR/VEGFR inhibitors (Cabozantinib, Erlotinib, Foretinib, and Sorafenib), and singletons Venetoclax and Sirolimus"
213	Wu, H and Cai, L and Li, D and Wang, X and Zhao, S and Zou, F and Zhou, K	Metagenomics Biomarkers Selected for Prediction of Three Different Diseases in Chinese Population	Biomed Res Int	2018		2936257-2936257	2018	China	http://dx.doi.org/10.1155/2018/2936257	article	"microbiome of 806 Chinese individuals (383 controls, 170 with type 2 diabetes, 130 with rheumatoid arthritis, and 123 with liver cirrhosis)"	Case-control study	AUC, F1-score (5-fold CV)	cross-validation	"The dysbiosis of human microbiome has been proven to be associated with the development of many human diseases. Metagenome sequencing emerges as a powerful tool to investigate the effects of microbiome on diseases. [...] Here, we developed a pipeline to address the challenging characterization of multilabel samples. In this study, a total of 300 biomarkers were selected from the microbiome of 806 Chinese individuals (383 controls, 170 with type 2 diabetes, 130 with rheumatoid arthritis, and 123 with liver cirrhosis), and then logistic regression prediction algorithm was applied to those markers as the model intrinsic features. The estimated model produced an score of 0.9142, which was better than other popular classification methods, and an average receiver operating characteristic (ROC) of 0.9475 showed a significant correlation between these selected biomarkers from microbiome and corresponding phenotypes."	"The dysbiosis of human microbiome has been proven to be associated with the development of many human diseases. Metagenome sequencing emerges as a powerful tool to investigate the effects of microbiome on diseases. [...] Here, we developed a pipeline to address the challenging characterization of multilabel samples. In this study, a total of 300 biomarkers were selected from the microbiome of 806 Chinese individuals (383 controls, 170 with type 2 diabetes, 130 with rheumatoid arthritis, and 123 with liver cirrhosis), and then logistic regression prediction algorithm was applied to those markers as the model intrinsic features. The estimated model produced an score of 0.9142, which was better than other popular classification methods, and an average receiver operating characteristic (ROC) of 0.9475 showed a significant correlation between these selected biomarkers from microbiome and corresponding phenotypes."
214	Xie, G and Dong, C and Kong, Y and Zhong, J F and Li, M and Wang, K	Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features	Genes	10	3		2019	USA	http://dx.doi.org/10.3390/genes10030240	article	the used dataset cover more than 50 samples per group	Case-control study			"To harness the rich information in multi-omics data, we developed GDP (Group lass regularized Deep learning for cancer Prognosis), a computational tool for survival prediction using both clinical and multi-omics data. GDP integrated a deep learning framework and Cox proportional hazard model (CPH) together, and applied group lasso regularization to incorporate gene-level group prior knowledge into the model training process. We evaluated its performance in both simulated and real data from The Cancer Genome Atlas (TCGA) project."	"To harness the rich information in multi-omics data, we developed GDP (Group lass regularized Deep learning for cancer Prognosis), a computational tool for survival prediction using both clinical and multi-omics data. GDP integrated a deep learning framework and Cox proportional hazard model (CPH) together, and applied group lasso regularization to incorporate gene-level group prior knowledge into the model training process. We evaluated its performance in both simulated and real data from The Cancer Genome Atlas (TCGA) project."
215	Yang, J S and Zhu, Z X and He, S and Ji, Z and Ieee	Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification	Proceedings of the 2013 Ieee Symposium on Computational Intelligence in Bioinformatics and Computational Biology			246-251	2013	USA	https://doi.org/10.1109/CIBCB.2013.6595417	article	the five selected datasets include multiple datasets with more than 50 samples per group	Case-control study	accuracy (10 runs of 10-fold CV)	cross-validation	"This paper studies a minimal-redundancy-maximal-relevance (MRMR) feature selection for omics data classification using three different relevance evaluation measures including mutual information (MI), correlation coefficient (CC), and maximal information coefficient (MIC). A linear forward search method is used to search the optimal feature subset. The experimental results on five real-world omics datasets indicate that MRMR feature selection with CC is more robust to obtain better (or competitive) classification accuracy than the other two measures."	"This paper studies a minimal-redundancy-maximal-relevance (MRMR) feature selection for omics data classification using three different relevance evaluation measures including mutual information (MI), correlation coefficient (CC), and maximal information coefficient (MIC). A linear forward search method is used to search the optimal feature subset. The experimental results on five real-world omics datasets indicate that MRMR feature selection with CC is more robust to obtain better (or competitive) classification accuracy than the other two measures."
216	Yang, S and Naiman, D Q	Multiclass cancer classification based on gene expression comparison	Stat Appl Genet Mol Biol	13	4	477-496	2014	United States	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4775275/	article	one of the considered datasets has more than 50 samples per group ("MILE is a two-stage study where a retrospective stage I generated expression profiles for 2,143 patients and was designed for biomarker discovery. A prospective stage II produced an independent cohort of 1,152 patients and was used for validation")	Case-control study	accuracy (LOOCV + test set)	cross-validation + test set	"[...] multiclass classification problems pose new methodological and computational challenges for developing novel and effective statistical approaches. In this paper, we introduce a new approach for classifying multiple disease states associated with cancer based on gene expression profiles. Our method focuses on detecting small sets of genes in which the relative comparison of their expression values leads to class discrimination. For an m-class problem, the classification rule typically depends on a small number of m-gene sets, which provide transparent decision boundaries and allow for potential biological interpretations. We first test our approach on seven common gene expression datasets and compare it with popular classification methods including support vector machines and random forests. We then consider an extremely large cohort of leukemia cancer to further assess its effectiveness. In both experiments, our method yields comparable or even better results to benchmark classifiers. In addition, we demonstrate that our approach can integrate pathway analysis of gene expression to provide accurate and biological meaningful classification."	"[...] multiclass classification problems pose new methodological and computational challenges for developing novel and effective statistical approaches. In this paper, we introduce a new approach for classifying multiple disease states associated with cancer based on gene expression profiles. Our method focuses on detecting small sets of genes in which the relative comparison of their expression values leads to class discrimination. For an m-class problem, the classification rule typically depends on a small number of m-gene sets, which provide transparent decision boundaries and allow for potential biological interpretations. We first test our approach on seven common gene expression datasets and compare it with popular classification methods including support vector machines and random forests. We then consider an extremely large cohort of leukemia cancer to further assess its effectiveness. In both experiments, our method yields comparable or even better results to benchmark classifiers. In addition, we demonstrate that our approach can integrate pathway analysis of gene expression to provide accurate and biological meaningful classification."
217	Yang, Y and Huang, N and Hao, L and Kong, W	A clustering-based approach for efficient identification of microRNA combinatorial biomarkers	BMC Genomics	18		210-210	2017	China	http://dx.doi.org/10.1186/s12864-017-3498-8	article	" There are a total of 120 samples collected from 64 breast cancer patients, including 56 pairs of matched tumor and adjacent peri-tumoral breast tissues, and 8 unmatched tissues in GSE 40525. And in GSE22220, there are 210 samples from 219 breast cancer patients, including 84 estrogen receptor (ER)-negative tissues, and 135 ER-positive tissues."	Case-control study	accuracy, sensitivity, specificity (5-fold cross-validation)	cross-validation	"This study aims to select combinatorial miRNA biomarkers, which have higher sensitivity and specificity than single-gene biomarkers. In order to avoid exhaustive search and redundant information, miRNAs are firstly clustered, then the combinations of representative cluster members are assessed as potential biomarkers. [...] Our experimental results demonstrate that the clustering-based method can identify microRNA combinatorial biomarkers with high accuracy and efficiency"	"This study aims to select combinatorial miRNA biomarkers, which have higher sensitivity and specificity than single-gene biomarkers. In order to avoid exhaustive search and redundant information, miRNAs are firstly clustered, then the combinations of representative cluster members are assessed as potential biomarkers. [...] Our experimental results demonstrate that the clustering-based method can identify microRNA combinatorial biomarkers with high accuracy and efficiency"

218	Yu, H and Samuels, D C and Zhao, Y Y and Guo, Y	Architectures and accuracy of artificial neural network for disease classification from omics data	Bmc Genomics	20		12-12	2019	China	http://dx.doi.org/10.1186/s12864-019-5546-z	article	37 omics datasets, including datasets with more than 50 samples per group	Case-control study	Accuracy, Kohen's Kappa (nested cross-validation)	cross-validation	"Deep learning has made tremendous successes in numerous artificial intelligence applications and is unsurprisingly penetrating into various biomedical domains. High-throughput omics data in the form of molecular profile matrices, such as transcriptomes and metabolomes, have long existed as a valuable resource for facilitating diagnosis of patient statuses/stages. It is timely imperative to compare deep learning neural networks against classical machine learning methods in the setting of matrix-formed omics data in terms of classification accuracy and robustness. [...] Using 37 high throughput omics datasets, covering transcriptomes and metabolomes, we evaluated the classification power of deep learning compared to traditional machine learning methods. Representative deep learning methods, Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN), were deployed and explored in seeking optimal architectures for the best classification performance. Together with five classical supervised classification methods (Linear Discriminant Analysis, Multinomial Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine), MLP and CNN were comparatively tested on the 37 datasets to predict disease stages or to discriminate diseased samples from normal samples. MLPs achieved the highest overall accuracy among all methods tested. More thorough analyses revealed that single hidden layer MLPs with ample hidden units outperformed deeper MLPs. Furthermore, MLP was one of the most robust methods against imbalanced class composition and inaccurate class labels. [...] Our results concluded that shallow MLPs (of one or two hidden layers) with ample hidden neurons are sufficient to achieve superior and robust classification performance in exploiting numerical matrix-formed omics data for diagnosis purpose."	"Deep learning has made tremendous successes in numerous artificial intelligence applications and is unsurprisingly penetrating into various biomedical domains. High-throughput omics data in the form of molecular profile matrices, such as transcriptomes and metabolomes, have long existed as a valuable resource for facilitating diagnosis of patient statuses/stages. It is timely imperative to compare deep learning neural networks against classical machine learning methods in the setting of matrix-formed omics data in terms of classification accuracy and robustness. [...] Using 37 high throughput omics datasets, covering transcriptomes and metabolomes, we evaluated the classification power of deep learning compared to traditional machine learning methods. Representative deep learning methods, Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN), were deployed and explored in seeking optimal architectures for the best classification performance. Together with five classical supervised classification methods (Linear Discriminant Analysis, Multinomial Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine), MLP and CNN were comparatively tested on the 37 datasets to predict disease stages or to discriminate diseased samples from normal samples. MLPs achieved the highest overall accuracy among all methods tested. More thorough analyses revealed that single hidden layer MLPs with ample hidden units outperformed deeper MLPs. Furthermore, MLP was one of the most robust methods against imbalanced class composition and inaccurate class labels. [...] Our results concluded that shallow MLPs (of one or two hidden layers) with ample hidden neurons are sufficient to achieve superior and robust classification performance in exploiting numerical matrix-formed omics data for diagnosis purpose."
219	Yu, J and Hu, Y and Xu, Y and Wang, J and Kuang, J and Zhang, W and Shao, J and Guo, D and Wang, Y	LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features	BMC Cancer	19	1	263-263	2019	China	http://dx.doi.org/10.1186/s12885-019-5433-7	article	more than 50 samples per group for "good" vs. "poor" distinction (Table S3)	Case-control study	ROC, accuracy, sensitivity, specificity (5-fold CV)	cross-validation	"Non–small cell lung cancer (NSCLC) is the most common type of lung cancer while adenocarcinoma (LUAD) is its most common subtype. [...] However, it remains difficult to find the most significant genetic features and build a high-effective predictive model for treatment outcomes. To confront the challenges, we collected the large-scale LUAD case data with both genome and clinic information (n = 371) from TCGA (The Cancer Genome Atlas) (http://cancergenome.nih.gov), analyzed the somatic mutation difference between the two groups categorized based on the 3-year overall survival, and developed a machine learning model to predict prognosis based on the most significant genetic markers. Through the analysis, we identified a list of genes with different mutation frequencies between different prognosis groups and many were involved in cell-cell adhesion and motility; an absolute majority of the genes showed higher mutation frequencies in the good prognosis group."	"Non–small cell lung cancer (NSCLC) is the most common type of lung cancer while adenocarcinoma (LUAD) is its most common subtype. [...] However, it remains difficult to find the most significant genetic features and build a high-effective predictive model for treatment outcomes. To confront the challenges, we collected the large-scale LUAD case data with both genome and clinic information (n = 371) from TCGA (The Cancer Genome Atlas) (http://cancergenome.nih.gov), analyzed the somatic mutation difference between the two groups categorized based on the 3-year overall survival, and developed a machine learning model to predict prognosis based on the most significant genetic markers. Through the analysis, we identified a list of genes with different mutation frequencies between different prognosis groups and many were involved in cell-cell adhesion and motility; an absolute majority of the genes showed higher mutation frequencies in the good prognosis group."
220	Yu, K H and Levine, D A and Zhang, H and Chan, D W and Zhang, Z and Snyder, M	Predicting Ovarian Cancer Patients' Clinical Response to Platinum-Based Chemotherapy by Their Tumor Proteomic Signatures	J Proteome Res	15	8	2455-2465	2016	United States	http://dx.doi.org/10.1021/acs.jproteome.5b01129	article	"Proteomic profiles of 130 ovarian serous carcinoma patients were analyzed by The Cancer Genome Atlas (TCGA) Clinical Proteomic Tumor Analysis Consortium (CPTAC)"	Cases only (drug response study)	AUC (LOOCV, hold-out test set)	training + test set	"Ovarian cancer is the deadliest gynecologic malignancy in the United States, with most patients diagnosed in the advanced stage of the disease. Platinum-based antineoplastic therapeutics is indispensable to treating advanced ovarian serous carcinoma. However, patients have heterogeneous response to platinum drugs, and it is difficult to predict these inter-individual differences before administering medication. In this study, we investigated the tumor proteomic profiles and clinical characteristics of 130 ovarian serous carcinoma patients analyzed by the Clinical Proteomic Tumor Analysis Consortium (CPTAC), predicted the platinum drug response using supervised machine learning methods [...]. Our data-driven feature selection approach indicated that tumor proteomics profiles contain information for predicting binarized platinum response (P<0.0001). We further built a least absolute shrinkage and selection operator (LASSO)-Cox proportional hazards model that stratified patients into early relapse and late relapse groups (P=0.00013). The top proteomic features indicative of platinum response were involved in ATP synthesis pathways and Ran GTPase binding"	"Ovarian cancer is the deadliest gynecologic malignancy in the United States, with most patients diagnosed in the advanced stage of the disease. Platinum-based antineoplastic therapeutics is indispensable to treating advanced ovarian serous carcinoma. However, patients have heterogeneous response to platinum drugs, and it is difficult to predict these inter-individual differences before administering medication. In this study, we investigated the tumor proteomic profiles and clinical characteristics of 130 ovarian serous carcinoma patients analyzed by the Clinical Proteomic Tumor Analysis Consortium (CPTAC), predicted the platinum drug response using supervised machine learning methods [...]. Our data-driven feature selection approach indicated that tumor proteomics profiles contain information for predicting binarized platinum response (P<0.0001). We further built a least absolute shrinkage and selection operator (LASSO)-Cox proportional hazards model that stratified patients into early relapse and late relapse groups (P=0.00013). The top proteomic features indicative of platinum response were involved in ATP synthesis pathways and Ran GTPase binding"
221	Zang, X and Jones, C M and Long, T Q and Monge, M E and Zhou, M and Walker, L D and Mezencev, R and Gray, A and McDonald, J F and Fernandez, F M	Feasibility of detecting prostate cancer by ultraperformance liquid chromatography-mass spectrometry serum metabolomics	J Proteome Res	13	7	3444-3454	2014	USA	http://dx.doi.org/10.1021/pr500409q	article	Age-matched blood serum samples were obtained from 64 PCa patients (age range 49-65, mean age 59 ± 4 years) and 50 healthy individuals (age range: 45-76, mean age 57 ± 7 years).	Case-control study	accuracy, sensitivity, specificity (10-fold cross-validation)	cross-validation	"Prostate cancer (PCa) is the second leading cause of cancer-related mortality in men. The prevalent diagnosis method is based on the serum Prostate-Specific Antigen (PSA) screening test, which suffers from low specificity, over-diagnosis and over-treatment. In this work, untargeted metabolomic profiling of age-matched serum samples from prostate cancer patients and healthy individuals was performed using ultra performance liquid chromatography coupled to high resolution tandem mass spectrometry (UPLC-MS/MS) and machine learning methods. A metabolite-based in vitro diagnostic multivariate index assay (IVDMIA) was developed to predict the presence of PCa in serum samples with high classification sensitivity, specificity and accuracy. A panel of 40 metabolic spectral features was found to be differential with 92.1% sensitivity, 94.3% specificity, and+P507 93.0% accuracy. The performance of the IVDMIA was higher than the prevalent PSA test. [...] The identification of fatty acids, amino acids, lysophospholipids, and bile acids provided further insights into the metabolic alterations associated with the disease."	"Prostate cancer (PCa) is the second leading cause of cancer-related mortality in men. The prevalent diagnosis method is based on the serum Prostate-Specific Antigen (PSA) screening test, which suffers from low specificity, over-diagnosis and over-treatment. In this work, untargeted metabolomic profiling of age-matched serum samples from prostate cancer patients and healthy individuals was performed using ultra performance liquid chromatography coupled to high resolution tandem mass spectrometry (UPLC-MS/MS) and machine learning methods. A metabolite-based in vitro diagnostic multivariate index assay (IVDMIA) was developed to predict the presence of PCa in serum samples with high classification sensitivity, specificity and accuracy. A panel of 40 metabolic spectral features was found to be differential with 92.1% sensitivity, 94.3% specificity, and+P507 93.0% accuracy. The performance of the IVDMIA was higher than the prevalent PSA test. [...] The identification of fatty acids, amino acids, lysophospholipids, and bile acids provided further insights into the metabolic alterations associated with the disease."
222	Zhang, L and Lv, C and Jin, Y and Cheng, G and Fu, Y and Yuan, D and Tao, Y and Guo, Y and Ni, X and Shi, T	Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma	Frontiers in Genetics	9			2018	China	http://dx.doi.org/10.3389/fgene.2018.00477	article	"The TARGET cohort is comprised of 407 high-risk neuroblastoma samples, including 217 samples with gene expression data and 380 samples with copy number alterations (CNA). Among these obtained samples, 190 has both gene expression and CNA data. The SEQC cohort has a total of 498 neuroblastoma samples, including 176 high-risk and 322 low- or intermediate-risk samples."	Cases only (prognostic sub-group stratification)	C-index, log-rank p-value, AUC (10-fold CV, external validation set)	cross-validation + external cohort validation	"High-risk neuroblastoma is a very aggressive disease, with excessive tumor growth and poor outcomes. A proper stratification of the high-risk patients by prognostic outcome is important for treatment. However, there is still a lack of survival stratification for the high-risk neuroblastoma. To fill the gap, we adopt a deep learning algorithm, Autoencoder, to integrate multi-omics data, and combine it with K-means clustering to identify two subtypes with significant survival differences. By comparing the Autoencoder with PCA, iCluster, and DGscore about the classification based on multi-omics data integration, Autoencoder-based classification outperforms the alternative approaches. Furthermore, we also validated the classification in two independent datasets by training machine-learning classification models, and confirmed its robustness. Functional analysis revealed that MYCN amplification was more frequently occurred in the ultra-high-risk subtype, in accordance with the overexpression of MYC/MYCN targets in this subtype."	"High-risk neuroblastoma is a very aggressive disease, with excessive tumor growth and poor outcomes. A proper stratification of the high-risk patients by prognostic outcome is important for treatment. However, there is still a lack of survival stratification for the high-risk neuroblastoma. To fill the gap, we adopt a deep learning algorithm, Autoencoder, to integrate multi-omics data, and combine it with K-means clustering to identify two subtypes with significant survival differences. By comparing the Autoencoder with PCA, iCluster, and DGscore about the classification based on multi-omics data integration, Autoencoder-based classification outperforms the alternative approaches. Furthermore, we also validated the classification in two independent datasets by training machine-learning classification models, and confirmed its robustness. Functional analysis revealed that MYCN amplification was more frequently occurred in the ultra-high-risk subtype, in accordance with the overexpression of MYC/MYCN targets in this subtype."
223	Zhang, S and Xu, Y and Hui, X and Yang, F and Hu, Y and Shao, J and Liang, H and Wang, Y	Improvement in prediction of prostate cancer prognosis with somatic mutational signatures	Journal of Cancer	8	16	3261-3267	2017	China	http://dx.doi.org/10.7150/jca.21261	article	more than 50 samples per group (both for recurrence status and tumor status)	Cases only (prognosis study)	ROC, accuracy (5-fold CV, training/test split)	training + test set	"Prostate cancer is a leading male malignancy worldwide, while the prognosis prediction remains quite inaccurate. The study aimed to observe whether there was an association between the prognosis of prostate cancer and genetic mutation profile, and to build an accurate prognostic predictor based on the genetic signatures. [...] No significant gene with somatic mutation rate difference was found between prognostic groups of prostate cancer. Total 43 atypical genes were screened for building a support vector machine model to predict prostate cancer prognosis, with an average accuracy of 66% and 64% for 5-fold cross-validation or training-testing evaluation respectively. When combined with the National Institute for Health and Care Excellence (NICE) features, the model could be further improved, with the 5-fold cross-validation accuracy of ~71%, much better than NICE itself (62%)."	"Prostate cancer is a leading male malignancy worldwide, while the prognosis prediction remains quite inaccurate. The study aimed to observe whether there was an association between the prognosis of prostate cancer and genetic mutation profile, and to build an accurate prognostic predictor based on the genetic signatures. [...] No significant gene with somatic mutation rate difference was found between prognostic groups of prostate cancer. Total 43 atypical genes were screened for building a support vector machine model to predict prostate cancer prognosis, with an average accuracy of 66% and 64% for 5-fold cross-validation or training-testing evaluation respectively. When combined with the National Institute for Health and Care Excellence (NICE) features, the model could be further improved, with the 5-fold cross-validation accuracy of ~71%, much better than NICE itself (62%)."
224	Zhao, H and Sun, Q and Li, L and Zhou, J and Zhang, C and Hu, T and Zhou, X and Zhang, L and Wang, B and Li, B and Zhu, T and Li, H	High expression levels of AGGF1 and MFAP4 predict primary platinum-based chemoresistance and are associated with adverse prognosis in patients with serous ovarian cancer	Journal of Cancer	10	2	397-407	2019	China	http://dx.doi.org/10.7150/jca.28127	article	The used TCGA data covers more than 50 samples per group	Cases only (drug resistance prediction)	log-rank test p-value, AUC (5-times 10-fold CV + external validation)	cross-validation + external cohort validation	"Primary platinum-based chemoresistance occurs in approximately one-third of patients with serous ovarian cancer (SOC); however, traditional clinical indicators are poor predictors of chemoresistance. So we aimed to identify novel genes as predictors of primary platinum-based chemoresistance. Gene expression microarray analyses were performed to identify the genes related to primary platinum resistance in SOC on two discovery datasets (GSE51373, GSE63885) and one validation dataset (TCGA). Univariate and multivariate analyses with logistic regression were performed to evaluate the predictive values of the genes for platinum resistance. Machine learning algorithms (linear kernel support vector machine and artificial neural network) were applied to build prediction models. Univariate and multivariate analyses with Cox proportional hazards regression and log-rank tests were used to assess the effects of these gene signatures for platinum resistance on prognosis in two independent datasets (GSE9891, GSE32062). AGGF1 and MFAP4 were found highly expressed in patients with platinum-resistant SOC and independently predicted platinum resistance. Platinum resistance prediction models based on these targets had robust predictive power (highest AUC: 0.8056, 95% CI: 0.6338-0.9773; lowest AUC: 0.7245, 95% CI: 0.6052-0.8438)."	"Primary platinum-based chemoresistance occurs in approximately one-third of patients with serous ovarian cancer (SOC); however, traditional clinical indicators are poor predictors of chemoresistance. So we aimed to identify novel genes as predictors of primary platinum-based chemoresistance. Gene expression microarray analyses were performed to identify the genes related to primary platinum resistance in SOC on two discovery datasets (GSE51373, GSE63885) and one validation dataset (TCGA). Univariate and multivariate analyses with logistic regression were performed to evaluate the predictive values of the genes for platinum resistance. Machine learning algorithms (linear kernel support vector machine and artificial neural network) were applied to build prediction models. Univariate and multivariate analyses with Cox proportional hazards regression and log-rank tests were used to assess the effects of these gene signatures for platinum resistance on prognosis in two independent datasets (GSE9891, GSE32062). AGGF1 and MFAP4 were found highly expressed in patients with platinum-resistant SOC and independently predicted platinum resistance. Platinum resistance prediction models based on these targets had robust predictive power (highest AUC: 0.8056, 95% CI: 0.6338-0.9773; lowest AUC: 0.7245, 95% CI: 0.6052-0.8438)."
225	Zhao, M and Tang, Y and Kim, H and Hasegawa, K	Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer	Cancer Informatics	17			2018	USA	http://dx.doi.org/10.1177/1176935118810215	article	"2509 adult female participants with breast cancer in a prospective cohort study"	Cases only (survival prediction)	ROC, accuracy (10-fold CV)	cross-validation	"Despite existing prognostic markers, breast cancer prognosis remains a difficult subject due to the complex relationships between many contributing factors and survival. This study seeks to integrate multiple clinicopathological and genomic factors with dimensional reduction across machine learning algorithms to compare survival predictions. [...] ROC and accuracy were not significantly different between models (ROC and accuracy around 0.67 and 0.72 across models, respectively). However, ensemble methods resulted in better fit (CS) with stable measures of variable importance across 10 random training/validation splits. K-means clustering of gene expression profiles on training data points along with KNN classification of validation data points was a robust method of dimensional reduction. Furthermore, the gene expression cluster with the highest mortality risk was an influential factor in model prediction. [...] Using machine learning methods to construct predictive models for 5-year survival in patients with breast cancer, we demonstrated discrimination ability across models with new insight into the stability and utility of dimensional reduction on genomic features in breast cancer survival prediction."	"Despite existing prognostic markers, breast cancer prognosis remains a difficult subject due to the complex relationships between many contributing factors and survival. This study seeks to integrate multiple clinicopathological and genomic factors with dimensional reduction across machine learning algorithms to compare survival predictions. [...] ROC and accuracy were not significantly different between models (ROC and accuracy around 0.67 and 0.72 across models, respectively). However, ensemble methods resulted in better fit (CS) with stable measures of variable importance across 10 random training/validation splits. K-means clustering of gene expression profiles on training data points along with KNN classification of validation data points was a robust method of dimensional reduction. Furthermore, the gene expression cluster with the highest mortality risk was an influential factor in model prediction. [...] Using machine learning methods to construct predictive models for 5-year survival in patients with breast cancer, we demonstrated discrimination ability across models with new insight into the stability and utility of dimensional reduction on genomic features in breast cancer survival prediction."
226	Zhu, W and Xie, L and Han, J and Guo, X	The application of deep learning in cancer prognosis prediction	Cancers	12	3		2020	China	http://dx.doi.org/10.3390/cancers12030603	article	review (not applicable)	review			"In this review, we reviewed the most recent published works that used deep learning to build models for cancer prognosis prediction. Deep learning has been suggested to be a more generic model, requires less data engineering, and achieves more accurate prediction when working with large amounts of data. The application of deep learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH. With the burst of multi-omics data, including genomics data, transcriptomics data and clinical information in cancer studies, we believe that deep learning would potentially improve cancer prognosis."	"In this review, we reviewed the most recent published works that used deep learning to build models for cancer prognosis prediction. Deep learning has been suggested to be a more generic model, requires less data engineering, and achieves more accurate prediction when working with large amounts of data. The application of deep learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH. With the burst of multi-omics data, including genomics data, transcriptomics data and clinical information in cancer studies, we believe that deep learning would potentially improve cancer prognosis."

227	Zou, M and Liu, Z and Zhang, X S and Wang, Y	NCC-AUC: an AUC optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data	Bioinformatics	31	20	3330-3338	2015	China	http://dx.doi.org/10.1093/bioinformatics/btv374	article	Dataset 1: 1981 patients with 328 basal-like tumors, 238 HER2+ tumors, 719 luminal A, 490 luminal B and 200 normal-like tumors; Dataset 2: 148 stage IB NSCLC patients	Cases only (prognosis study)	AUC (training + validation data)	training + test set	"In this study, we propose a novel Area Under Curve (AUC) optimization method for multi-biomarker panel identification named Nearest Centroid Classifier for AUC optimization (NCC-AUC). Our method is motived by the connection between AUC score for classification accuracy evaluation and Harrell's concordance index in survival analysis. This connection allows us to convert the survival time regression problem to a binary classification problem. Then an optimization model is formulated to directly maximize AUC and meanwhile minimize the number of selected features to construct a predictor in the nearest centroid classifier framework. NCC-AUC shows its great performance by validating both in genomic data of breast cancer and clinical data of stage IB Non-Small-Cell Lung Cancer (NSCLC)."	"In this study, we propose a novel Area Under Curve (AUC) optimization method for multi-biomarker panel identification named Nearest Centroid Classifier for AUC optimization (NCC-AUC). Our method is motived by the connection between AUC score for classification accuracy evaluation and Harrell's concordance index in survival analysis. This connection allows us to convert the survival time regression problem to a binary classification problem. Then an optimization model is formulated to directly maximize AUC and meanwhile minimize the number of selected features to construct a predictor in the nearest centroid classifier framework. NCC-AUC shows its great performance by validating both in genomic data of breast cancer and clinical data of stage IB Non-Small-Cell Lung Cancer (NSCLC)."
228	Zou, M and Zhang, P J and Chen, L and Tian, Y P and Wang, Y	Identifying joint biomarker panel from multiple level dataset by an optimization model	Biomark Med	10	6	567-575	2016	China	http://dx.doi.org/10.2217/bmm-2015-0022	article	"101 colorectal cancer and 95 benign samples"	Case-control study	accuracy (LOOCV + test set)	cross-validation + test set	"Joint biomarker panel takes advantage of coherence across multiple level datasets and holds the promise to improve disease diagnosis accuracy. [...] We collected 101 colorectal cancer and 95 benign samples, measured the molecular concentrations by serum assays and mass spectra, and developed LPGLO (Linear Programming based on Group Lasso Optimization) to identify the joint biomarker panel. [...] A joint biomarker panel, with ten serum biomarkers and six mass spectra peaks, achieves LOOCV accuracy 0.8724, which is better than the optimal panels identified from separate datasets (LOOCV = 0.7551 for mass spectra only or 0.8265 for serum assay only) and the simply merged dataset (LOOCV = 0.8622)."	"Joint biomarker panel takes advantage of coherence across multiple level datasets and holds the promise to improve disease diagnosis accuracy. [...] We collected 101 colorectal cancer and 95 benign samples, measured the molecular concentrations by serum assays and mass spectra, and developed LPGLO (Linear Programming based on Group Lasso Optimization) to identify the joint biomarker panel. [...] A joint biomarker panel, with ten serum biomarkers and six mass spectra peaks, achieves LOOCV accuracy 0.8724, which is better than the optimal panels identified from separate datasets (LOOCV = 0.7551 for mass spectra only or 0.8265 for serum assay only) and the simply merged dataset (LOOCV = 0.8622)."
229	Shamir, Ron and Klein, Christine and Amar, David and Vollstedt, Eva Juliane and Bonin, Michael and Usenovic, Marija and Wong, Yvette C. and Maver, Ales and Poths, Sven and Safer, Hershel and Corvol, Jean Christophe and Lesage, Suzanne and Lavi, Ofer and Deuschl, Günther and Kuhlenbaeumer, Gregor and Pawlack, Heike and Ulitsky, Igor and Kasten, Meike and Riess, Olaf and Brice, Alexis and Peterlin, Borut and Krainc, Dimitri	Analysis of blood-based gene expression in idiopathic Parkinson disease	Neurology	89	16	1676-1683	2017	Croatia	http://dx.doi.org/10.1212/wnl.0000000000004516	article	"Whole-blood gene expression profiles were collected from a total of 523 individuals. After preprocessing, the data contained 486 gene profiles (n = 205 PD, n = 233 controls, n = 48 other neurodegenerative diseases) that were partitioned into training, validation, and independent test cohorts to identify and validate a gene signature."	Case-control study	AUC (cross-validation + external test set)	cross-validation + external cohort validation	The objective was "To examine whether gene expression analysis of a large-scale Parkinson disease (PD) patient cohort produces a robust blood-based PD gene signature compared to previous studies that have used relatively small cohorts (≤220 samples). [...] gene signature of 100 probes that mapped to 87 genes, corresponding to 64 upregulated and 23 downregulated genes differentiating between patients with idiopathic PD and controls, was identified with the training cohort and successfully replicated in both an independent validation cohort (area under the curve [AUC] = 0.79, p = 7.13E–6) and a subsequent independent test cohort (AUC = 0.74, p = 4.2E–4)."	The objective was "To examine whether gene expression analysis of a large-scale Parkinson disease (PD) patient cohort produces a robust blood-based PD gene signature compared to previous studies that have used relatively small cohorts (≤220 samples). [...] gene signature of 100 probes that mapped to 87 genes, corresponding to 64 upregulated and 23 downregulated genes differentiating between patients with idiopathic PD and controls, was identified with the training cohort and successfully replicated in both an independent validation cohort (area under the curve [AUC] = 0.79, p = 7.13E–6) and a subsequent independent test cohort (AUC = 0.74, p = 4.2E–4)."
230	Glaab, E	Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification	Briefings in Bioinformatics	17	3	440-452	2016	England	http://dx.doi.org/10.1093/bib/bbv044	article	review (not applicable)	review			"In recent years, new approaches for building multivariate biomarker models on omics data have been proposed, which exploit prior biological knowledge from molecular networks and cellular pathways to address these limitations. This survey provides an overview of these recent developments and compares pathway- and network-based specimen classification approaches in terms of their utility for improving model robustness, accuracy and biological interpretability. Different routes to translate omics-based multifactorial biomarker models into clinical diagnostic tests are discussed, and a previous study is presented as example."	"In recent years, new approaches for building multivariate biomarker models on omics data have been proposed, which exploit prior biological knowledge from molecular networks and cellular pathways to address these limitations. This survey provides an overview of these recent developments and compares pathway- and network-based specimen classification approaches in terms of their utility for improving model robustness, accuracy and biological interpretability. Different routes to translate omics-based multifactorial biomarker models into clinical diagnostic tests are discussed, and a previous study is presented as example."
231	Shreve, J and Meggendorfer, M and Awada, H and Mukherjee, S and Walter, W and Hutter, S and Makhoul, A and Hilton, C B and Radakovich, N and Nagata, Y and Rouphail, Y and Adema, V and Kerr, C M and Patel, B J and Kuzmanovic, T and Maciejewski, J P and Haferlach, C and Sekeres, M A and Haferlach, T and Nazha, A	A personalized prediction model to risk stratify patients with acute myeloid leukemia (AML) using artificial intelligence	Blood	134			2019	USA	http://dx.doi.org/10.1182/blood-2019-128066	meeting abstract	"A total of 792,779 genomic and clinical data points from 3,421 pts were analyzed. The cohort was comprised of five independent datasets: 443 pts from the Beat AML Master Trial (Tyner et al, Nature, 2018), 855 pts from Cleveland Clinic, 414 pts from Munich Leukemia Laboratory (MLL), 1,509 pts from the German-Austrian Study Group (Papaemmanuil et al, NEJM, 2016), and 200 pts from The Cancer Genome Atlas (NEJM, 2013)."	Subtype categorization	C-index (training + test cohorts)	external cohort validation	"Genomic alterations have a differential impact on OS [overall survival] in each cytogenetic risk group, highlighting the complexity of incorporating these mutations into risk stratification. A personalized prediction model based on clinical-genomic data can accurately provide survival unique to each individual pt and can significantly outperform ELN [European LeukemiaNet] classifications or any currently available models."	"Genomic alterations have a differential impact on OS [overall survival] in each cytogenetic risk group, highlighting the complexity of incorporating these mutations into risk stratification. A personalized prediction model based on clinical-genomic data can accurately provide survival unique to each individual pt and can significantly outperform ELN [European LeukemiaNet] classifications or any currently available models."
232	Baer, C and Walter, W and Stengel, A and Hutter, S and Meggendorfer, M and Kern, W and Haferlach, C and Haferlach, T	Molecular classification of AML-MRC reveals a distinct profile and identifies MRC-like patients with poor overall survival	Blood	134			2019	Germany	http://dx.doi.org/10.1182/blood-2019-128234	article	"According to WHO standards 165/739 (22%) cases fulfilled MRC criteria (96 male; 69 female). The non-MRC cohort (n=574) represents a heterogeneous AML population incl. the WHO defined recurrent cytogenetic abnormalities or t-AML (301 male, 273 female). "	Case-control study	true positive rate, false positive rate (10-fold cross-validation)	cross-validation	"Using patients' history and genetic information instead of morphology allow to identify 96-99% of AML-MRC as defined in WHO today"	"Using patients' history and genetic information instead of morphology allow to identify 96-99% of AML-MRC as defined in WHO today"
233	Zhu, W and Xie, L and Han, J and Guo, X	The application of deep learning in cancer prognosis prediction	Cancers	12	3		2020	China	http://dx.doi.org/10.3390/cancers12030603	article	review (not applicable)	Review			"In this review, we reviewed the most recent published works that used deep learning to build models for cancer prognosis prediction. Deep learning has been suggested to be a more generic model, requires less data engineering, and achieves more accurate prediction when working with large amounts of data. The application of deep learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH."	"In this review, we reviewed the most recent published works that used deep learning to build models for cancer prognosis prediction. Deep learning has been suggested to be a more generic model, requires less data engineering, and achieves more accurate prediction when working with large amounts of data. The application of deep learning in cancer prognosis has been shown to be equivalent or better than current approaches, such as Cox-PH."
234	Angelino, P and Hosseinian Ehrensberger, S and Ciarloni, L and Despraz, J and Dorta, G and Perez-Uribe, A and Morgenthaler, S and Delorenzi, M	Discovery of an immunotranscriptomics signature in blood for early colorectal cancer detection	Annals of Oncology	30		v45-v45	2019	Switzerland	http://dx.doi.org/10.1093/annonc/mdz239.056	meeting abstract	"The cohort included 189 subjects with CRC, 115 with advanced adenoma (AA), 39 with other types of cancer (OC) as well as 218 individuals without any colorectal lesions (CON)."	Case-control study	AUC, sensitivity, specificity (independent set validation)	training + test set	"Mapping the reaction of the immune system to onset of cancer and disease identification through application ML methods is a new approach to ensure an unbiased, genome-wide, unsupervised ene expression analysis for a highly specific biomarker identification."	"Mapping the reaction of the immune system to onset of cancer and disease identification through application ML methods is a new approach to ensure an unbiased, genome-wide, unsupervised ene expression analysis for a highly specific biomarker identification."
235	Kuwabara, H and Iwabuthi, A and Soya, R and Enomoto, M and Ishizaki, T and Tsuchida, A and Nagakawa, Y and Katsumata, K and Sugimoto, M	Salivary metabolomics for colorectal cancer detection	Annals of Oncology	30		v46-v46	2019	Japan	http://dx.doi.org/10.1093/annonc/mdz239.058	meeting abstract	"A total of 2,602 unstimulated saliva samples were collected from 231 subjects with CRC, 99 subjects with polyps, and 2272 subjects with healthy controls"	Case-control study	AUC	training + test set	"Combinations of salivary metabolites show high potential as a screening tool for CRC"	"Combinations of salivary metabolites show high potential as a screening tool for CRC"
236	Hilton, C B and Meggendorfer, M and Sekeres, M A and Shreve, J and Radakovich, N and Rouphail, Y and Walter, W and Hutter, S and Padron, E and Savona, M R and Gerds, A T and Mukherjee, S and Smith, B N and Nagata, Y and Hirsch, C M and Komrokji, R S and Jha, B K and Haferlach, C and Maciejewski, J P and Haferlach, T and Nazha, A	Geno-clinical model for the diagnosis of bone marrow myeloid neoplasms	Blood	134			2019	USA	http://dx.doi.org/10.1182/blood-2019-126967	meeting abstract	"Of 2471 pts, 1306 had MDS, 223 had ICUS, 107 had CCUS, 478 had CMML, 89 had MDS/MPN, 79 had PV, 90 had ET, and 99 had PMF."	Case-control study	AUC ("The cohort was randomly divided into learner (80%) and validation (20%) cohorts")	training + test set	"We propose a new approach using interpretable, individualized modeling to predict myeloid neoplasm phenotypes based on genomic and clinical data without bone marrow biopsy data. This approach can aid clinicians and hematopathologists when encountering pts with cytopenias and suspicion for these disorders. The model also provides feature attributions that allow for quantitative understanding of the complex interplay among genotypes, clinical variables, and phenotypes."	"We propose a new approach using interpretable, individualized modeling to predict myeloid neoplasm phenotypes based on genomic and clinical data without bone marrow biopsy data. This approach can aid clinicians and hematopathologists when encountering pts with cytopenias and suspicion for these disorders. The model also provides feature attributions that allow for quantitative understanding of the complex interplay among genotypes, clinical variables, and phenotypes."
237	Liu, X D and Wu, H and Li, Y and Liu, X and Zhang, Z and Yu, L and Qin, Z and Su, Z and Liu, R and He, Q and Dai, M and Liang, Z	Early detection of pancreatic ductal adenocarcinoma using methylation signatures in circulating tumour DNA	Annals of Oncology	30		v261-v262	2019	China	http://dx.doi.org/10.1093/annonc/mdz247.013	meeting abstract	"We have collected freshly frozen clinical PDAC tissues (N=46), paratumour pancreas tissues (N=30), PDAC plasma samples (N=120), chronical pancreatitis plasma samples (N=90), and normal plasma samples (N=100)."	Case-control study	AUC	cross-validation	"Using multiple metrics onMHBs, we identified PDAC-specific DNAmethylation markers, some of which are functionally overlapped with previously reported markers. These markers are candidate biomarkers or non-invasive PDAC screening."	"Using multiple metrics onMHBs, we identified PDAC-specific DNAmethylation markers, some of which are functionally overlapped with previously reported markers. These markers are candidate biomarkers or non-invasive PDAC screening."
238	Adom, D and Rowan, C and Adeniyan, T and Yang, J and Paczesny, S	Biomarkers for Allogeneic HCT Outcomes	Front Immunol	11		673-673	2020	USA	http://dx.doi.org/10.3389/fimmu.2020.00673	article	review (not applicable)	Review			"Advances in technology in the field of omics have permitted the discovery of numerous biomarkers for identification of complications post-HCT and signature of the beneficial GVT. A good biomarker has several features: has been developed through the different phases including discovery and validation in large independent cohorts, use a cost efficient non-invasive robust assay that has been qualified."	"Advances in technology in the field of omics have permitted the discovery of numerous biomarkers for identification of complications post-HCT and signature of the beneficial GVT. A good biomarker has several features: has been developed through the different phases including discovery and validation in large independent cohorts, use a cost efficient non-invasive robust assay that has been qualified."
239	Ahmed, K T and Park, S and Jiang, Q and Yeu, Y and Hwang, T and Zhang, W	Network-based drug sensitivity prediction	BMC Med Genomics	13		193-193	2020	USA	http://dx.doi.org/10.1186/s12920-020-00829-3	article	"The feature selection methods and prediction models were tested on 144 NSCLC cell lines RNA-seq gene expression dataset [34]. All the 144 cell lines were screened by the same drugs and the AUC and ED50 scores for each drug on each cell line are available in this study."	Drug response prediction	correlation (70% as the training set, and 30% as the test set)	training + test set	"This study introduced a network-based feature selection method and two graph-based neural network models for drug sensitivity prediction. Comparing to the Pearson correlation coefficients for feature selection, four canonical prediction methods, and deep neural network on an NSCLC cell line dataset, we have made several useful observations. First, the network-based feature selection method identifies more representative features based on gene co-expression network for drug sensitivity prediction. Second, Random Forest outperforms all the other canonical prediction methods and deep neural network models, Third, the graph-based neural network models show better drug response prediction performance compared to DNN, however, it is still worse than the performance of the canonical prediction methods and a dataset with larger sample size is needed to further increase the prediction accuracy."	"This study introduced a network-based feature selection method and two graph-based neural network models for drug sensitivity prediction. Comparing to the Pearson correlation coefficients for feature selection, four canonical prediction methods, and deep neural network on an NSCLC cell line dataset, we have made several useful observations. First, the network-based feature selection method identifies more representative features based on gene co-expression network for drug sensitivity prediction. Second, Random Forest outperforms all the other canonical prediction methods and deep neural network models, Third, the graph-based neural network models show better drug response prediction performance compared to DNN, however, it is still worse than the performance of the canonical prediction methods and a dataset with larger sample size is needed to further increase the prediction accuracy."
240	Ahmed, Z and Mohamed, K and Zeeshan, S and Dong, X	Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine	Database (Oxford)	2020			2020	USA	http://dx.doi.org/10.1093/database/baaa010	article	review (not applicable)	Review			"Precision medicine is progressing but with many challenges lying ahead (255), which require addition of useful analytic tools, technologies, databases and approaches (4,6) to efficiently augment networking and interoperability of clinical, laboratory and public health systems, as well as address ethical and social issues related to the privacy and protection of healthcare and omics data with effective balance. This will also require more efficient management of massive amounts of generated data, as well as earlier mined consensus and actionable data."	"Precision medicine is progressing but with many challenges lying ahead (255), which require addition of useful analytic tools, technologies, databases and approaches (4,6) to efficiently augment networking and interoperability of clinical, laboratory and public health systems, as well as address ethical and social issues related to the privacy and protection of healthcare and omics data with effective balance. This will also require more efficient management of massive amounts of generated data, as well as earlier mined consensus and actionable data."
241	Ambati, A and Ju, Y E and Lin, L and Olesen, A N and Koch, H and Hedou, J J and Leary, E B and Sempere, V P and Mignot, E and Taheri, S	Proteomic biomarkers of sleep apnea	Sleep	43	11		2020	Qatar	http://dx.doi.org/10.1093/sleep/zsaa086	article	"serum samples from 713 individuals in the Stanford Sleep Cohort"	Case-control study	AUC (10-fold CV + validation)	cross-validation + test set	"An OAHI [Obstructive Apnea Hypopnea Index] machine learning classifier (OAHI >=15 vs OAHI<15) trained on SomaScan protein measures alone performed robustly, achieving 76% accuracy in a validation dataset. Multiplex protein assays offer diagnostic potential and provide new insights into the biological basis of sleep disordered breathing."	"An OAHI [Obstructive Apnea Hypopnea Index] machine learning classifier (OAHI >=15 vs OAHI<15) trained on SomaScan protein measures alone performed robustly, achieving 76% accuracy in a validation dataset. Multiplex protein assays offer diagnostic potential and provide new insights into the biological basis of sleep disordered breathing."
242	Aslam, M A and Xue, C and Wang, K and Chen, Y and Zhang, A and Cai, W and Ma, L and Yang, Y and Sun, X and Liu, M and Pan, Y and Munir, M A and Song, J and Cui, D	SVM based classification and prediction system for gastric cancer using dominant features of saliva	Nano Biomedicine and Engineering	12	1	1-13	2020	China	http://dx.doi.org/10.5101/nbe.v12i1.p1-13	article	"220 saliva samples were collected from the non-cancerous and gastric cancerous persons"	Case-control study	AUC, accuracy, specificity, sensitivity (10-fold cross-validation. "To avoid the overfitting issue, we used an ES approach. We controlled the error of the network during the training phase and stopped the training if the model undergoes the overfitting")	cross-validation	"Based on the achieved results, the sigmoid kernel has produced the best classification results with an accuracy of 97.18% %, specificity of 97.44%, and specificity of 96.88% during the testing phase, whereas the results of the other kernel-based models, SVM has not produced good results. Our proposed method for the classification of GC is non-invasive, cheap, and faster."	"Based on the achieved results, the sigmoid kernel has produced the best classification results with an accuracy of 97.18% %, specificity of 97.44%, and specificity of 96.88% during the testing phase, whereas the results of the other kernel-based models, SVM has not produced good results. Our proposed method for the classification of GC is non-invasive, cheap, and faster."
243	Awada, H and Durmaz, A and Gurnari, C and Kishtagari, A and Meggendorfer, M and Kerr, C M and Kuzmanovic, T and Durrani, J and Nagata, Y and Radivoyevitch, T and Advani, A S and Ravandi, F and Carraway, H E and Nazha, A and Haferlach, C and Saunthararajah, Y and Scott, J and Visconte, V and Kantarjian, H M and Kadia, T M and Sekeres, M A and Haferlach, T and Maciejewski, J P	The application of machine learning to improve the subclassification and prognostication of acute myeloid leukemia	Blood	136		28-28	2020	USA	http://dx.doi.org/10.1182/blood-2020-139013	meeting abstract	"We collected and analyzed genomic data from a multicenter cohort of 6788 AML patients"	Case-control study	accuracy (cross-validation)	cross-validation	"In conclusion, the heterogeneity inherent in the genomic changes across nearly 7000 AML patients is too vast for traditional prediction methods. Using newer ML methods, however, we were able to decipher a set of prognostic subgroups predictive of survival, allowing us to move AML into the era of personalized medicine."	"In conclusion, the heterogeneity inherent in the genomic changes across nearly 7000 AML patients is too vast for traditional prediction methods. Using newer ML methods, however, we were able to decipher a set of prognostic subgroups predictive of survival, allowing us to move AML into the era of personalized medicine."
244	Bader, J M and Geyer, P E and Müller, J B and Strauss, M T and Koch, M and Leypoldt, F and Koertvelyessy, P and Bittner, D and Schipke, C G and Incesoy, E I and Peters, O and Deigendesch, N and Simons, M and Jensen, M K and Zetterberg, H and Mann, M	Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease	Mol Syst Biol	16	6	e9356-e9356	2020	Germany	http://dx.doi.org/10.15252/msb.20199356	article	"From three independent studies (197 individuals), we characterize differences in proteins by AD status)	Case-control study	AUC (k-fold CV, k = 6)	cross-validation	"We found that an ensemble method-based classifier reached high specificity (87%) and sensitivity (82%), while showing promising generalizability. Intriguingly, tau itself, one of the glycolysis-related proteins, and an immunological factor were selected by the machine learning algorithm as the most important features for classification, proving further validation of our biomarker panel and biomarker identification pipeline. The modeling also indicated that additional and more uniform training data could further improve diagnostic performance."	"We found that an ensemble method-based classifier reached high specificity (87%) and sensitivity (82%), while showing promising generalizability. Intriguingly, tau itself, one of the glycolysis-related proteins, and an immunological factor were selected by the machine learning algorithm as the most important features for classification, proving further validation of our biomarker panel and biomarker identification pipeline. The modeling also indicated that additional and more uniform training data could further improve diagnostic performance."

245	Bergamaschi, A and Ku, J and Ning, Y and Collin, F and Ellison, C and Phillips, T and McCarthy, E and Wang, W and Antoine, M and Haan, D and Scott, A and Lloyd, P and Guler, G and Ashworth, A and Quake, S and Levy, S	Epigenomic detection of multiple cancers in plasma derived cell free DNA	Cancer Research	80		16		2020	USA	http://dx.doi.org/10.1158/1538-7445.AM2020-783	meeting abstract	"Cancer and control patient cfDNA cohorts were accrued from multiple sites consisting of 48 breast, 55 lung, 32 prostate and 2 pancreatic datasets consisting of 41 and 53 cancer subjects (Set 1 and 2)"	Case-control study	AUC (5-foldCV)	cross-validation	"These findings suggest that 5hmC changes in cfDNA enable non-invasive detection of early stage breast, pancreatic, prostate, and lung cancers. Furthermore, 5hmC profiling in cfDNA may enable the prediction of clinically relevant features such as tumor size in breast adenocarcinoma or indolent disease in prostate cancer. Finally, this study identified a suite of 5hmC biomarkers that may be further validated in larger, and more diverse, patient cohorts."	
246	Berry, S E and Valdes, A M and Drew, D A and Asnicar, F and Mazidi, M and Wolf, J and Capdevila, J and Hadjigeorgiou, G and Davies, R and Al Khatib, H and Bonnett, C and Ganesh, S and Bakker, E and Hart, D and Mangino, M and Merino, J and Linenberg, I and Wyatt, P and Ordovas, J M and Gardner, C D and Delahanty, L M and Chan, A T and Segata, N and Franks, P W and Spector, T D	Human postprandial responses to food and potential for precision nutrition	Nat Med	26	6		964-973	2020	UK	https://www.nature.com/articles/s41591-020-0934-0	article	"We recruited n = 1,002 twins and unrelated healthy adults in the United Kingdom to the PREDICT 1 study and assessed postprandial metabolic responses in a clinical setting and at home"	Response to food intake prediction	correlation	training + test set	"We developed a machine-learning model that predicted both triglyceride (r = 0.47) and glycemic (r = 0.77) responses to food intake. These findings may be informative for developing personalized diet strategies."	"We developed a machine-learning model that predicted both triglyceride (r = 0.47) and glycemic (r = 0.77) responses to food intake. These findings may be informative for developing personalized diet strategies."
247	Bhalla, S and Kaur, H and Kaur, R and Sharma, S and Raghava, G P S	Expression based biomarkers and models to classify early and late-stage samples of Papillary Thyroid Carcinoma	PLoS One	15	4		e0231629-e0231629	2020	India	http://dx.doi.org/10.1371/journal.pone.0231629	article	"In this study, we used the THCA RNA-Seq dataset of The Cancer Genome Atlas, consisting of 500 cancer and 58 normal (adjacent non-tumorous) samples"	Case-control study	AUC (cross-validation + external validation)	cross-validation + test set	"In conclusion, 36 RNA-transcripts based SVC prediction model attained considerable performance in segregating the early-stage and late-stage PTC [Papillary Thyroid Carcinoma] tissue samples with F1 score of 0.75. In addition, prediction models based on five protein-coding transcripts categorized tumorous and non-tumorous samples of patients with high F1 score of 0.97."	"In conclusion, 36 RNA-transcripts based SVC prediction model attained considerable performance in segregating the early-stage and late-stage PTC [Papillary Thyroid Carcinoma] tissue samples with F1 score of 0.75. In addition, prediction models based on five protein-coding transcripts categorized tumorous and non-tumorous samples of patients with high F1 score of 0.97."
248	Bigelow, E G and Baras, A and Yarchoan, M and Jaffee, E M	Machine learning methods to identifysalient genomic predictors of clinical responses toimmune checkpoint inhibitor therapy	Cancer Research	80		16		2020	USA	http://dx.doi.org/10.1158/1538-7445.AM2020-5670	meeting abstract	"paired tumor and normal whole-exome sequencing (WES) data from clinically annotated tumor specimens treated with anti-CTLA4 (n=145, melanoma) and anti-PD1 therapies (n=94, NSCLC, melanoma, head and neck, bladder cancer, and others)"	Case-control study	precision, recall ("A random forest classifier was trained and tested on various subsets of the dataset")	training + test set	"Relative feature importance indicates that tumor mutational burden is the main predictive feature, with number of frameshift mutations, patient age, splice site mutations, and the novel tumor heterogeneity score contributing roughly an order of magnitude less weight for this dataset. Cancer type and various aggregate neoantigen counts did not contribute to predictions across analyses. Novel findings include an evaluation of the ability to predict ICI patient response using random forest classifiers, and a semi-quantitative evaluation of the relative contributions of features included the model. Additionally, analyses suggest the ability to transfer learning on one tumor type to identify responders in other diseases."	"Relative feature importance indicates that tumor mutational burden is the main predictive feature, with number of frameshift mutations, patient age, splice site mutations, and the novel tumor heterogeneity score contributing roughly an order of magnitude less weight for this dataset. Cancer type and various aggregate neoantigen counts did not contribute to predictions across analyses. Novel findings include an evaluation of the ability to predict ICI patient response using random forest classifiers, and a semi-quantitative evaluation of the relative contributions of features included the model. Additionally, analyses suggest the ability to transfer learning on one tumor type to identify responders in other diseases."
249	Brown, E and Karrar, A and Hellings, S and Stepanova, M and Warrack, B and Lam, B and Onorato, J and Felix, S and Apfel, A and Jeffers, T and Rajput, B and Charles, E and Nader, F and Luo, Y and Reily, M and Zhao, L and Thompson, J and Goodman, Z and Younossi, Z	Metabolomics composite biomarkers selected by machine learning predicts NASH	Journal of Hepatology	73			S409-S410	2020	USA	http://dx.doi.org/10.1016/S0168-8278(20)31304-0	meeting abstract	"Serum samples were obtained from 100 biopsy-proven NASH (F0-F4) and 50 NAFLD patients without NASH"	Case-control study	AUC, repeated loops of crossvalidation	cross-validation	"Applying elastic nets with cross-validation, we find that 12 metabolites plus patient diabetes status classify NASH vs NAFL patients with predicted out-ofsample power AUC = 0.824. In addition to dicarboxylic and bile acids, predictors include lipid and microbiome-derived metabolites."	"Applying elastic nets with cross-validation, we find that 12 metabolites plus patient diabetes status classify NASH vs NAFL patients with predicted out-ofsample power AUC = 0.824. In addition to dicarboxylic and bile acids, predictors include lipid and microbiome-derived metabolites."
250	Cai, W Y and Dong, Z N and Fu, X T and Lin, L Y and Wang, L and Ye, G D and Luo, Q C and Chen, Y C	Identification of a Tumor Microenvironment-relevant Gene set-based Prognostic Signature and Related Therapy Targets in Gastric Cancer	Theranostics	10		19	8633-8647	2020	China	http://dx.doi.org/10.7150/thno.47938	article	"a prognostic model was established based on gastric cancer gene expression datasets from 1699 patients from five independent cohorts"	Prognostic study	concordance index (C-index; was calculated to determine the discrimination of the nomogram via a bootstrap method with 1000 resamples)	external cohort validation	"As a tumor microenvironment-relevant gene set-based prognostic signature, the GPSGC model provides an effective approach to evaluate GC [Gastric Cancer] patient survival outcomes and may prolong overall survival by enabling the selection of individualized targeted therapy."	"As a tumor microenvironment-relevant gene set-based prognostic signature, the GPSGC model provides an effective approach to evaluate GC [Gastric Cancer] patient survival outcomes and may prolong overall survival by enabling the selection of individualized targeted therapy."
251	Cammarota, G and Ianiro, G and Ahern, A and Carbone, C and Temko, A and Claesson, M J and Gasbarrini, A and Tortora, G	Gut microbiome, big data and machine learning to promote precision medicine for cancer	Nat Rev Gastroenterol Hepatol	17		10	635-648	2020	Italy	http://dx.doi.org/10.1038/s41575-020-0327-3	article	review (not applicable)	Review			"In this Perspective, we provide a brief overview on the role of gut microbiome in cancer and focus on the need, role and limitations of a machine learning-driven approach to analyse large amounts of complex health-care information in the era of big data. "	"In this Perspective, we provide a brief overview on the role of gut microbiome in cancer and focus on the need, role and limitations of a machine learning-driven approach to analyse large amounts of complex health-care information in the era of big data. "
252	Cascianelli, S and Molineris, I and Isella, C and Masseroli, M and Medico, E	Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer	Sci Rep	10	1		14071-14071	2020	Italy	http://dx.doi.org/10.1038/s41598-020-70832-2	article	"A 220-sample training set was extracted randomly from the TCGA dataset, respecting the same 60/40 ER+/ER− proportion of the PAM50 training set. All the remaining 597 cases were instead included in the TCGA test set"	Tumor subtype prediction	accuracy (10-fold CV, training/test set)	cross-validation + external cohort validation	"Thus, in conclusion, the main contribution of this paper is twofold: 1. Propose the AWCA reference construction approach to face the proved issues of the standard PAM50 algorithm; 2. Define RNA-seq-based classification approaches to perform single-sample BC intrinsic subtyping with external-AWCA-based PAM50 or regularized mLR methods. These strategies appeared valuable to favor the use of RNA-seq in BC [Breast Cancer] clinical practice and are worthy of other studies on heterogeneous RNA-seq data, to evaluate and strengthen the reliability of their intrinsic subtyping methods."	"Thus, in conclusion, the main contribution of this paper is twofold: 1. Propose the AWCA reference construction approach to face the proved issues of the standard PAM50 algorithm; 2. Define RNA-seq-based classification approaches to perform single-sample BC intrinsic subtyping with external-AWCA-based PAM50 or regularized mLR methods. These strategies appeared valuable to favor the use of RNA-seq in BC [Breast Cancer] clinical practice and are worthy of other studies on heterogeneous RNA-seq data, to evaluate and strengthen the reliability of their intrinsic subtyping methods."
253	Catalina, M D and Bachali, P and Yeo, A E and Geraci, N S and Petri, M A and Grammer, A C and Lipsky, P E	Patient ancestry significantly contributes to molecular heterogeneity of systemic lupus erythematosus	JCI Insight	5		15		2020	USA	http://dx.doi.org/10.1172/jci.insight.140380	article	"The ILL1 and ILL2 clinical trials had microarray expression data for 1566 female patients of self-described ancestry: AA (n = 216), EA (n = 1118), and NAA (mostly from South America [n = 232]; top 3 countries of origin Peru [n = 81], Ecuador [n = 30], and Guatemala [n = 27]) and 124 male patients of self-described ancestry: AA (n = 14), EA (n = 93), NAA (n = 17)"	Subtype categorization	accuracy (10-fold CV)	cross-validation	"Although standard therapy affected every gene signature and significantly increased myeloid cell signatures, logistic regression analysis determined that ancestral background significantly changed 23 of 34 gene signatures. Additionally, the strongest association to gene expression changes was found with autoantibodies, and this also had etiology in ancestry: the AA predisposition to have both RNP and dsDNA autoantibodies compared with EA predisposition to have only anti-dsDNA. A machine learning approach was used to determine a gene signature characteristic to distinguish AA SLE and was most influenced by genes characteristic of the perturbed B cell axis in AA SLE patients."	"Although standard therapy affected every gene signature and significantly increased myeloid cell signatures, logistic regression analysis determined that ancestral background significantly changed 23 of 34 gene signatures. Additionally, the strongest association to gene expression changes was found with autoantibodies, and this also had etiology in ancestry: the AA predisposition to have both RNP and dsDNA autoantibodies compared with EA predisposition to have only anti-dsDNA. A machine learning approach was used to determine a gene signature characteristic to distinguish AA SLE and was most influenced by genes characteristic of the perturbed B cell axis in AA SLE patients."
254	Cedars, A M and Manlhiot, C and Ko, J and Bottiglieri, T and Weingarten, A and Opotowsky, A and Kutty, S	ARTIFICIAL INTELLIGENCE FACILITATED METABOLOMIC PROFILING IN ADULT CONGENITAL HEART DISEASE (ACHD)	Journal of the American College of Cardiology	75		11	552-552	2020	USA	http://dx.doi.org/10.1016/S0735-1097(20)31179-7	meeting abstract	"We analyzed 674 metabolites in 186 serum samples from 101 non-fasting ACHD patients followed regularly at a single institution, including repeated samples at different times."	Subtype categorization	AUC (cross-validation)	cross-validation	"After correction for repeated measures, clustering identified 3 separate metabolic/clinical profiles: 1) right ventricular (RV) dysfunction, arrhythmia and dyspnea (n=107), 2) complex biventricular disease with hypoxia and lower educational level (n=79) and 3) individuals managing well with valvular and septal defects (n=42). Metabolomic data permitted the creation of models associated with prevalent arrhythmia (cross validated AUC 0.67), patient-reported exertion-associated shortness of breath (AUC 0.58) and RV dysfunction (AUC 0.62)."	"After correction for repeated measures, clustering identified 3 separate metabolic/clinical profiles: 1) right ventricular (RV) dysfunction, arrhythmia and dyspnea (n=107), 2) complex biventricular disease with hypoxia and lower educational level (n=79) and 3) individuals managing well with valvular and septal defects (n=42). Metabolomic data permitted the creation of models associated with prevalent arrhythmia (cross validated AUC 0.67), patient-reported exertion-associated shortness of breath (AUC 0.58) and RV dysfunction (AUC 0.62)."
255	Chan, S and Reddy, V and Myers, B and Thibodeaux, Q and Brownstone, N and Liao, W	Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations	Dermatology and Therapy	10	3		365-386	2020	USA	http://dx.doi.org/10.1007/s13555-020-00372-0	article	review (not applicable)	Review			" While ML methods are powerful, they are still similar to previous clinical tools in that physician interpretation is crucial for implementation in a real-world setting. We should also be cognizant of how potential biases can interfere with the black box nature of these algorithms. It is also important to make these technologies inclusive of skin of color. Further research in ML should be transparent by making algorithms and datasets available to the public for further validation and testing."	" While ML methods are powerful, they are still similar to previous clinical tools in that physician interpretation is crucial for implementation in a real-world setting. We should also be cognizant of how potential biases can interfere with the black box nature of these algorithms. It is also important to make these technologies inclusive of skin of color. Further research in ML should be transparent by making algorithms and datasets available to the public for further validation and testing."
256	Chavarriaga, J and Moreno, C	Precision Medicine, Artificial Intelligence, and Genomic Markers in UrologyDo we need to Tailor our Clinical Practice?	Urologia Colombiana	29	3		158-167	2020	Colombia	http://dx.doi.org/10.1055/s-0040-1714148	article	review (not applicable)	Review			"Urology is a constantly changing specialty with a wide range of therapeutic breakthroughs, a huge understanding of the genomic expression profiles for each urological cancer and a tendency to use cutting-edge technology to treat our patients. All of these major developments must be analyzed objectively, taking into account costs to the health systems, risks and benefits to the patients, and the legal background that comes with them. A critical analysis of these new technologies and pharmacological breakthroughs should be made before considering changing our clinical practice."	"Urology is a constantly changing specialty with a wide range of therapeutic breakthroughs, a huge understanding of the genomic expression profiles for each urological cancer and a tendency to use cutting-edge technology to treat our patients. All of these major developments must be analyzed objectively, taking into account costs to the health systems, risks and benefits to the patients, and the legal background that comes with them. A critical analysis of these new technologies and pharmacological breakthroughs should be made before considering changing our clinical practice."
257	Chierici, M and Bussola, N and Marcolini, A and Francescatto, M and Zandonà, A and Trastulla, L and Agostinelli, C and Jurman, G and Furlanello, C	Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling	Frontiers in Oncology	10				2020	Italy	http://dx.doi.org/10.3389/fonc.2020.01065	article	"Gene expression, protein expression and copy number variants are used to predict estrogen receptor status (BRCA-ER, N = 381) and breast invasive carcinoma subtypes (BRCA-subtypes, N = 305), while gene expression, miRNA expression and methylation data is used as predictor layers for acute myeloid leukemia and renal clear cell carcinoma survival (AML-OS, N = 157; KIRC-OS, N = 181)"	Disease status, subtype and survival prediction	Matthews Correlation Coefficient (10 × stratified Monte Carlo cross-validation (50% training/validation proportion))	cross-validation + external cohort validation	"Although it is clear that no single method is consistently preferable, and that most of the proposed approaches are task and/or data dependent, the complexity of tumor analysis suggests that network-based approaches are needed [...] In this context, it is clear that omics-integration is one of the most promising and demanding challenges of the modern bioinformatics, and that there is an urgent need to prove the reproducibility, interpretability, and generalization capability of the proposed methods"	"Although it is clear that no single method is consistently preferable, and that most of the proposed approaches are task and/or data dependent, the complexity of tumor analysis suggests that network-based approaches are needed [...] In this context, it is clear that omics-integration is one of the most promising and demanding challenges of the modern bioinformatics, and that there is an urgent need to prove the reproducibility, interpretability, and generalization capability of the proposed methods"
258	Czolk, R and Klueber, J and Sørensen, M and Wilmes, P and Codreanu-Morel, F and Skov, P S and Hilger, C and Bindslev-Jensen, C and Ollert, M and Kuehn, A	IgE-Mediated Peanut Allergy: Current and Novel Predictive Biomarkers for Clinical Phenotypes Using Multi-Omics Approaches	Front Immunol	11			594350-594350	2020	Luxembourg	http://dx.doi.org/10.3389/fimmu.2020.594350	article	review (not applicable)	Review			"The immune mechanism underlying acute food-allergic events remains elusive until today. Deciphering this immunological response shall enable to identify novel biomarker for stratification of patients into reaction endotypes. The availability of powerful multi-omics technologies, together with integrated data analysis, network-based approaches and unbiased machine learning holds out the prospect of providing clinically useful biomarkers or biomarker signatures being predictive for reaction phenotypes."	"The immune mechanism underlying acute food-allergic events remains elusive until today. Deciphering this immunological response shall enable to identify novel biomarker for stratification of patients into reaction endotypes. The availability of powerful multi-omics technologies, together with integrated data analysis, network-based approaches and unbiased machine learning holds out the prospect of providing clinically useful biomarkers or biomarker signatures being predictive for reaction phenotypes."
259	de Oliveira Lima, E and Navarro, L C and Morishita, K N and Kamikawa, C M and Rodrigues, R G M and Dabaja, M Z and de Oliveira, D N and Delafiori, J and Dias-Audibert, F L and da Silva Ribeiro, M and Vicentini, A P and Rocha, A and Catharino, R R	Metabolomics and machine learning approaches combined in pursuit for more accurate paracoccidioidomycosis diagnoses	mSystems	5	3			2020	Brazil	http://dx.doi.org/10.1128/mSystems.00258-20	article	"In total, 343 individuals were included in this study, regardless of age and gender, in two main groups: the test group, consisting of PCM patients (n = 85), and the control group (n = 258)"	Case-control study	accuracy, sensitivity, specificity (patients' samples were randomly split into fit partition (Pfit) and test partition (Ptest) in the proportion of 80% and 20%, respectively. Classifiers were trained and validated in all steps of the method using 10 experiments of Pfit randomly shuffled and divided into training partition (Ptrain) and validation partition (Pval) in the proportions of 80% and 20%, respectively)	cross-validation	"Paracoccidioidomycosis (PCM) is a fungal infection typically found in Latin American countries, especially in Brazil. The identification of this disease is based on techniques that may fail sometimes. Intending to improve PCM detection in patient samples, this study used the combination of two of the newest technologies, artificial intelligence and metabolomics. This combination allowed PCM detection, independently of disease form, through identification of a set of molecules present in patients' blood. The great difference in this research was the ability to detect disease with better confidence than the routine methods employed today."	"Paracoccidioidomycosis (PCM) is a fungal infection typically found in Latin American countries, especially in Brazil. The identification of this disease is based on techniques that may fail sometimes. Intending to improve PCM detection in patient samples, this study used the combination of two of the newest technologies, artificial intelligence and metabolomics. This combination allowed PCM detection, independently of disease form, through identification of a set of molecules present in patients' blood. The great difference in this research was the ability to detect disease with better confidence than the routine methods employed today."
260	Eddy, S and Mariani, L H and Kretzler, M	Integrated multi-omics approaches to improve classification of chronic kidney disease	Nat Rev Nephrol	16		11	657-668	2020	USA	http://dx.doi.org/10.1038/s41581-020-0286-5	article	review (not applicable)	Review			"Technological advances now enable large-scale datasets, including DNA and RNA sequence data, proteomics and metabolomics data, to be captured from individuals and groups of patients along the genotype–phenotype continuum of chronic kidney disease (CKD). The ability to combine these high-dimensional datasets, in which the number of variables exceeds the number of clinical outcome observations, using computational approaches such as machine learning, provides an opportunity to re-classify patients into molecularly defined subgroups that better reflect underlying disease mechanisms. Patients with CKD are uniquely poised to benefit from these integrative, multi-omics approaches since the kidney biopsy, blood and urine samples used to generate these different types of molecular data are frequently obtained during routine clinical care."	"Technological advances now enable large-scale datasets, including DNA and RNA sequence data, proteomics and metabolomics data, to be captured from individuals and groups of patients along the genotype–phenotype continuum of chronic kidney disease (CKD). The ability to combine these high-dimensional datasets, in which the number of variables exceeds the number of clinical outcome observations, using computational approaches such as machine learning, provides an opportunity to re-classify patients into molecularly defined subgroups that better reflect underlying disease mechanisms. Patients with CKD are uniquely poised to benefit from these integrative, multi-omics approaches since the kidney biopsy, blood and urine samples used to generate these different types of molecular data are frequently obtained during routine clinical care."
261	Fu, S and Zarrinpar, A	Recent advances in precision medicine for individualized immunosuppression	Curr Opin Organ Transplant	25	4		420-425	2020	USA	http://dx.doi.org/10.1097/mot.0000000000000771	article	review (not applicable)	Review			"Newly proposed biomarkers offer precise and noninvasive ways to monitor patient's status. Cell-free DNA quantitation is increasingly explored as an indicator of allograft injury and rejection, which can help avoid unneeded biopsies and more frequently monitor graft function."	"Newly proposed biomarkers offer precise and noninvasive ways to monitor patient's status. Cell-free DNA quantitation is increasingly explored as an indicator of allograft injury and rejection, which can help avoid unneeded biopsies and more frequently monitor graft function."
262	Gacesa, R and Vich Vila, A and Collij, V and Imhann, F and Wijmenga, C and Jonkers, D M A E and Kurilshikov, A and Fu, J and Zhernakova, A and Weersma, R	Microbiome and fecal biomarkers can diagnose and classify inflammatory bowel disease	United European Gastroenterology Journal	7	8		166-167	2019	Netherlands	https://onlinelibrary.wiley.com/doi/10.1177/2050640619854670	meeting abstract	"We used whole metagenome sequencing to analyse composition and function of microbiome of fecal samples of 181 IBS patient, 380 IBD cases and 859 healthy controls"	Case-control study	AUC, sensitivity, specificity (independent set validation)	training + test set	"we show that features of gut microbiome, in combination with already used fecal biomarkers, are strong predictors for differentiating IBD and IBS, with additional potential of classifying location and type of IBD. These results have a potential to improve non-invasive pre-screening for IBD in clinical practice."	"we show that features of gut microbiome, in combination with already used fecal biomarkers, are strong predictors for differentiating IBD and IBS, with additional potential of classifying location and type of IBD. These results have a potential to improve non-invasive pre-screening for IBD in clinical practice."
263	Garcia, S and Lauritsen, J and Zhang, Z and Dalgaard, M D and Nielsen, R L and Daugaard, G and Gupta, R	Prediction of nephrotoxicity associated with cisplatinbased chemotherapy in testicular cancer patients	JNCI Cancer Spectrum		4	3		2020	Denmark	https://doi.org/10.1093/jncics/pkaa032	meeting abstract	"Of 433 patients assessed in this study, 26.8% developed nephrotoxicity after bleomycin-etoposide-cisplatin treatment."	Case-control study	AUC ("Training and testing of the algorithm was performed with a 5 outer, 2 inner fold nested cross-validation")	cross-validation + external cohort validation	"In this study, we were able to predict patients at risk of developing nephrotoxicity after BEP chemotherapy based on clinical and genetic features with a machine learning algorithm. Clinical features selected on the random forests–driven baseline clinical model were known risk factors of renal toxicity and were statistically significant in univariate analysis."	"In this study, we were able to predict patients at risk of developing nephrotoxicity after BEP chemotherapy based on clinical and genetic features with a machine learning algorithm. Clinical features selected on the random forests–driven baseline clinical model were known risk factors of renal toxicity and were statistically significant in univariate analysis."

264	Gindin, Y and Chuang, J C and Billin, A and Camargo, M and Huss, R and Chung, C and Myers, R P and Younossi, Z M and Harrison, S A and Anstee, Q M and Loomba, R	A random forest classifier based on a 30-gene signature distinguishes patients with bridging fibrosis from those with cirrhosis due to NASH	Hepatology	72	1	43A-44A	2020	USA	http://dx.doi.org/10.1002/hep.31578	meeting abstract	"The study included 1,120 adults with advanced fibrosis (F3-F4) due to NASH enrolled in the simtuzumab and STELLAR trials (discovery cohort, n=994) and the ATLAS trial (validation cohort, n=126)"	Case-control study	AUC (training + validation cohort)	training + test set	"A machine learning technique applied to hepatic transcriptomic data identified a 30-gene expression signature that accurately differentiates NASH patients with cirrhosis from those with bridging fibrosis. The functional activities of these genes may suggest novel drivers of fibrosis progression in NASH."	"A machine learning technique applied to hepatic transcriptomic data identified a 30-gene expression signature that accurately differentiates NASH patients with cirrhosis from those with bridging fibrosis. The functional activities of these genes may suggest novel drivers of fibrosis progression in NASH."
265	Goswami, C and Chawla, S and Thakral, D and Pant, H and Verma, P and Malik, P S and Jayadeva and Gupta, R and Ahuja, G and Sengupta, D	Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC	BMC Genomics	21	1	744-744	2020	India	https://doi.org/10.1186/s12864-020-07147-z	article	"we obtained 273 TEP expression profiles spanning six cancer types: non-small-cell lung cancer (NSCLC): 59, colorectal cancer (CRC): 44, glioblastoma multiforme (GBM): 40, breast cancer (BRCA): 38, pancreatic cancer (PC): 33, hepatobiliary cancer (HBC): 5. In addition to the cancer samples, platelets from 54 healthy individuals were also profiled."	Case-control study	AUC (LOOCV)	cross-validation	"In this article, we demonstrated the predictive power of a small set of platelet genes in determining the existence of cancer. Similar strategies can be developed for inferring the potential cancer types. In all these cases, the gene panels need to be validated on larger patient and control samples' cohorts. An orthogonal application of such panels could be tracking the treatment responses, as well as the recurrence of the disease."	"In this article, we demonstrated the predictive power of a small set of platelet genes in determining the existence of cancer. Similar strategies can be developed for inferring the potential cancer types. In all these cases, the gene panels need to be validated on larger patient and control samples' cohorts. An orthogonal application of such panels could be tracking the treatment responses, as well as the recurrence of the disease."
266	Graham, S A and Lee, E E and Jeste, D V and Van Patten, R and Twamley, E W and Nebeker, C and Yamada, Y and Kim, H C and Depp, C A	Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review	Psychiatry Research	284			2020	USA	http://dx.doi.org/10.1016/j.psychres.2019.112732	article	review (not applicable)	Review			"Artificial Intelligence has great potential to advance diagnosis and treatment of patients with neurocognitive disorders. Multi-feature datasets can improve personalization and predictive ability of machine learning algorithms in healthcare. Development of Explainable Artificial Intelligence is warranted to establish trust in models for clinical decision-making."	"Artificial Intelligence has great potential to advance diagnosis and treatment of patients with neurocognitive disorders. Multi-feature datasets can improve personalization and predictive ability of machine learning algorithms in healthcare. Development of Explainable Artificial Intelligence is warranted to establish trust in models for clinical decision-making."
267	Gumaei, A and Sammouda, R and Al-Rakhami, M and AlSalman, H and El-Zaart, A	Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression	Health Informatics Journal	27	1		2021	Saudi Arabia	http://dx.doi.org/10.1177/1460458221989402	article	"The experiment of this study is conducted on a public dataset of microarray prostate cancer gene expression, consisting of 102 tissue samples (52 prostate tumor and 50 normal tissues)"	Case-control study	accuracy (10-fold CV)	cross-validation	"In this paper, we propose to use a correlation feature selection (CFS) method with random committee (RC) ensemble learning to detect prostate cancer from microarray data of gene expression. A set of experiments are conducted on a public benchmark dataset using 10-fold cross-validation technique to evaluate the proposed approach. The experimental results revealed that the proposed approach attains 95.098% accuracy, which is higher than related work methods on the same dataset."	"In this paper, we propose to use a correlation feature selection (CFS) method with random committee (RC) ensemble learning to detect prostate cancer from microarray data of gene expression. A set of experiments are conducted on a public benchmark dataset using 10-fold cross-validation technique to evaluate the proposed approach. The experimental results revealed that the proposed approach attains 95.098% accuracy, which is higher than related work methods on the same dataset."
268	Guo, L Y and Wu, A H and Wang, Y X and Zhang, L P and Chai, H and Liang, X F	Deep learning-based ovarian cancer subtypes identification using multi-omics data	BioData Mining	13	1		2020	China	http://dx.doi.org/10.1186/s13040-020-00222-x	article	"The R package TCGA-assemble2 [13] was used for data collection and we obtained 298 samples concluded three types of omics data: mRNA-seq data (UNC Illumina HiSeq_RNASeq V2), miRNA-seq data (BCGSC Illumina HiSeq) and copy number variation (CNV) data (BROAD-MIT Genome wide SNP_6)"	Tumor stratification	silhouette score (external test datasets)	training + test set	"We identified 34 biomarkers and 19 KEGG pathways associated with ovarian cancer. The independent test results in three GEO datasets proved the robustness of our model. The literature reviewing show 19 (56%) biomarkers and 8 (42.1%) KEGG pathways identified based on the classification ubtypes have been proved to be associated with ovarian cancer.	"We identified 34 biomarkers and 19 KEGG pathways associated with ovarian cancer. The independent test results in three GEO datasets proved the robustness of our model. The literature reviewing show 19 (56%) biomarkers and 8 (42.1%) KEGG pathways identified based on the classification ubtypes have been proved to be associated with ovarian cancer.
269	Hajirasouliha, I and Elemento, O	Precision medicine and artificial intelligence: overview and relevance to reproductive medicine	Fertil Steril	114	5	908-913	2020	USA	http://dx.doi.org/10.1016/j.fertnstert.2020.09.156	article	review (not applicable)	Review			"Precision medicine matches each individual to the best treatment in a way that is tailored to his or her genetic uniqueness. To further personalize medicine, precision medicine increasingly incorporates and integrates data beyond genomics, such as epigenomics and metabolomics, as well as imaging. Increasingly, the robust use and integration of these modalities in precision medicine require the use of artificial intelligence and machine learning. This modern view of precision medicine, adopted early in certain areas of medicine such as cancer, has started to impact the field of reproductive medicine."	"Precision medicine matches each individual to the best treatment in a way that is tailored to his or her genetic uniqueness. To further personalize medicine, precision medicine increasingly incorporates and integrates data beyond genomics, such as epigenomics and metabolomics, as well as imaging. Increasingly, the robust use and integration of these modalities in precision medicine require the use of artificial intelligence and machine learning. This modern view of precision medicine, adopted early in certain areas of medicine such as cancer, has started to impact the field of reproductive medicine."
270	Hao, S and Bai, J and Liu, H and Wang, L and Liu, T and Lin, C and Luo, X and Gao, J and Zhao, J and Li, H and Tang, H	Comparison of machine learning tools for the prediction of AMD based on genetic, age, and diabetes-related variables in the Chinese population	Regenerative Therapy	15		180-186	2020	China	http://dx.doi.org/10.1016/j.reth.2020.09.001	article	"This study included a total of 202 subjects, comprising 82 AMD patients and 120 control subjects."	Case-control study	AUC (4-fold CV)	cross-validation	"In summary, we constructed and evaluated AMD [Age-related macular degeneration] prediction models integrating 3 SNPs and 2 clinical factors. The four models showed AUROCs above 0.72 in the training set. Machine learning tools have the potential to aid in the early diagnosis and treatment of patients with AMD. There is still a way to go before the models can be applied in the clinic for AMD prediction, and they should be validated in a larger cohort."	"In summary, we constructed and evaluated AMD [Age-related macular degeneration] prediction models integrating 3 SNPs and 2 clinical factors. The four models showed AUROCs above 0.72 in the training set. Machine learning tools have the potential to aid in the early diagnosis and treatment of patients with AMD. There is still a way to go before the models can be applied in the clinic for AMD prediction, and they should be validated in a larger cohort."
271	He, Z and Zhang, J and Yuan, X and Zhang, Y	Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods	Frontiers in Genetics	11			2020	China	http://dx.doi.org/10.3389/fgene.2020.632901	article	" We finally obtained 488 primary breast tumors together with survival time, and all samples of them included all of the five aforementioned genomic data types. The details of our dataset are illustrated in Table 1."	Case-control study	AUC (entire datasets were randomly divided into a learning dataset (80% of the entire dataset) and validation dataset (20%))		"We integrated somatic mutations and previously used data types, including Exp, CNV, Methy, and protein, using MKL to predict breast cancer patient survival. Applying mRMR-selected features and MKL classification, we found that the integration of somatic mutations enriched the diversity of features and was conducive to the improvement of the prediction model. In all, integrating promising data sources such as somatic mutations and harnessing the powerful feature selection method mRMR and the effective data fusion method MKL can increase the prediction accuracy of breast cancer patient survival."	"We integrated somatic mutations and previously used data types, including Exp, CNV, Methy, and protein, using MKL to predict breast cancer patient survival. Applying mRMR-selected features and MKL classification, we found that the integration of somatic mutations enriched the diversity of features and was conducive to the improvement of the prediction model. In all, integrating promising data sources such as somatic mutations and harnessing the powerful feature selection method mRMR and the effective data fusion method MKL can increase the prediction accuracy of breast cancer patient survival."
272	Hong, S and Su, Z and Li, J and Yu, S and Lin, B and Ke, Z and Zhang, Q and Guo, Z and Lv, W and Peng, S and Cheng, L and He, Q and Liu, R and Xiao, H	Development of circulating free DNA methylation markers for thyroid nodule diagnostics	Annals of Oncology	31		S1362-S1362	2020	China	http://dx.doi.org/10.1016/j.annonc.2020.10.301	meeting abstract	"Candidate markers were developed into a targeted sequencing panel and were validated on plasma DNA samples (115 PTC, 102 BTN)"	Case-control study	accuracy, sensitivity, specificity (training/test set)	training + test set+extcohort	"Our study demonstrates that DNA methylation markers can robustly differentiate thyroid nodules based on their malignancy. They are thus promising candidates to develop non-invasive diagnostics for thyroid cancer screening"	"Our study demonstrates that DNA methylation markers can robustly differentiate thyroid nodules based on their malignancy. They are thus promising candidates to develop non-invasive diagnostics for thyroid cancer screening"
273	Hoshino, A and Kim, H S and Bojmar, L and Gyan, K E and Cioffi, M and Hernandez, J and Zambirinis, C P and Rodrigues, G and Molina, H and Heissel, S and Mark, M T and Steiner, L and Benito-Martin, A and Lucotti, S and Di Giannatale, A and Offer, K and Nakajima, M and Williams, C and Nogués, L and Pelissier Vatter, F A and Hashimoto, A and Davies, A E and Freitas, D and Kenific, C M and Ararso, Y and Buehring, W and Lauritzen, P and Ogitani, Y and Sugiura, K and Takahashi, N and Alečković, M and Bailey, K A and Jolissant, J S and Wang, H and Harris, A and Schaeffer, L M and García-Santos, G and Posner, Z and Balachandran, V P and Khakoo, Y and Raju, G P and Scherz, A and Sagi, I and Scherz-Shouval, R and Yarden, Y and Oren, M and Malladi, M and Petriccione, M and De Braganca, K C and Donzelli, M and Fischer, C and Vitolano, S and Wright, G P and Ganshaw, L and Marrano, M and Ahmed, A and DeStefano, J and Danzer, E and Roehrl, M H A and Lacayo, N J and Vincent, T C and Weiser, M R and Brady, M S and Meyers, P A and Wexler, L H and Ambati, S R and Chou, A J and Slotkin, E K and Modak, S and Roberts, S S and Basu, E M and Diolaiti, D and Krantz, B A and Cardoso, F and Simpson, A L and Berger, M and Rudin, C M and Simeone, D M and Jain, M and Ghajar, C M and Batra, S K and Stanger, B Z and Bui, J and Brown, K A and Rajasekhar, V K and Healey, J H and de Sousa, M and Kramer, K and Sheth, S and Baisch, J and Pascual, V and Heaton, T E and La Quaglia, M P and Pisapia, D J and Schwartz, R and Zhang, H and Liu, Y and Shukla, A and Blavier, L and DeClerck, Y A and LaBarge, M and Bissell, M J and Caffrey, T C and Grandgenett, P M and Hollingsworth, M A and Bromberg, J and Costa-Silva, B and Peinado, H and Kang, Y and Garcia, B A and O'Reilly, E M and Kelsen, D and Trippett, T M and Jones, D R and Matei, I R and Jarnagin, W R and Lyden, D	Extracellular Vesicle and Particle Biomarkers Define Multiple Human Cancers	Cell	182	4	1044-1061.e18	2020	Japan	http://dx.doi.org/10.1016/j.cell.2020.07.009	article	"To confirm that EVPs are ideal diagnostic tools, we analyzed proteomes of TE- (n = 151) and plasma-derived (n = 120) EVPs"	Case-control study	sensitivity, specificity (10-fold CV + external test set)	cross-validation + external cohort validation	"Machine-learning classification of plasma-derived EVP [extracellular vesicles and particles] cargo, including immunoglobulins, revealed 95% sensitivity/90% specificity in detecting cancer. Finally, we defined a panel of tumor-type-specific EVP proteins in TEs and plasma, which can classify tumors of unknown primary origin. Thus, EVP proteins can serve as reliable biomarkers for cancer detection and determining cancer type."	"Machine-learning classification of plasma-derived EVP [extracellular vesicles and particles] cargo, including immunoglobulins, revealed 95% sensitivity/90% specificity in detecting cancer. Finally, we defined a panel of tumor-type-specific EVP proteins in TEs and plasma, which can classify tumors of unknown primary origin. Thus, EVP proteins can serve as reliable biomarkers for cancer detection and determining cancer type."
274	Huang, J and Huth, C and Covic, M and Troll, M and Adam, J and Zukunft, S and Prehn, C and Wang, L and Nano, J and Scheerer, M F and Neschen, S and Kastenmüller, G and Suhre, K and Laxy, M and Schliess, F and Gieger, C and Adamski, J and Hrabe de Angelis, M and Peters, A and Wang-Sattler, R	Machine Learning Approaches Reveal Metabolic Signatures of Incident Chronic Kidney Disease in Individuals With Prediabetes and Type 2 Diabetes	Diabetes	69	12	2756-2765	2020	Germany	http://dx.doi.org/10.2337/db20-0586	article	"We investigated the two follow-ups of the longitudinal cohort KORA survey 4, conducted in the area of Augsburg, Southern Germany. The first follow-up (F4) involved 3,080 individuals (aged 32–81 years) examined between 2006 and 2008. For the second follow-up (FF4), 2,269 participants were examined from 2013 to 2014"	Case-control study	AUC (three-step feature selection with 100 random repeats of 10-fold cross validation)	cross-validation	"This longitudinal study revealed significant accumulation of sphingo- and glycerophospholipids (SM C18:1 and PC aa C38:0) in individuals with prediabetes and T2D up to 6.5 years before their clinical onset of CKD. These candidate metabolite biomarkers of incident CKD were specific for hyperglycemic state, i.e., individuals with increased fasting and/or 2-h glucose levels. Highly stable performances of the sets of predictors for incident CKD developed from 125 metabolites and 14 clinical variables were furthermore independently confirmed with three machine learning algorithms."	"This longitudinal study revealed significant accumulation of sphingo- and glycerophospholipids (SM C18:1 and PC aa C38:0) in individuals with prediabetes and T2D up to 6.5 years before their clinical onset of CKD. These candidate metabolite biomarkers of incident CKD were specific for hyperglycemic state, i.e., individuals with increased fasting and/or 2-h glucose levels. Highly stable performances of the sets of predictors for incident CKD developed from 125 metabolites and 14 clinical variables were furthermore independently confirmed with three machine learning algorithms."
275	Huang, Z and Johnson, T S and Han, Z and Helm, B and Cao, S and Zhang, C and Salama, P and Rizkalla, M and Yu, C Y and Cheng, J and Xiang, S and Zhan, X and Zhang, J and Huang, K	Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations	BMC Med Genomics	13		41-41	2020	USA	http://dx.doi.org/10.1186/s12920-020-0686-1	article	"The performance comparison was conducted at pan-cancer level using 12 cancer from The Cancer Genome Atlas (TCGA). These 12 cancers were chosen due to their relatively large sample sizes and sufficient information about patient outcomes. The specific cancers analyzed in this paper were (1) Urothelial Bladder Carcinoma (BLCA); (2) Breast Invasive Carcinoma (BRCA); (3) Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC); (4) Head-Neck Squamous Cell Carcinoma (HNSC); (5) Kidney Renal Clear Cell Carcinoma (KIRC); (6) Kidney Renal Papillary Cell Carcinoma (KIRP); (7) Liver Hepatocellular Carcinoma (LIHC); (8) Lung Adenocarcinoma (LUAD); (9) Lung Squamous Cell Carcinoma (LUSC); (10) Ovarian Cancer (OV); (11) Pancreatic Adenocarcinoma (PAAD); and (12) Stomach Adenocarcinoma (STAD). In this paper, we used the expression data of Illumina Hi-Seq RNA-seq v2 RSEM normalized genes from TCGA."	Cancer survival prognosis	C-index, p-value of log-rank test (Each dataset was split into training, validation, and testing sets in a proportion of 60, 20, and 20% respectively)	training + test set	"Overall our study demonstrated that the Deep Learning architecture can be effectively applied for cancer prognosis prediction with Cox-proportional hazard model incorporated. We found that Deep Learning-based model demonstrated superior performances comparing to traditional machine learning models. Among the three Deep Learning-based models tested, we observed that Cox-nnet, which has the most succinct neural network structure, resulted in better prognosis performances in the measurement of concordance index and p-value of log-rank test. We showed that integrating autoencoder with Cox regression network does not significantly improve the prognosis performances."	"Overall our study demonstrated that the Deep Learning architecture can be effectively applied for cancer prognosis prediction with Cox-proportional hazard model incorporated. We found that Deep Learning-based model demonstrated superior performances comparing to traditional machine learning models. Among the three Deep Learning-based models tested, we observed that Cox-nnet, which has the most succinct neural network structure, resulted in better prognosis performances in the measurement of concordance index and p-value of log-rank test. We showed that integrating autoencoder with Cox regression network does not significantly improve the prognosis performances."
276	Jiang, J and Yao, H and Yang, L and Li, J and Xin, J and Shi, D and Liang, X and Cai, Q and Ren, K and Chen, X and Li, J	Proteome predicts progression and prognosis of hepatitis B virus-related acute-on-chronic liver failure	Hepatology v70 suppl.1 2019	70		162A-163A	2019	China	https://aasldpubs.onlinelibrary.wiley.com/doi/10.1002/hep.30940	meeting abstract	"Serum of 520 subjects with HBVACLF (n=141), acute-on-chronic hepatic dysfunction (ACHD n=102), liver cirrhosis (LC n=110), chronic hepatitis B (CHB n=102), and normal controls (NC n=65) from a prospective multi-center cohort were subjected to a robust and highly streamlined single-run quantification proteomic analysis"	Cancer progression and prognosis prediction	AUC (training + validation cohort)	external cohort validation	"Quantitative proteomic study reveals an overall path of HBV-ACLF [Hepatitis B Virus-Related Acute-on-Chronic Liver Failure] disease progression. And the combinatorial predict model provides fundamental information of multiple biomarkers response the disease progression, severity and prognosis."	"Quantitative proteomic study reveals an overall path of HBV-ACLF [Hepatitis B Virus-Related Acute-on-Chronic Liver Failure] disease progression. And the combinatorial predict model provides fundamental information of multiple biomarkers response the disease progression, severity and prognosis."
277	Jovčevska, I	Next Generation Sequencing and Machine Learning Technologies Are Painting the Epigenetic Portrait of Glioblastoma	Frontiers in Oncology	10			2020	Slovenia	http://dx.doi.org/10.3389/fonc.2020.00798	article	review (not applicable)	Review			"Using a combination of phenotypic, genotypic, and epigenetic parameters in glioblastoma diagnostics will bring us closer to precision medicine where therapies will be tailored to suit the genetic profile and epigenetic signature of the tumor"	"Using a combination of phenotypic, genotypic, and epigenetic parameters in glioblastoma diagnostics will bring us closer to precision medicine where therapies will be tailored to suit the genetic profile and epigenetic signature of the tumor"

278	Kandimalla, R and Xu, J and Link, A and Matsuyama, T and Yamamura, K and Parker, I and Uetake, H and Hernandez-Illan, E and Lozano, J and Borazanci, E and Tsai, S and Evans, D and Meltzer, S J and Baba, H and Brand, R and Von Hoff, D and Balaguer, F and Li, W and Goel, A	EpiPanGI-Dx: A cell-free DNA methylation fingerprint for the early detection of gastrointestinal cancers	Cancer Research	80	16		2020	USA	http://dx.doi.org/10.1158/1538-7445.AM2020-1084	meeting abstract	"Using this approach, we sequenced 300 plasma specimens from all GI cancers, as well as age-matched healthy control"	Case-control study	AUC (training + validation cohort)	external cohort validation	"Utilizing a novel biomarker discovery approach, we provide first evidence for cell-free DNA methylation biomarkers that offer a robust diagnostic accuracy for the identification of specific cancer types, and demonstrate their potential clinical application as a Pan-cancer panel for the early detection of all gastrointestinal cancers."	"Utilizing a novel biomarker discovery approach, we provide first evidence for cell-free DNA methylation biomarkers that offer a robust diagnostic accuracy for the identification of specific cancer types, and demonstrate their potential clinical application as a Pan-cancer panel for the early detection of all gastrointestinal cancers."
279	Kaur, H and Bhalla, S and Garg, D and Mehta, N and Raghava, G P S	Analysis and prediction of cholangiocarcinoma from transcriptomic profile of patients	Journal of Hepatology	73		S16-S17	2020	India	http://dx.doi.org/10.1016/S0168-8278(20)30693-6	meeting abstract	"Eight datasets containing a total of 350 CCA, 133 adjacentnon-tumorous and 90 HCC samples"	Case-control study	AUC, accuracy (training + validation set)	training + test set	"Prediction models developed based on three genes categorized CCA [Cholangiocarcinoma] with high precision. Thus, they can be further explored for their diagnostic and therapeutic potential for CCA."	"Prediction models developed based on three genes categorized CCA [Cholangiocarcinoma] with high precision. Thus, they can be further explored for their diagnostic and therapeutic potential for CCA."
280	Khoshnejat, M and Kavousi, K and Banaei-Moghaddam, A M and Moosavi-Movahedi, A A	Unraveling the molecular heterogeneity in type 2 diabetes: a potential subtype discovery followed by metabolic modeling	BMC Med Genomics	13	1	119-119	2020	Iran	http://dx.doi.org/10.21203/rs.2.20464/v2	article	"The dataset contains gene expression data from participants with glucose tolerance ranging from normal to newly diagnosed T2DM, in which 91 and 63 individuals were healthy and diabetic, respectively"	Case-control study	AUC, accuracy, F1 score, precision, recall (10-fold CV)	cross-validation	"Using only gene expression data, it is possible to discriminate T2DM individuals from healthy controls with approximately 90 percent accuracy. Clustering of diabetic patients according to their gene expression patterns and subsequent more in-depth analysis of each cluster unraveled specific abnormalities leading to insulin resistance in each cluster."	"Using only gene expression data, it is possible to discriminate T2DM individuals from healthy controls with approximately 90 percent accuracy. Clustering of diabetic patients according to their gene expression patterns and subsequent more in-depth analysis of each cluster unraveled specific abnormalities leading to insulin resistance in each cluster."
281	Kong, J and Lee, H and Kim, D and Han, S K and Ha, D and Shin, K and Kim, S	Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients	Nat Commun	11	1	5485-5485	2020	Republic of Korea	http://dx.doi.org/10.1038/s41467-020-19313-8	article	"We downloaded the FPKM-UQ (upper quartile) dataset from TCGA data portal for expression analysis"	Case-control study	"The final predictive performance was measured by comparing the correlation between the observed and predicted drug responses in the test set (R2)" ("split the organoid dataset into training (60%), validation (10%), and test (30%) sets", 3-fold CV for training)	training + test set	"In this study, we tested if the incorporation of network analysis into an ML framework could accurately identify robust drug-response biomarkers using organoid models. Indeed, we found that our method accurately predicted cancer patient–drug responses, whereas conventional ML approaches showed less optimal predictive performances. Importantly, our network-based ML model provided interpretable results for drug-response prediction, which were further tested in external experimental datasets."	"In this study, we tested if the incorporation of network analysis into an ML framework could accurately identify robust drug-response biomarkers using organoid models. Indeed, we found that our method accurately predicted cancer patient–drug responses, whereas conventional ML approaches showed less optimal predictive performances. Importantly, our network-based ML model provided interpretable results for drug-response prediction, which were further tested in external experimental datasets."
282	Koras, K and Juraeva, D and Kreis, J and Mazur, J and Staub, E and Szczurek, E	Feature selection strategies for drug sensitivity prediction	Sci Rep	10	1	9377-9377	2020	Poland	http://dx.doi.org/10.1038/s41598-020-65927-9	article	"The total set of samples consisted of 983 cancer cell lines originated from 13 tissue sites"	Drug response prediction	Correlation, RMSE (3-fold CV on training data + test set evaluation)	training + test set	"For many compounds, even a very small subset of drug-related features is highly predictive of drug sensitivity. Small feature sets selected using prior knowledge are more predictive for drugs targeting specific genes and pathways, while models with wider feature sets perform better for drugs affecting general cellular mechanisms. Appropriate feature selection strategies facilitate the development of interpretable models that are indicative for therapy design."	"For many compounds, even a very small subset of drug-related features is highly predictive of drug sensitivity. Small feature sets selected using prior knowledge are more predictive for drugs targeting specific genes and pathways, while models with wider feature sets perform better for drugs affecting general cellular mechanisms. Appropriate feature selection strategies facilitate the development of interpretable models that are indicative for therapy design."
283	Koureas, M and Kirgou, P and Amoutzias, G and Hadjichristodoulou, C and Gourgoulianis, K and Tsakalof, A	Target analysis of volatile organic compounds in exhaled breath for lung cancer discrimination from other pulmonary diseases and healthy persons	Metabolites	10	8	1-18	2020	Greece	http://dx.doi.org/10.3390/metabo10080317	article	"The population sample consisted of 51 patients with confirmed LC, 38 patients with pathological computed tomography (CT) findings not diagnosed with LC, and 53 healthy controls"	Case-control study	AUC (10-fold CV + validation)	cross-validation + test set	"The random forest machine learning algorithm achieved a correct classification of patients of 88.5% (area under the curve—AUC 0.94). However, none of the methods used achieved adequate discrimination between LC patients and patients with abnormal computed tomography (CT) findings. Biomarker sets, consisting mainly of the exogenous monoaromatic compounds and 1- and 2- propanol, adequately discriminated LC patients from healthy controls."	"The random forest machine learning algorithm achieved a correct classification of patients of 88.5% (area under the curve—AUC 0.94). However, none of the methods used achieved adequate discrimination between LC patients and patients with abnormal computed tomography (CT) findings. Biomarker sets, consisting mainly of the exogenous monoaromatic compounds and 1- and 2- propanol, adequately discriminated LC patients from healthy controls."
284	Lai, Y H and Chen, W N and Hsu, T C and Lin, C and Tsao, Y and Wu, S	Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning	Sci Rep	10	1	4679-4679	2020	Taiwan	http://dx.doi.org/10.1038/s41598-020-61588-w	article	" We separated 256 patients as the training set, 85 patients as the validation set, and 171 patients as the test set"	Case-control study	AUC, accuracy (10-fold CV + validation set)	training + test set	"In this study, we applied the concept of bimodal learning to construct an integrative DNN where two heterogeneous modalities (gene expression and clinical data) were integrated for predicting ADC patient overall survival. By using two modalities, the integrative DNN approach is capable of providing the missing information left by the other observed modality. Compared with our microarray DNN, we observed an increase in AUC and accuracy from the integrative DNN."	"In this study, we applied the concept of bimodal learning to construct an integrative DNN where two heterogeneous modalities (gene expression and clinical data) were integrated for predicting ADC patient overall survival. By using two modalities, the integrative DNN approach is capable of providing the missing information left by the other observed modality. Compared with our microarray DNN, we observed an increase in AUC and accuracy from the integrative DNN."
285	Lee, S and Deasy, J O and Oh, J H and Di Meglio, A and Dumas, A and Menvielle, G and Charles, C and Boyault, S and Rousseau, M and Besse, C and Thomas, E and Boland, A and Cottu, P and Tredan, O and Levy, C and Martin, A L and Everhard, S and Ganz, P A and Partridge, A H and Michiels, S and Deleuze, J F and Andre, F and Vaz-Luis, I	Prediction of breast cancer treatment–induced fatigue by machine learning using genome-wide association data	JNCI Cancer Spectrum	4	5		2020	USA	http://dx.doi.org/10.1093/JNCICS/PKAA039	article	"We accessed germline genome-wide data of 2799 early-stage breast cancer patients from the Cancer Toxicity study (NCT01993498)"	Treatment response prediction	AUC (training + test set)	training + test set	"This study highlighted the potential of machine learning-aided deep methylation sequencing as a sensitive ctDNA profiling approach for early cancer detection. Further investigation in large-scale clinical studies is ongoing."	"This study highlighted the potential of machine learning-aided deep methylation sequencing as a sensitive ctDNA profiling approach for early cancer detection. Further investigation in large-scale clinical studies is ongoing."
286	Li, B and Wang, C and Xu, J and Fang, S and Qiu, F and Su, J and Chu, H and Han-Zhang, H and Mao, X and Liu, H and Liu, X and Zhang, W and Zhao, H and Zhang, Z	Multiplatform analysis of early-stage cancer signatures in blood	Clinical Cancer Research	26		11	2020	China	http://dx.doi.org/10.1158/1557-3265.LiqBiop20-A06	meeting abstract	"The study was conducted among 452 surgery-resectable patients with lung cancer (N=180), colorectal cancer (N=210), liver cancer (N=62), and 290 age-/sex- matched non-cancer controls"	Case-control study	AUC (training + test set)	training + test set	"In conclusion, we proposed a boosting ensemble predictive framework with the wrapper-based feature selection algorithm for predicting antidepressant treatment response and remission in Taiwanese patients with MDD. The present results suggest that our boosting ensemble predictive framework with the wrapper-based feature selection algorithm may leverage a feasible way to create predictive algorithms for forecasting antidepressant treatment response and remission with clinically meaningful accuracy."	"In conclusion, we proposed a boosting ensemble predictive framework with the wrapper-based feature selection algorithm for predicting antidepressant treatment response and remission in Taiwanese patients with MDD. The present results suggest that our boosting ensemble predictive framework with the wrapper-based feature selection algorithm may leverage a feasible way to create predictive algorithms for forecasting antidepressant treatment response and remission with clinically meaningful accuracy."
287	Lin, E and Kuo, P H and Liu, Y L and Yu, Y W Y and Yang, A C and Tsai, S J	Prediction of antidepressant treatment response and remission using an ensemble machine learning framework	Pharmaceuticals Experimental Hematology and Oncology	13		10 1-12	2020	USA	http://dx.doi.org/10.3390/ph13100305	article	"we retained 421 MDD patients for the subsequent analysis"	Case-control study	AUC (repeated 10-fold CV)	cross-validation	"Our work demonstrated 3 immune clusters with different features. More importantly, multi-omics signatures, such as MMP9 was identified based on three clusters to help us recognize patients with different prognosis and responses to immunotherapy in HCC."	"Our work demonstrated 3 immune clusters with different features. More importantly, multi-omics signatures, such as MMP9 was identified based on three clusters to help us recognize patients with different prognosis and responses to immunotherapy in HCC."
288	Liu, F and Qin, L and Liao, Z and Song, J and Yuan, C and Liu, Y and Wang, Y and Xu, H and Zhang, Q and Pei, Y and Zhang, H and Pan, Y and Chen, X and Zhang, Z and Zhang, W and Zhang, B	Microenvironment characterization and multi-omics signatures related to prognosis and immunotherapy response of hepatocellular carcinoma		9	1		2020	China	http://dx.doi.org/10.1186/s40164-020-00165-3	article	"We performed unsupervised clustering of total 1000 HCC [hepatocellular carcinoma] samples including discovery and validation group from available public datasets"	Prognostic subtype stratification	correlation (discovery + validation cohorts)	external cohort validation	"Our work demonstrated 3 immune clusters with different features. More importantly, multi-omics signatures, such as MMP9 was identified based on three clusters to help us recognize patients with different prognosis and responses to immunotherapy in HCC."	"Our work demonstrated 3 immune clusters with different features. More importantly, multi-omics signatures, such as MMP9 was identified based on three clusters to help us recognize patients with different prognosis and responses to immunotherapy in HCC."
289	Liu, P and Tian, W	Identification of DNA methylation patterns and biomarkers for clear-cell renal cell carcinoma by multi-omics data analysis	PeerJ	8			2020	China	http://dx.doi.org/10.7717/peerj.9654	article	"The discovery stage involved 160 pairs of ccRCC [clear-cell renal cell carcinoma] and matched normal tissues for investigation of DNAm and biomarkers as well as 318 cases of ccRCC including clinical signatures"	Case-control study	AUC (10-fold CV + validation cohort)	cross-validation + external cohort validation	"The present study provides a comprehensive analysis of ccRCC using multi-omics data. These findings indicated that multi-omics analysis could identify some novel epigenetic factors, which were the most important causes of advanced cancer and poor clinical prognosis."	"The present study provides a comprehensive analysis of ccRCC using multi-omics data. These findings indicated that multi-omics analysis could identify some novel epigenetic factors, which were the most important causes of advanced cancer and poor clinical prognosis."
290	Liu, X Y and Wu, S B and Zeng, W Q and Yuan, Z J and Xu, H B	LogSum + L(2) penalized logistic regression model for biomarker selection and cancer classification	Sci Rep	10	1	22125-22125	2020	China	http://dx.doi.org/10.1038/s41598-020-79233/tho-218026	article	"This data describes 20,501 genes in 806 different breast cancer samples. We retained only samples with complete information. After that, 85 TNBC and 460 non-TNBC were further divided into two groups: training (n= 327; 51 TNBC, 276 non-TNBC) and testing (n= 218; 34 TNBC, 184 non- TNBC) sets"	Case-control study	AUC (10-fold CV + validation)	cross-validation + test set	"The logsum method is a powerful method for feature selection. However, it is unable to use any previous biological structure information. To overcome this drawback, in this paper we first propose Logsum-Net regularization to integrate biological network knowledge. Then, we suggest the penalized logsum-net regularization logistic regression model (logsum-NL) for gene selection and cancer classification. For a real large dataset, the proposed method has achieved 89.66% (training) and 90.02% (testing) AUC performance which are, on average, 5.17% (training) and 4.49% (testing) better than mainstream methods"	"The logsum method is a powerful method for feature selection. However, it is unable to use any previous biological structure information. To overcome this drawback, in this paper we first propose Logsum-Net regularization to integrate biological network knowledge. Then, we suggest the penalized logsum-net regularization logistic regression model (logsum-NL) for gene selection and cancer classification. For a real large dataset, the proposed method has achieved 89.66% (training) and 90.02% (testing) AUC performance which are, on average, 5.17% (training) and 4.49% (testing) better than mainstream methods"
291	Liu, Y and Liu, F and Hu, X and He, J and Jiang, Y	Combining Genetic Mutation and Expression Profiles Identifies Novel Prognostic Biomarkers of Lung Adenocarcinoma	Clinical Medicine Insights: Oncology	14			2020	China	http://dx.doi.org/10.1177/1179554920902260	article	"272 samples of the TCGA LUAD cohort were selected according to the overall survival and were partitioned into the training set and testing set [75%/25%]"	Prognostic study	AUC (10-fold CV + validation)	cross-validation + test set	"The most significant contribution of this article is the integration of the genetic mutation and expression profiles to determine prognostic genes for LUAD patients. If genetic expression and mutation profiles are available, the pipeline of determining DEGs and DMGs in this article can be applied to other types of cancers."	"The most significant contribution of this article is the integration of the genetic mutation and expression profiles to determine prognostic genes for LUAD patients. If genetic expression and mutation profiles are available, the pipeline of determining DEGs and DMGs in this article can be applied to other types of cancers."
292	Lu, Y and Wu, S and Cui, C and Yu, M and Wang, S and Yue, Y and Liu, M and Sun, Z	Gene expression along with genomic copy number variation and mutational analysis were used to develop a 9-gene signature for estimating prognosis of coad	OncoTargets and Therapy	13		10393-10408	2020	China	http://dx.doi.org/10.2147/OTT.S255599	article	"The TCGA training set contains 226 samples and the test set contains 227 samples. As an external validation set, the GSE17538 data set contains a total of 244 samples, including 6 mouse samples, while among the 238 human samples, 38 samples recorded the survival status of NA, and finally used for follow-up analysis"	Prognostic study	AUC (training + test set + external validation set)	external cohort validation	"our research results indicate that the 9 gene prognostic signature is a reliable tool for predicting the OS [overall survival] of COAD [Colon adenocarcinoma] patients"	"In this study, the AUC of 9-gene signature screened by multi-omics in the training set and validation set for five years is more than 0.64, which is more effective in predicting the prognosis of patients."
293	Luca, B A and Moulton, V and Ellis, C and Edwards, D R and Campbell, C and Cooper, R A and Clark, J and Brewer, D S and Cooper, C S	A novel stratification framework for predicting outcome in patients with prostate cancer	Br J Cancer	122		10 1467-1476	2020	United Kingdom	http://dx.doi.org/10.1038/s41416-020-0799-5	article	"There were 1785 samples from primary malignant tissue, and 173 from normal tissue"	Prognostic subtype stratification	Correlation, log-rank p-value (hold-out validation)	training + test set	"We have confirmed a key prediction of the DESNT cancer model by demonstrating that the presence of a small proportion of the DESNT cancer signature confers poorer outcome. The proportion of DESNT signature can be considered a continuous variable, such that as DESNT cancer content increases, the outcome became worse. This observation led to the development of nomograms for estimating PSA failure at 3, 5 and 7 years following prostatectomy."	"We have confirmed a key prediction of the DESNT cancer model by demonstrating that the presence of a small proportion of the DESNT cancer signature confers poorer outcome. The proportion of DESNT signature can be considered a continuous variable, such that as DESNT cancer content increases, the outcome became worse. This observation led to the development of nomograms for estimating PSA failure at 3, 5 and 7 years following prostatectomy."
294	Luo, R and Song, J and Xiao, X and Xie, Z and Zhao, Z and Zhang, W and Miao, S and Tang, Y and Ran, L	Identifying CpG methylation signature as a promising biomarker for recurrence and immunotherapy in non-small-cell lung carcinoma	Aging (Albany NY)	12		14649-14 14676	2020	China	http://dx.doi.org/10.18632/aging.103517	article	"a total of 901 TCGA NSCLC samples were available using the Illumina Infinium HumanMethylation450 platform, including 827 tumor tissues and 74 non-tumor tissues"	Tumor recurrence and immunotherapy response prediction	AUC (training + external validation)	external cohort validation	"In this study, we initially identified 4 CpG biomarkers associated with recurrence of NSCLC. Base on TCGA NSCLC cohort comprised of lung adenocarcinomas (LUAD) and lung squamous cell carcinomas (LUSC), a promising DNAm-based risk score model predictive of relapse was constructed and then validated in the other 3 datasets."	"In this study, we initially identified 4 CpG biomarkers associated with recurrence of NSCLC. Base on TCGA NSCLC cohort comprised of lung adenocarcinomas (LUAD) and lung squamous cell carcinomas (LUSC), a promising DNAm-based risk score model predictive of relapse was constructed and then validated in the other 3 datasets."
295	Makarious, M and Iwaki, H and Blauwendraat, C and Leonard, H and Hashemi, S and Kim, J and Van Keuren-Jensen, K and Craig, D and Appelmans, E and Smolensky, L and Bookman, M and Singleton, A and Faghri, F and Nalls, M	Biomarker discovery in parkinson's disease using machine learning on public multi-OMIC datasets: A pilot study	Movement Disorders	35		5207-5207	2020	USA	http://dx.doi.org/10.1002/mds.28268	meeting abstract	"This included 872 samples that had sequenced genomes, clinical data, and ~50K normalized transcripts from RNA sequencing"	Case-control study	AUC (30% test samples after training on 70% of samples)	training + test set	"When assessing the performance in 30% of test samples after training on 70% of samples, multiple modalities implemented in the same predictive model performs best. By incorporating different modalities, we can develop more comprehensive predictive models to better understand the complex disease and identify better biomarkers."	"When assessing the performance in 30% of test samples after training on 70% of samples, multiple modalities implemented in the same predictive model performs best. By incorporating different modalities, we can develop more comprehensive predictive models to better understand the complex disease and identify better biomarkers."
296	Mantha, S and Dunbar, A and Bolton, K L and Devlin, S and Gorenshteyn, D and Donoghue, M and Arcila, M E and Soff, G A	Machine learning for prediction of cancer-associated venous thromboembolism	Blood	136		37-37	2020	USA	http://dx.doi.org/10.1182/blood-2020-138579	meeting abstract	"12,040 patients were included in the final analysis. There were 855 CAT events during the observation period"	Prognostic study	C-index (cross-validation)	cross-validation	"Machine learning is a promising approach in the search of more accurate and generalizable models for prediction of CAT [Cancer-Associated Venous Thromboembolism]. In the application described here, the use of random survival forests performed well without information about future chemotherapy administration. Additional work is needed to identify the optimal algorithm and covariates, including better delineation of which cancer genomic information should be retained."	"Machine learning is a promising approach in the search of more accurate and generalizable models for prediction of CAT [Cancer-Associated Venous Thromboembolism]. In the application described here, the use of random survival forests performed well without information about future chemotherapy administration. Additional work is needed to identify the optimal algorithm and covariates, including better delineation of which cancer genomic information should be retained."
297	Manzi, M and Palazzo, M and Knott, M E and Beauseroy, P and Yankilevich, P and Giménez, M I and Monge, M E	Coupled Mass-Spectrometry-Based Lipidomics Machine Learning Approach for Early Detection of Clear Cell Renal Cell Carcinoma	J Proteome Res	20	1	841-857	2021	Argentina	http://dx.doi.org/10.1021/acs.jproteome.0c00663	article	"patients with clear cell renal cell carcinoma (ccRCC) stages I, II, III, and IV (n = 112) and controls (n = 52)"	Case-control study	accuracy (training/test set)	training + test set	"Multivariate models based on support vector machines and the LASSO variable selection method yielded two discriminant lipid panels for ccRCC detection and early diagnosis. A 16-lipid panel allowed discriminating ccRCC patients from controls with 95.7% accuracy in a training set under cross-validation and 77.1% accuracy in an independent test set."	"Multivariate models based on support vector machines and the LASSO variable selection method yielded two discriminant lipid panels for ccRCC detection and early diagnosis. A 16-lipid panel allowed discriminating ccRCC patients from controls with 95.7% accuracy in a training set under cross-validation and 77.1% accuracy in an independent test set."
298	McCarthy, C P and Neumann, J T and Michelhaugh, S A and Ibrahim, N E and Gaggin, H K and Sörensen, N A and Schäefer, S and Zeller, T and Magaret, C A and Barnes, G and Rhyne, R F and Westermann, D and Januzzi Jr., J L	Derivation and External Validation of a High-Sensitivity Cardiac Troponin-Based Proteomic Model to Predict the Presence of Obstructive Coronary Artery Disease	J Am Heart Assoc	9		e017221-16 e017221	2020	USA	http://dx.doi.org/10.1161/jaha.120.017221	article	"In a derivation cohort of 636 patients referred for coronary angiography, predictors of ≥70% coronary stenosis were identified from 6 clinical variables and 109 biomarkers. The final model was first internally validated on a separate cohort (n=275) and then externally validated on a cohort of 241 patients"	Case-control study	AUC (train + test + external validation)	external cohort validation	"We have derived and externally validated a clinical/proteomic panel that can predict the presence of obstructive CAD [Coronary Artery Disease] with high accuracy. The score performs similarly well in the evaluation of acute chest pain in the ED (including patients who had MI neither ruled in nor ruled out) and in outpatients presenting for evaluation of stable angina including those with renal injury."	"We have derived and externally validated a clinical/proteomic panel that can predict the presence of obstructive CAD with high accuracy. The score performs similarly well in the evaluation of acute chest pain in the ED (including patients who had MI neither ruled in nor ruled out) and in outpatients presenting for evaluation of stable angina including those with renal injury."

299	Miao, R and Chen, H H and Dang, Q and Xia, L Y and Yang, Z Y and He, M F and Hao, Z F and Liang, Y	Beyond the limitation of targeted therapy: Improve the application of targeted drugs combining genomic data with machine learning	Pharmacol Res	159			104932-104932	2020	China	http://dx.doi.org/10.1016/j.phrs.2020.104932	article	"The GDSC dataset contains 140 drug sensitivity experiments results in 624 cell lines"	Drug response prediction	AUC, sensitivity, specificity (5-fold CV)	cross-validation	"The proposed model of this paper used statistical methods and Machine Learning methods combined with genomics data to accurately predict the performance of oncology drugs on cancer cell lines."	"The proposed model of this paper used statistical methods and Machine Learning methods combined with genomics data to accurately predict the performance of oncology drugs on cancer cell lines."
300	Michuda, J and Leibowitz, B and Amar-Farkash, S and Bevis, C and Breschi, A and Kapilivsky, J and Igartua, C and Bell, J S and Beauchamp, K A and White, K and Stumpe, M and Beaubier, N and Taxter, T	Multimodal prediction of diagnosis for cancers of unknown primary	Cancer Research	80		16		2020	USA	http://dx.doi.org/10.1158/1538-7445.AM2020-5423	meeting abstract	"Targeted DNA sequencing for more than 500 cancer-associated genes and exome-capture RNA sequencing was carried out in more than 25,000 fresh frozen or paraffin embedded tumor samples, including both primary and metastatic tumors"	Differential diagnosis prediction	accuracy (training/test set)	training + test set	"The incorporation of multiple modes of omics data can improve the interpretability and robustness of machine learning models to predict cancer diagnosis"	"The incorporation of multiple modes of omics data can improve the interpretability and robustness of machine learning models to predict cancer diagnosis"
301	Miller, J B and Kauwe, J S K	Predicting Clinical Dementia Rating Using Blood RNA Levels	Genes (Basel)	11	6			2020	USA	http://dx.doi.org/10.3390/genes11060706	article	"We divided 741 ADNI participants with blood microarray data into three groups based on their most recent CDR assessment: cognitive normal (CDR = 0), mild cognitive impairment (CDR = 0.5), and probable Alzheimer's disease (CDR ≥ 1.0)"	Differential diagnosis prediction	AUC (10-fold CV)	cross-validation	"Our analyses indicate that machine learning may be able to predict cognitive decline in individuals using RNA levels from a blood microarray by taking into account small differences in expression that are individually nonsignificant. A support vector machine was able to increase predictive accuracy of AD from a 55% baseline to almost 90%."	"Our analyses indicate that machine learning may be able to predict cognitive decline in individuals using RNA levels from a blood microarray by taking into account small differences in expression that are individually nonsignificant. A support vector machine was able to increase predictive accuracy of AD from a 55% baseline to almost 90%."
302	Miller-Atkins, G and Acevedo-Moreno, L A and Grove, D and Dweik, R A and Tonelli, A R and Brown, J M and Allende, D S and Aucejo, F and Rotroff, D M	Breath Metabolomics Provides an Accurate and Noninvasive Approach for Screening Cirrhosis, Primary, and Secondary Liver Tumors	Hepatology Communications	4	7		1041-1055	2020	USA	http://dx.doi.org/10.1002/hep4.1499	article	"Here, we investigate whether 22 VOCs from the breath of 296 patients can distinguish those with no liver disease (n = 54), cirrhosis (n = 30), HCC (n = 112), pulmonary hypertension (n = 49), or colorectal cancer liver metastases (n = 51)"	Differential diagnosis prediction	balanced accuracy (cross-validation)	cross-validation	"The use of machine learning and breath VOCs [Volatile organic compounds] shows promise as an approach to develop improved, noninvasive screening tools for chronic liver disease and primary and secondary liver tumors"	
303	Mongan, D and Focking, M and Healy, C and Susai, S R and Cagney, G and Cannon, M and Zammit, S and Nelson, B and McGorry, P and Nordentoft, M and Krebs, M O and Riecher-Rossler, A and Bressan, R and Barrantes-Vidal, N and Borgwardt, S and Ruhrmann, S and Sachs, G and Van Der Gaag, M and Rutten, B and Pantelis, C and De Haan, L and Valmaggia, L and Kempton, M and McGuire, P and Cotter, D	Development of proteomic prediction models for outcomes in the clinical high risk state and psychotic experiences in adolescence: Machine learning analyses in two nested casecontrol studies	Schizophrenia Bulletin	46			S238-S239	2020	Ireland	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7234235/	meeting abstract	"The sample comprised 133 CHR participants who were followed clinically for up to 6 years, of whom 49 transitioned to psychosis and 84 did not"	Case-control study	AUC, PPV, NPV (training + test set)	training + test set	"With external validation, models incorporating proteomic data may contribute to improved prediction of clinical outcomes in individuals at risk of psychosis"	"With external validation, models incorporating proteomic data may contribute to improved prediction of clinical outcomes in individuals at risk of psychosis"
304	Mostavi, M and Chiu, Y C and Huang, Y and Chen, Y	Convolutional neural network models for cancer type prediction based on gene expression	BMC Med Genomics	13			44-44	2020	USA	http://dx.doi.org/10.1186/s12920-020-0677-2	article	"The models were trained and tested on gene expression profiles from combined 10,340 samples of 33 cancer types and 713 matched normal tissues of The Cancer Genome Atlas (TCGA)"	Differential diagnosis prediction	accuracy (6x 5-fold CV, 80–20% splitting for training and validation)	cross-validation	"Taken together, we have presented three unique CNN architectures that take high dimension gene expression inputs and perform cancer type prediction while considering their tissue of origin. Our model achieved an equivalent 95.7% prediction accuracy comparing to earlier published studies, however with a drastically simplified CNN construction and with a reduced influence of the tissue origin."	"Taken together, we have presented three unique CNN architectures that take high dimension gene expression inputs and perform cancer type prediction while considering their tissue of origin. Our model achieved an equivalent 95.7% prediction accuracy comparing to earlier published studies, however with a drastically simplified CNN construction and with a reduced influence of the tissue origin."
305	Myasoedova, E and Athreya, A and Crowson, C and Weinshilboum, R and Wang, L and Matteson, E	Individualized Prediction of Response to Methotrexate Treatment in Patients with Rheumatoid Arthritis: A Pharmacogenomics-driven Machine Learning Approach	Arthritis and Rheumatology	72			4014-4015	2020	USA	http://dx.doi.org/10.1002/art.41538	meeting abstract	"A total of 647 patients were included: 336 recruited in the United Kingdom [UK]; 307 recruited across Europe (70% female; 72% rheumatoid factor [RF] positive; mean age 54 years; mean baseline Disease Activity Score with 28-joint count [DAS28] 5.65)"	Drug response prediction	AUC (5x 10-fold CV + external validation)	cross-validation + external cohort validation	"Pharmacogenomic biomarkers including gene variants for cancer susceptibility genes (CASC15) and important MTX pathway enzymes (ATIC) combined with baseline DAS28 score predicted MTX response in patients with early RA more reliably than demographics and baseline DAS28 alone, with replication in an independent cohort"	"Pharmacogenomic biomarkers including gene variants for cancer susceptibility genes (CASC15) and important MTX pathway enzymes (ATIC) combined with baseline DAS28 score predicted MTX response in patients with early RA more reliably than demographics and baseline DAS28 alone, with replication in an independent cohort"
306	Naz, H and Ahuja, S	Deep learning approach for diabetes prediction using PIMA Indian dataset	Journal of Diabetes and Metabolic Disorders	19	1		391-403	2020	India	http://dx.doi.org/10.1007/s40200-020-00520-5	article	"a total of 768 instances, from which 268 samples were identified as diabetic and 500 were non-diabetics"	Case-control study	accuracy ("splits in an 80/20% ratio into the training and validation set")	training + test set	"The outcome of the study confirms that DL provides the best results with the most promising extracted features. DL achieves the accuracy of 98.07%which can be used for further development of the automatic prognosis tool."	"The outcome of the study confirms that DL provides the best results with the most promising extracted features. DL achieves the accuracy of 98.07%which can be used for further development of the automatic prognosis tool."
307	Nazha, A and Sekeres, M A and Bejar, R and Rauh, M J and Othus, M and Komrokji, R S and Barnard, J and Hilton, C B and Kerr, C M and Steensma, D P and DeZern, A and Roboz, G and Garcia-Manero, G and Erba, H and Ebert, B L and Maclejewski, J P	Genomic biomarkers to predict resistance to hypomethylating agents in patients with myelodysplastic syndromes using artificial intelligence	JCO Precision Oncology	3				2019	USA	http://dx.doi.org/10.1200/PO.19.00119	article	"Among 433 patients, 193 (45%) received azacitidine, 176 (40%) received decitabine, and 64 (15%) received HMA alone or in combination"	Drug response prediction	accuracy (training/test set)	training + test set	"Genomic biomarkers can identify, with high accuracy, approximately one third of patients with MDS who will not respond to HMAs. This study highlights the importance of machine learning technologies such as the recommender system algorithm in translating genomic data into useful clinical tools."	"Genomic biomarkers can identify, with high accuracy, approximately one third of patients with MDS who will not respond to HMAs. This study highlights the importance of machine learning technologies such as the recommender system algorithm in translating genomic data into useful clinical tools."
308	Nielsen, R L and Helenius, M and Garcia, S L and Roager, H M and Aytan-Aktug, D and Hansen, L B S and Lind, M V and Vogt, J K and Dalgaard, M D and Bahl, M I and Jensen, C B and Muktupavela, R and Warinner, C and Aaskov, V and Gøbel, R and Kristensen, M and Frøkiær, H and Sparholt, M H and Christensen, A F and Vestergaard, H and Hansen, T and Kristiansen, K and Brix, S and Petersen, T N and Lauritzen, L and Licht, T R and Pedersen, O and Gupta, R	Data integration for prediction of weight loss in randomized controlled dietary trials	Sci Rep	10	1		20103-20103	2020	Denmark	http://dx.doi.org/10.1038/s41598-020-76097-z	article	"Here, we classify weight loss responders (N = 106) and non-responders (N = 97) of overweight non-diabetic middle-aged Danes to two earlier reported dietary trials over 8 weeks"	Treatment response prediction	AUC ("50 shuffle-split fivefold cross-validations was used")	cross-validation	"By identifying the propensity of study participants likely to experience weight loss, a more effective individual targeting of dietary interventions can be facilitated, eventually in concert with comprehensive population weight loss strategies. Furthermore, understanding predictive features of weight loss response will drive improved understanding of the interplay between gut microbiota, diet and individual predisposition."	"By identifying the propensity of study participants likely to experience weight loss, a more effective individual targeting of dietary interventions can be facilitated, eventually in concert with comprehensive population weight loss strategies. Furthermore, understanding predictive features of weight loss response will drive improved understanding of the interplay between gut microbiota, diet and individual predisposition."
309	Nyamundanda, G and Eason, K and Guinney, J and Lord, C J and Sadanandam, A	A machine-learning tool concurrently models single omics and phenome data for functional subtyping and personalized cancer medicine	Cancers	12		10	1-14	2020	United Kingdom	http://dx.doi.org/10.3390/cancers12102811	article	"in total, 2043 breast cancer samples were used in this work "	Subgroup stratification	Silhouette width, cophenetic correlation (external test datasets)	external cohort validation	"Overall, this genome-phenome machine-learning integration tool, PhenMap identifies functional and phenotype-integrated discrete or continuous subtypes with clinical translational potential."	"Overall, this genome-phenome machine-learning integration tool, PhenMap identifies functional and phenotype-integrated discrete or continuous subtypes with clinical translational potential."
310	Ozer, M E and Sarica, P O and Arga, K Y	New Machine Learning Applications to Accelerate Personalized Medicine in Breast Cancer: Rise of the Support Vector Machines	Omics	24	5		241-246	2020	Turkey	http://dx.doi.org/10.1089/omi.2020.0001	article	review (not applicable)	Review			"This expert review describes and examines, first, the SVM models employed to forecast breast cancer subtypes using diverse systems science data, including transcriptomics, epigenetics, proteomics, and radiomics, as well as biological pathway, clinical, pathological, and biochemical data. Then, we compare the performance of the present SVM and other diagnostic and therapeutic prediction models across the data types. We conclude by emphasizing that data integration is a critical bottleneck in systems science, cancer research and development, and health care innovation and that SVM and machine learning approaches offer new solutions and ways forward in biomedical, bioengineering, and clinical applications."	"This expert review describes and examines, first, the SVM models employed to forecast breast cancer subtypes using diverse systems science data, including transcriptomics, epigenetics, proteomics, and radiomics, as well as biological pathway, clinical, pathological, and biochemical data. Then, we compare the performance of the present SVM and other diagnostic and therapeutic prediction models across the data types. We conclude by emphasizing that data integration is a critical bottleneck in systems science, cancer research and development, and health care innovation and that SVM and machine learning approaches offer new solutions and ways forward in biomedical, bioengineering, and clinical applications."
311	Pai, S and Weber, P and Isserlin, R and Kaka, H and Hui, S and Shah, M A and Giudice, L and Giugno, R and Nøhr, A K and Baumbach, J and Bader, G D	netDx: Software for building interpretable patient classifiers by multi-'omic data integration using patient similarity networks	F1000Res	9			1239-1239	2020	Canada	http://dx.doi.org/10.12688/f1000research.26429.2	article	"including 154 Luminal A and 194 tumours of other subtypes"	Case-control study	AUROC, AUPR, and accuracy (an approximately 70:30 split of samples was used for cross validation)	cross-validation	"the netDx Bioconductor package provides a novel workflow for pathway-based patient classification from sparse genetic data"	"the netDx Bioconductor package provides a novel workflow for pathway-based patient classification from sparse genetic data"
312	Rachid Zaim, S and Kenost, C and Berghout, J and Chiu, W and Wilson, L and Zhang, H H and Lussier, Y A	binomialRF: interpretable combinatoric efficiency of random forests to identify biomarker interactions	BMC Bioinformatics	21	1		374-374	2020	USA	https://doi.org/10.1186/s12859-020-03718-9	article	"we conduct a variety of simulations and trials against the Madelon benchmark dataset from the University of California – Irvine (UCI), and clinical datasets from The Cancer Genome Atlas (TCGA)"	Case-control study	Precision, recall, test error (training and test set)	training + test set	"binomialRF extends upon previous methods for identifying interpretable features in RFs and brings them together under a correlated binomial distribution to create an efficient hypothesis testing algorithm that identifies biomarkers' main effects and interactions. Preliminary results in simulations demonstrate computational gains while retaining competitive model selection and classification accuracies."	"binomialRF extends upon previous methods for identifying interpretable features in RFs and brings them together under a correlated binomial distribution to create an efficient hypothesis testing algorithm that identifies biomarkers' main effects and interactions. Preliminary results in simulations demonstrate computational gains while retaining competitive model selection and classification accuracies."
313	Radakovich, N and Meggendorfer, M and Malcovati, L and Sekeres, M A and Shreve, J and Beau Hilton, C and Rouphail, Y and Walter, W and Hutter, S and Mukherjee, S and Kerr, C M and Jha, B K and Galli, A and Pozzi, S and Gerds, A T and Haferlach, C and Maciejewski, J P and Haferlach, T and Nazha, A	A personalized clinical-decision tool to improve the diagnostic accuracy of myelodysplastic syndromes	Blood	136			33-35	2020	USA	http://dx.doi.org/10.1182/blood-2020-139412	meeting abstract	"Clinical and genomic data, including commercially available next-generation sequencing panels, were obtained for patients (pts) treated at the Cleveland Clinic (CC; 652 pts), Munich Leukemia Laboratory (MLL; 1509 pts), and the University of Pavia in Italy (UP, 536 pts)"	Case-control study	AUC (training + external validation)	external cohort validation	"We developed and externally validated a highly accurate and interpretable model that can distinguish MDS from other myeloid malignancies using clinical and mutational data from a large international cohort. The model can provide personalized interpretations of its outcome and can aid physicians and hematopathologists in recognizing MDS with high accuracy when encountering pts with pancytopenia and with a suspected diagnosis of MDS."	"We developed and externally validated a highly accurate and interpretable model that can distinguish MDS from other myeloid malignancies using clinical and mutational data from a large international cohort. The model can provide personalized interpretations of its outcome and can aid physicians and hematopathologists in recognizing MDS with high accuracy when encountering pts with pancytopenia and with a suspected diagnosis of MDS."
314	Rajula, H S R and Verlato, G and Manchia, M and Antonucci, N and Fanos, V	Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment	Medicina-Lithuania	56	9			2020	Italy	http://dx.doi.org/10.3390/medicina56090455	article	review (not applicable)	Review			"In this article, we compare the usefulness and limitations of traditional statistical methods and ML, when applied to the medical field. Traditional statistical methods seem to be more useful when the number of cases largely exceeds the number of variables under study and a priori knowledge on the topic under study is substantial such as in public health. ML could be more suited in highly innovative fields with a huge bulk of data, such as omics, radiodiagnostics, drug development, and personalized treatment. Integration of the two approaches should be preferred over a unidirectional choice of either approach."	"In this article, we compare the usefulness and limitations of traditional statistical methods and ML, when applied to the medical field. Traditional statistical methods seem to be more useful when the number of cases largely exceeds the number of variables under study and a priori knowledge on the topic under study is substantial such as in public health. ML could be more suited in highly innovative fields with a huge bulk of data, such as omics, radiodiagnostics, drug development, and personalized treatment. Integration of the two approaches should be preferred over a unidirectional choice of either approach."
315	Rauschert, S and Melton, P E and Heiskala, A and Karhunen, V and Burdge, G and Craig, J M and Godfrey, K M and Lillycrop, K and Mori, T A and Beilin, L J and Oddy, W H and Pennell, C and Järvelin, M R and Sebert, S and Huang, R C	Machine Learning-Based DNA Methylation Score for Fetal Exposure to Maternal Smoking: Development and Validation in Samples Collected from Adolescents and Adults	Environ Health Perspect	128	9		97003-97003	2020	Australia	http://dx.doi.org/10.1289/ehp6076	article	"The score was developed and tested in the Raine Study with data from 995 white 17-y-old participants using 10-fold cross-validation"	Case-control study	Sensitivity, specificity (10-fold CV)	cross-validation	"We have developed a DNA methylation score for exposure to maternal smoking during pregnancy, outperforming the three previously developed scores. One possible application of the current score could be for model adjustment purposes or to assess its association with distal health outcomes where part of the effect can be attributed to maternal smoking. Further, it may provide a biomarker for fetal exposure to maternal smoking."	"We have developed a DNA methylation score for exposure to maternal smoking during pregnancy, outperforming the three previously developed scores. One possible application of the current score could be for model adjustment purposes or to assess its association with distal health outcomes where part of the effect can be attributed to maternal smoking. Further, it may provide a biomarker for fetal exposure to maternal smoking."
316	Ristori, M V and Mortera, S L and Marzano, V and Guerrera, S and Vernocchi, P and Ianiro, G and Gardini, S and Torre, G and Valeri, G and Vicari, S and Gasbarrini, A and Putignani, L	Proteomics and Metabolomics Approaches towards a Functional Insight onto AUTISM Spectrum Disorders: Phenotype Stratification and Biomarker Discovery	Int J Mol Sci	21		17		2020	Italy	http://dx.doi.org/10.3390/ijms21176274	article	review (not applicable)	Review			"studies are limited in their evaluation of biomarkers by comparisons of patients with ASD and healthy controls, without considering the family and specific characteristics of the pathology. Often, the sample cohort is also highly limited. From the point of view of omics data, the biggest limit is that all of the data from the omics are not considered and the data are not integrated with collected clinical data."	"studies are limited in their evaluation of biomarkers by comparisons of patients with ASD and healthy controls, without considering the family and specific characteristics of the pathology. Often, the sample cohort is also highly limited. From the point of view of omics data, the biggest limit is that all of the data from the omics are not considered and the data are not integrated with collected clinical data."
317	Romero-Rosales, B L and Tamez-Pena, J G and Nicolini, H and Moreno-Treviño, M G and Trevino, V	Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling	PLoS One	15	4		e0232103-e0232103	2020	Mexico	http://dx.doi.org/10.1371/journal.pone.0232103	article	"The study has four groups accounting for 5,220 individuals"	Case-control study	AUC (20 rounds of internal cross-validation (CV) to 80% of the dataset for training and 20% for testing)	cross-validation	"In this research, we compared three machine learning methods that have been proved to construct powerful predictive models (genetic algorithms, LASSO, and step-wise) and propose the inclusion of markers from misclassified samples to improve overall prediction accuracy. Our results show that the addition of markers from an initial model plus the markers of the model fitted to misclassified samples improves the area under the receiving operative curve by around 5%, reaching ~0.84, which is highly competitive using only genetic information"	"In this research, we compared three machine learning methods that have been proved to construct powerful predictive models (genetic algorithms, LASSO, and step-wise) and propose the inclusion of markers from misclassified samples to improve overall prediction accuracy. Our results show that the addition of markers from an initial model plus the markers of the model fitted to misclassified samples improves the area under the receiving operative curve by around 5%, reaching ~0.84, which is highly competitive using only genetic information"
318	Rychkov, D and Neely, J and Sirota, M	Uncovering Novel Biomarkers for Rheumatoid Arthritis from Feature Selection and Machine Learning Approaches on Synovium and Blood Gene Expression Data	Arthritis and Rheumatology	72			1503-1504	2020	USA	https://acrabstracts.org/abstract/uncovering-novel-biomarkers-for-rheumatoid-arthritis-from-feature-selection-and-machine-learning-approaches-on-synovium-and-blood-gene-expression-data/	meeting abstract	"The raw data from 13 synovium datasets with 284 samples and 14 blood datasets with 1,885 samples were downloaded and processed"	Case-control study	AUC (training + test set)	training + test set	"This novel list of biomarkers, identified through a robust feature selection procedure on public data and validated using multiple independent data sets, coupled with the RAScore may be useful in the early diagnosis and disease and treatment monitoring of RA."	"This novel list of biomarkers, identified through a robust feature selection procedure on public data and validated using multiple independent data sets, coupled with the RAScore may be useful in the early diagnosis and disease and treatment monitoring of RA."

319	Saorin, A and Di Gregorio, E and Miolo, G and Steffan, A and Corona, G	Emerging role of metabolomics in ovarian cancer diagnosis	Metabolites	10	10	1-15	2020	Italy	http://dx.doi.org/10.3390/metabo10100419	article	review (not applicable)	Review			"The most promising circulating signatures of OC [Ovarian Cancer] involve metabolites belonging to lipids and AA pathways. These metabolic fingerprints find agreement in many studies, making them relevant for OC diagnosis. However, their clinical application appears to be limited because a lack of independent, large validation studies prevents their effective use for OC screening and monitoring. Future research should include better designed studies on large homogeneous populations that include proper external validation in order to further improve the translational success of metabolomics for OC diagnosis."	"The most promising circulating signatures of OC [Ovarian Cancer] involve metabolites belonging to lipids and AA pathways. These metabolic fingerprints find agreement in many studies, making them relevant for OC diagnosis. However, their clinical application appears to be limited because a lack of independent, large validation studies prevents their effective use for OC screening and monitoring. Future research should include better designed studies on large homogeneous populations that include proper external validation in order to further improve the translational success of metabolomics for OC diagnosis."
320	Schaack, D and Brenner, T and Weigand, M and Uhle, F	Deep-learning neural networks for accurate diagnosis of sepsis using microarray gene expression data	Intensive Care Medicine Experimental	7			2019	Germany	http://dx.doi.org/10.1186/s40635-019-0265-y	meeting abstract	"septic patients (n=1,354), trauma patients (n=478), and healthy controls (n=383)"	Case-control study	AUC, sensitivity, specificity (training, validation and test set)	training + test set	"The presented approach surpasses the most advanced solutions for sepsis classification and closes the gap to comparable studies on cancer diagnosis. By limiting the number of available genes for prediction, we can prove that, instead of learning idiosyncratic features tailored to specific data series, generalized strategies for sample discrimination have developed in the trained artificial neural networks. The combination of artificial neural networks and microarray gene expression data is therefore capable of achieving sepsis diagnosis with superior accuracy and thus augments the current diagnostic scope."	"The presented approach surpasses the most advanced solutions for sepsis classification and closes the gap to comparable studies on cancer diagnosis. By limiting the number of available genes for prediction, we can prove that, instead of learning idiosyncratic features tailored to specific data series, generalized strategies for sample discrimination have developed in the trained artificial neural networks. The combination of artificial neural networks and microarray gene expression data is therefore capable of achieving sepsis diagnosis with superior accuracy and thus augments the current diagnostic scope."
321	Schperberg, A V and Boichard, A and Tsigelny, I F and Richard, S B and Kurzrock, R	Machine learning model to predict oncologic outcomes for drugs in randomized clinical trials	Int J Cancer	147	9	2537-2549	2020	USA	http://dx.doi.org/10.1002/ijc.33240	article	"A total of 467 progression-free survival (PFS) and 369 overall survival (OS) data points were used as training sets to build our ML (random forest) model"	Treatment response prediction	Spearman correlation (cross-validation)	cross-validation	"In this study, we determine whether machine learning (ML) can extract meaningful associations between oncologic outcome and clinical trial, drug-related biomarker and molecular profile information. [...] The Spearman correlation (rs) between predicted and actual outcomes was statistically significant (PFS: rs = 0.879, OS: rs = 0.878, P < .0001)."	"In this study, we determine whether machine learning (ML) can extract meaningful associations between oncologic outcome and clinical trial, drug-related biomarker and molecular profile information. [...] The Spearman correlation (rs) between predicted and actual outcomes was statistically significant (PFS: rs = 0.879, OS: rs = 0.878, P < .0001)."
322	Senturk, N and Tuncel, G and Koseoglu, S and Dogan, B and Sag, S O and Mocan, G and Temel, S G and Dundar, M and Ergoren, M C	Developing evidence based computerized diagnostic tools for breast cancer early prediction	Gazi Medical Journal	31	3	P44-P44	2020	Turkey	https://www.embase.com/search/results?subaction=viewrecord&id=L633285107&from=export	article	"268 different BRCA1/2 positive breast cancer patients"	Case-control study	accuracy (training/test set)	training + test set	"Overall, our developed models will provide the early prediction for BRCA1/BRCA2 related breast cancer cases and will improve to be beneficial for preventive medicine and a unique example for today's genetic-based personalized medicine software"	"Overall, our developed models will provide the early prediction for BRCA1/BRCA2 related breast cancer cases and will improve to be beneficial for preventive medicine and a unique example for today's genetic-based personalized medicine software"
323	Singh, M and Singh, S P and Dubey, P K and Rachana, R and Mani, S and Yadav, D and Agarwal, M and Agarwal, S and Agarwal, V and Kaur, H	Advent of Proteomic Tools for Diagnostic Biomarker Analysis in Alzheimer's Disease	Curr Protein Pept Sci	21	10	965-977	2020	India	http://dx.doi.org/10.2174/1389203721666200615173213	article	review (not applicable)	Review			"the molecular diagnosis of AD incorporates various sophisticated techniques including immuno-sensing, machine learning, nano conjugation-based detections, etc. In the current review description, we have summarized the various diagnostic approaches and their relevance in mitigating the long-standing urgency of targeted diagnostic tools for detection of AD"	"the molecular diagnosis of AD incorporates various sophisticated techniques including immuno-sensing, machine learning, nano conjugation-based detections, etc. In the current review description, we have summarized the various diagnostic approaches and their relevance in mitigating the long-standing urgency of targeted diagnostic tools for detection of AD"
324	Singh, N P and Vinod, P K	Integrative analysis of DNA methylation and gene expression in papillary renal cell carcinoma	Mol Genet Genomics	295	3	807-824	2020	India	http://dx.doi.org/10.1007/s00438-020-01664-y	article	"The pathological stages are known for 250 samples (common across both the platforms) with the following distributions: Stage I—167, Stage II—19, Stage III—50, and Stage IV—14. We divided the dataset containing these 250 samples into training (80%) and test (20%) datasets"	Tumor stage prediction	PR AUC, MCC, Accuracy, Sensitivity and Specificity ("The performance of the models was evaluated on the 20% test dataset")	training + test set	"In this study, we performed an integrative analysis of DNA methylation and gene expression to characterize the patterns of DNA methylation in PRCC. Our analysis showed that most probes are hypermethylated in RCC, and both hyper- and hypo-methylated probes can distinguish normal from cancer samples."	"In this study, we performed an integrative analysis of DNA methylation and gene expression to characterize the patterns of DNA methylation in PRCC. Our analysis showed that most probes are hypermethylated in RCC, and both hyper- and hypo-methylated probes can distinguish normal from cancer samples."
325	Song, X and Yang, X and Narayanan, R and Shankar, V and Ethiraj, S and Wang, X and Duan, N and Ni, Y H and Hu, Q and Zare, R N	Oral squamous cell carcinoma diagnosed from saliva metabolic profiling	Proc Natl Acad Sci U S A	117	28	16167-16173	2020	China	http://dx.doi.org/10.1073/pnas.2001395117	article	"Saliva samples from 373 volunteers, 124 who are healthy, 124 who have premalignant lesions, and 125 who are OSCC patients"	Differential diagnosis prediction	Accuracy ("20-fold cross-validation was carried out", external validation samples)	cross-validation + test set	"The salivary metabolic profile can reflect oral cancer development. Most discovered metabolites in saliva were found to be highly linked to their expression levels within the primary oncological site of oral cavity tissues, demonstrating the potential of saliva for in vitro molecular diagnosis of OSCC."	"The salivary metabolic profile can reflect oral cancer development. Most discovered metabolites in saliva were found to be highly linked to their expression levels within the primary oncological site of oral cavity tissues, demonstrating the potential of saliva for in vitro molecular diagnosis of OSCC."
326	Stein, A S and Watson, D and Nair, P R and Basu, K and Ullal, Y S and Ghosh, A and Narvekar, Y and Grover, H and Sahu, D and Prakash, A and Behura, L and Balakrishnan, V and Ghosh Roy, K and Rajagopalan, S and Alam, A and Parashar, R and Mundkur, N and Christie, J and Macpherson, M D and Kapoor, S and Marcucci, G	Superior therapy response predictions for patients withmyelodysplastic syndrome (MDS) using cellworks Singula™: Mycare-020-02	Blood	136		9-10	2020	USA	http://dx.doi.org/10.1182/blood-2020-142214	meeting abstract	"The performance of Singula™ was evaluated in an independent, randomly selected, retrospective cohort of 144 MDS patients "	Therapy response prediction	Accuracy + confidence intervals (training + external cohort)	external cohort validation	"Cellworks Singula™ has high accuracy and sensitivity in predicting CR [complete response] for MDS [Myelodysplastic Syndrome] patient response to physician prescribed therapies. Singula™ also has high specificity in identifying patients who are unlikely to respond to physician prescribed therapies and provides alternative treatment recommendations for these patients."	"Cellworks Singula™ has high accuracy and sensitivity in predicting CR [complete response] for MDS [Myelodysplastic Syndrome] patient response to physician prescribed therapies. Singula™ also has high specificity in identifying patients who are unlikely to respond to physician prescribed therapies and provides alternative treatment recommendations for these patients."
327	Tabares-Soto, R and Orozco-Arias, S and Romero-Cano, V and Bucheli, V S and Rodriguez-Sotelo, J L and Jimenez-Varon, C F	A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data	Peerj Computer Science				2020	Colombia	http://dx.doi.org/10.7717/peerj-cs.270	article	"This database consists of 174 samples with 12,533 gene expression microarrays for 11 different types of cancer."	Case-control study	accuracy, confusion matrix (10-fold CV)	cross-validation	"In this work, we show the application of unsupervised and supervised learning approaches of ML and DL for the classification of 11 cancer types based on a microarray dataset. We observed that the best average results using the training and validation data are obtained using the raw dataset and the LR algorithm, yielding an accuracy value of 100% (validation set, using the hold-out splitting method). [...] Additional tests with independent data should be done to discard potential overfitting."	"In this work, we show the application of unsupervised and supervised learning approaches of ML and DL for the classification of 11 cancer types based on a microarray dataset. We observed that the best average results using the training and validation data are obtained using the raw dataset and the LR algorithm, yielding an accuracy value of 100% (validation set, using the hold-out splitting method). [...] Additional tests with independent data should be done to discard potential overfitting."
328	Talhouk, A and George, J and Wang, C and Goode, E and Ramus, S and Doherty, J and Bowtell, D and Anglesio, M	PrOTYPE (Predictor of high-grade-serous Ovarian carcinoma molecular subTYPE): The development and validation of a clinical-grade consensus classifier for the molecular subtypes of high-grade serous tubo-ovarian cancer	Clinical Cancer Research	26	13		2020	Canada	http://dx.doi.org/10.1158/1557-3265.OVCA19-A03	meeting abstract	Adopting two independent approaches, we derived and internally validated algorithms for molecular subtype prediction from gene-expression array data in 1,650 tumors.	Differential diagnosis prediction	accuracy (training + external validation)	external cohort validation	"We validated the Predictor of high-grade-serous Ovarian carcinoma molecular subTYPE, or PrOTYPE, following the Institute of Medicine guidelines for the development of omics-based tests. This simple-to-use, cost-effective, fully defined, and locked-down clinical-grade assay will facilitate molecular subtype stratification into clinical trial design."	"We validated the Predictor of high-grade-serous Ovarian carcinoma molecular subTYPE, or PrOTYPE, following the Institute of Medicine guidelines for the development of omics-based tests. This simple-to-use, cost-effective, fully defined, and locked-down clinical-grade assay will facilitate molecular subtype stratification into clinical trial design."
329	Talwar, J and Carter, H	Assessing cancer drug responseprediction from gene expression	Cancer Research	80	16		2020	Canada	http://dx.doi.org/10.1158/1538-7445.AM2020-2099	meeting abstract	"Gene expression data for 17,737 genes across 1014 human cancer cell-lines with IC50 concentrations for 251 anti-cancer drugs were obtained from the Genomics of Drug Sensitivity in Cancer Project"	Drug response prediction	accuracy (train/test set)	training + test set	"Overall, our analysis shows that utilizing gene expression profiles independent of other -omics data for cancer drug response prediction through machine learning frameworks offers modest predictive capabilities. To increase performance, we suggest augmenting training size through shared pathway cross-training, optimizing feature encoding to maximize neural network predictive capabilities, and incorporating other -omics data."	"Overall, our analysis shows that utilizing gene expression profiles independent of other -omics data for cancer drug response prediction through machine learning frameworks offers modest predictive capabilities. To increase performance, we suggest augmenting training size through shared pathway cross-training, optimizing feature encoding to maximize neural network predictive capabilities, and incorporating other -omics data."
330	Tang, B and Wang, Y and Chen, Y and Li, M and Tao, Y	A Novel Early-Stage Lung Adenocarcinoma Prognostic Model Based on Feature Selection With Orthogonal Regression	Frontiers in Cell and Developmental Biology	8			2020	China	http://dx.doi.org/10.3389/fcell.2020.620746	article	"A total of four LUAD expression profiles (GSE32036, GSE32867, GSE33532, and GSE75037) were retrieved from GEO. And the corresponding accession number, platform, and sample information are listed in Table 1. Based on survival information from a total of 479 LUAD samples, the risk score was stratified into high and low groups."	Case-control study	AUC (cross-validation + external validation)	cross-validation + external cohort validation	"In conclusion, the proposed FSOR [feature selection with orthogonal regression] method can deliver better prediction performance for the early-stage prognosis and has the potential to improve therapy strategy, but with few predictor consideration and computation burden."	"In conclusion, the proposed FSOR [feature selection with orthogonal regression] method can deliver better prediction performance for the early-stage prognosis and has the potential to improve therapy strategy, but with few predictor consideration and computation burden."
331	Tang, W and Cao, Y and Ma, X	Novel prognostic prediction model constructed through machine learning on the basis of methylation-driven genes in kidney renal clear cell carcinoma	Biosci Rep	40	7		2020	China	http://dx.doi.org/10.1042/bsr20201604	article	"normal samples = 160, tumor samples = 325"	Prognostic study	AUC (10-fold CV, test set)	cross-validation + test set	"We used the machine learning method to establish a multivariate methylation prognostic prediction model and combined with clinical information to build the trans-omics prognostic nomogram. [...] These results can help in the accurate evaluation of the prognosis of KIRC patients and provide new clues and data resources for the further study of the pathogenesis and the development of the disease."	"We used the machine learning method to establish a multivariate methylation prognostic prediction model and combined with clinical information to build the trans-omics prognostic nomogram. [...] These results can help in the accurate evaluation of the prognosis of KIRC patients and provide new clues and data resources for the further study of the pathogenesis and the development of the disease."
332	Tawk, B and Wirkner, U and Schwager, C and Herpel, E and Tinhofer, I and Budach, V and Krause, M and Stuschke, M and Balermpas, P and Roedel, C and Grosu, A and Zips, D and Combs, S E and Weichert, W and Belka, C and Baumann, M and Herold-Mende, C and Debus, J and Abdollahi, A	Hypoxia Methylome Classifier (HDMC) Outperforms Gene Signatures in Identifying HPV-Negative HNSCC Patients at Risk for Locoregional Failure Post Primary Radiochemotherapy: A German Cancer Consortium Radiation Oncology Group (DKTK-ROG) Multicenter Trial	International Journal of Radiation Oncology Biology Physics	108	3	e552-e553	2020	Germany	http://dx.doi.org/10.1016/j.ijrobp.2020.07.1715	article	"To identify the prognostic impact of tumor hypoxia on increased risk for loco-regional recurrence (LR) and all event progression in Human Papilloma Virus DNA negative (HPV-) HNSCC, methylation data (450K Illumina, FFPE material) was obtained from the homogeneous primary RCHT DKTK-ROG validation cohort (total n = 117, HPV- n = 88). HDMC was trained in the TCGA-HNSCC cohort (total n = 278, HPV- n = 242) where matching RNA-seq based assignments for 15- and 30 GEsOH as well as methylome data were available. A random forest machine learning based HDMC was developed based on differentially methylated probes (5129, FDR<0.05) between tumors with high GEsOH (consensus, n = 96) vs those with intermediate-to low GEsOH (n = 146). HDMC was validated in the DKTK-ROG primary cohort."	Differential diagnosis and survival prediction	correlation (training + validation cohort)	external cohort validation	"A methylation-based classifier of tumor hypoxia is successfully developed and validated to be prognostic for LR [loco-regional recurrence], progression and OS [overall survival] in HPV-HNSCC patients treated with primary RCHT."	"A methylation-based classifier of tumor hypoxia is successfully developed and validated to be prognostic for LR [loco-regional recurrence], progression and OS [overall survival] in HPV-HNSCC patients treated with primary RCHT."
333	Tiedt, S and Brandmaier, S and Düring, M and Artati, A and Klein, M and Liebig, T and Holdt, L and Teupser, D and Wang-Sattler, R and Schwedhelm, E and Gieger, C and Dichgans, M	Circulating metabolites differentiate acute ischemic stroke from stroke mimics	International Journal of Stroke	15	1	77-78	2020	Germany	http://dx.doi.org/10.1177/1747493020963387	meeting abstract	"We performed untargeted metabolomics on serum samples obtained from patients with ischemic stroke (N = 219) and stroke mimics (N = 138; as defined by absence of a DWI positive lesion on MRI)"	Case-control study	AUC (training + test set)	training + test set	"We performed untargeted metabolomics on serum samples obtained from patients with ischemic stroke (N = 219) and stroke mimics (N = 138; as defined by absence of a DWI positive lesion on MRI)"	"We performed untargeted metabolomics on serum samples obtained from patients with ischemic stroke (N = 219) and stroke mimics (N = 138; as defined by absence of a DWI positive lesion on MRI)"
334	Tran, A and Walsh, C J and Batt, J and Dos Santos, C C and Hu, P	A machine learning-based clinical tool for diagnosing myopathy using multi-cohort microarray expression profiles	J Transl Med	18	1	454-454	2020	Canada	http://dx.doi.org/10.1186/s12967-020-02630-3	article	"Muscle tissue samples originating from 1260 patients with muscle weakness."	Differential diagnosis prediction	AUC (training + test set)	training + test set	"Our results present a well-performing molecular classification tool with the selected gene markers for muscle disease classification. In practice, this tool addresses an important gap in the literature on myopathies and presents a potentially useful clinical tool for muscle disease subtype diagnosis."	"Our results present a well-performing molecular classification tool with the selected gene markers for muscle disease classification. In practice, this tool addresses an important gap in the literature on myopathies and presents a potentially useful clinical tool for muscle disease subtype diagnosis."
335	Tran, P M H and Tran, L K H and Nechtman, J and Dos Santos, B and Purohit, S and Satter, K B and Dun, B and Kolhe, R and Sharma, S and Bollag, R and She, J X	Comparative analysis of transcriptomic profile, histology, and IDH mutation for classification of gliomas	Sci Rep	10	1	20651-20651	2020	USA	http://dx.doi.org/10.1038/s41598-020-77777-6	article	"RNAseq and microarray data were obtained for 1032 gliomas from the TCGA and 395 gliomas from REMBRANDT"	Differential diagnosis and survival prediction	accuracy, log rank test p-value (cross-validation)	cross-validation	"We developed and validated an unbiased, automated pipeline for transcriptomic clustering. Without any domain knowledge, our classifier recapitulated known glioma subtypes from histology and mutation status. Our analytical pipeline avoids the potential of overfitting a supervised model to misclassified or mishandled samples [...] and can be used in establishing gold standard datasets devoid of erroneous and questionable samples for the development of automated tumor classifiers"	"We developed and validated an unbiased, automated pipeline for transcriptomic clustering. Without any domain knowledge, our classifier recapitulated known glioma subtypes from histology and mutation status. Our analytical pipeline avoids the potential of overfitting a supervised model to misclassified or mishandled samples [...] and can be used in establishing gold standard datasets devoid of erroneous and questionable samples for the development of automated tumor classifiers"
336	van der Heijden, M and Essers, P B M and Verhagen, C V M and Willems, S M and Sanders, J and de Roest, R H and Vossen, D M and Leemans, C R and Verheij, M and Brakenhoff, R H and van den Brekel, M W M and Vens, C	Epithelial-to-mesenchymal transition is a prognostic marker for patient outcome in advanced stage HNSCC patients treated with chemoradiotherapy	Radiother Oncol	147		186-194	2020	Netherlands	http://dx.doi.org/10.1016/j.radonc.2020.05.013	article	"Pretreatment tumor material from patients of two cohorts, totalling 174 cisplatin-based chemoradiotherapy treated HPV-negative HNSCC patients, was RNA-sequenced"	Prognostic study	AUC (cross-validation + external validation)	cross-validation + external cohort validation	"EMT in HPV-negative HNSCC co-defines patient outcome after chemoradiotherapy. The generated HNSCC-EMT prediction models can function as strong prognostic biomarkers."	"EMT in HPV-negative HNSCC co-defines patient outcome after chemoradiotherapy. The generated HNSCC-EMT prediction models can function as strong prognostic biomarkers."
337	Vittrant, B and Leclercq, M and Martin-Magniette, M L and Collins, C and Bergeron, A and Fradet, Y and Droit, A	Identification of a Transcriptomic Prognostic Signature by Machine Learning Using a Combination of Small Cohorts of Prostate Cancer	Frontiers in Genetics	11			2020	Canada	http://dx.doi.org/10.3389/fgene.2020.550894	article	"Gene expression data were extracted from three RNA-Seq datasets cumulating a total of 171 PCa patients"	Prognostic study	balanced error rate ("The resampling strategy was run 100 times with a split of 2/3 for training and 1/3 for test sets")	stratified resampling training and test sets	"Determining which treatment to provide to men with prostate cancer (PCa) is a major challenge for clinicians. [...] This study demonstrates the feasibility to regroup different small datasets in one larger to identify a predictive genomic signature that would benefit PCa patients."	"Determining which treatment to provide to men with prostate cancer (PCa) is a major challenge for clinicians. [...] This study demonstrates the feasibility to regroup different small datasets in one larger to identify a predictive genomic signature that would benefit PCa patients."

338	Wang, S and Su, W and Zhong, C and Yang, T and Chen, W and Chen, G and Liu, Z and Wu, K and Zhong, W and Li, B and Mao, X and Lu, J	An Eight-CircRNA Assessment Model for Predicting Biochemical Recurrence in Prostate Cancer	Frontiers in Cell and Developmental Biology	8			2020	China	http://dx.doi.org/10.3389/fcell.2020.599494	article	"The dataset is from the GEO database, using a cohort of 144 patients in Canada" "The training cohort consists of 148 hepatocellular carcinoma cases (median age of 63) and 84 healthy individuals (median age of 60)"	Prognostic study	AUC (10-fold CV)	cross-validation	"We constructed an eight circRNAs risk score model to reliably predict the BCR [Biochemical Recurrence] of PCa [Prostate Cancer] patients. We found that the BCR predicting effect may be related to the tumor microenvironment. At the same time, we preliminarily verified the function of circ_14736 and circ_17720 in vitro. Further experiments are necessary to clarify their roles in PCa."	"We constructed an eight circRNAs risk score model to reliably predict the BCR [Biochemical Recurrence] of PCa [Prostate Cancer] patients. We found that the BCR predicting effect may be related to the tumor microenvironment. At the same time, we preliminarily verified the function of circ_14736 and circ_17720 in vitro. Further experiments are necessary to clarify their roles in PCa."
339	Wang, Y and Wang, Y and Huang, A and Jiang, R and Zheng, J and Li, Z and Peng, J and Sun, J and Liu, C and Yang, G and Yuan, J and Yang, X and Zhou, J and Fan, J	The genetic and epigenetic abnormalities of plasma cfDNA as liquid biopsy biomarkers to diagnose hepatocellular carcinoma	Cancer Research	80	16		2020	China	http://dx.doi.org/10.1158/1538-7445.AM2020-782	meeting abstract	"Using a pre-specified mutation scoring system, we found that cfDNA mutation profiling achieved a sensitivity of 59.6%, 67.2%, and 46.8% for detecting HCC (n = 322), CRC (n = 244) and PC (n = 141) respectively, with a specificity of 95% in healthy controls (n = 207)"	Case-control study	AUC (10-fold CV + validation cohort)	cross-validation + external cohort validation	"In conclusion, our results suggest that cancer-derived abnormal methylation pattern of cfDNA provides promising biomarkers for the diagnosis of HCC [hepatocellular carcinoma] with high sensitivity and specificity."	"In conclusion, our results suggest that cancer-derived abnormal methylation pattern of cfDNA provides promising biomarkers for the diagnosis of HCC [hepatocellular carcinoma] with high sensitivity and specificity."
340	Wang, Y and Zheng, J and Li, Z and Jiang, R and Peng, J and Sun, J and Yang, G and Yang, X R and Huang, A and Wang, Y and Jie, Y and Liu, X and Gao, F and Wu, X and Wang, D and Wu, W and Lou, W and Zhou, J and Fan, J	Development of a novel liquid biopsy test to diagnose and locate gastrointestinal cancers	Journal of Clinical Oncology	38	15		2020	China	http://dx.doi.org/10.1200/JCO.2020.38.15_suppl.1557	meeting abstract	"Based on the metabolomics data from cohort 1 (504 HBV associated liver fibrosis patients and 502 normal controls, NC), we selected a panel of 4 predictive metabolite markers"	Case-control study	AUC, sensitivity, specificity (10-fold CV, training + validation cohort)	cross-validation + external cohort validation	" Plasma cfDNA methylome profiling identified effective biomarkers for the detection and tissue-of-origin determination of GI cancers, and outperformed mutation-based detection approach. Therefore, a liquid biopsy test capable of detecting and locating GI cancers is feasible and may serve as a valuable tool for early detection and intervention."	" Plasma cfDNA methylome profiling identified effective biomarkers for the detection and tissue-of-origin determination of GI cancers, and outperformed mutation-based detection approach. Therefore, a liquid biopsy test capable of detecting and locating GI cancers is feasible and may serve as a valuable tool for early detection and intervention."
341	Xie, G and Wang, X and Wei, R and Wang, J and Zhao, A and Chen, T and Wang, Y and Zhang, H and Xiao, Z and Liu, X and Deng, Y and Wong, L and Rajani, C and Kwee, S and Bian, H and Gao, X and Liu, P and Jia, W	Serum metabolite profiles are associated with the presence of advanced liver fibrosis in Chinese patients with chronic hepatitis B viral infection	BMC Med	18	1	144-144	2020	China	http://dx.doi.org/10.1186/s12916-020-01595-w	article	"Based on the metabolomics data from cohort 1 (504 HBV associated liver fibrosis patients and 502 normal controls, NC), we selected a panel of 4 predictive metabolite markers"	Differential diagnosis prediction	AUC (10-fold CV + validation cohort)	cross-validation + external cohort validation	"Our study showed that this 4-metabolite panel has potential usefulness in clinical assessments of CLD progression in patients with chronic hepatitis B virus infection." "In this study, we proposed a multi-scale clustering-based feature selection method for gene expression data, MCBFS, which performs clustering and feature weighting in a supervised manner. In the algorithm, a multi-scale distance function designed by us was used as a dissimilarity measure. Based on the experimental results, MCBFS has significant advantages in terms of classification performance compared with 7 benchmark and 6 state-of-the-art feature selection algorithms." "Our results demonstrate that the objective measure for alcohol consumption is a more informative phenotype than self-reported data in revealing peripheral epigenetic mechanisms of alcohol consumption. The DNAm signature associated with PEth shows greater utility on prediction hazardous alcohol drinking in comparison with self-report phenotype. These findings suggest that DNA methylation in blood is a robust biomarker for alcohol consumption."	"Our study showed that this 4-metabolite panel has potential usefulness in clinical assessments of CLD progression in patients with chronic hepatitis B virus infection." "In this study, we proposed a multi-scale clustering-based feature selection method for gene expression data, MCBFS, which performs clustering and feature weighting in a supervised manner. In the algorithm, a multi-scale distance function designed by us was used as a dissimilarity measure. Based on the experimental results, MCBFS has significant advantages in terms of classification performance compared with 7 benchmark and 6 state-of-the-art feature selection algorithms." "Our results demonstrate that the objective measure for alcohol consumption is a more informative phenotype than self-reported data in revealing peripheral epigenetic mechanisms of alcohol consumption. The DNAm signature associated with PEth shows greater utility on prediction hazardous alcohol drinking in comparison with self-report phenotype. These findings suggest that DNA methylation in blood is a robust biomarker for alcohol consumption."
342	Xu, D and Zhang, J and Xu, H and Zhang, Y and Chen, W and Gao, R and Dehmer, M	Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data	BMC Genomics	21	1	650-650	2020	China	http://dx.doi.org/10.1186/s12864-020-07038-3	article	see Table 1	Differential diagnosis prediction	accuracy (10-fold CV)	cross-validation		
343	Xu, K and Liang, X and Justice, A and So-Armah, K and Krystal, J and Sinha, R	DNA methylation biosignature in blood predicts alcohol consumption in Two distinct populations	Neuropsychopharmacology	44		509-510	2019	USA	http://dx.doi.org/10.1038/s41386-019-0547-6	meeting abstract	"In this study, we aimed to select DNAm signatures in blood to predict HAD from two demographically and clinically distinct populations (Ntotal =1,530)" "In this study, we enrolled 70 patients that were diagnosed with non-ischemic heart failure at Second Hospital of Jilin University, Changchun, China, From January 2018 to August 2018 [...] In addition, 77 matched control subjects without heart failure were used as control"	Surrogate biomarker study	AUC, correlation (training + external cohort)	external cohort validation	"Our results demonstrate that the objective measure for alcohol consumption is a more informative phenotype than self-reported data in revealing peripheral epigenetic mechanisms of alcohol consumption. The DNAm signature associated with PEth shows greater utility on prediction hazardous alcohol drinking in comparison with self-report phenotype. These findings suggest that DNA methylation in blood is a robust biomarker for alcohol consumption."	"Our results demonstrate that the objective measure for alcohol consumption is a more informative phenotype than self-reported data in revealing peripheral epigenetic mechanisms of alcohol consumption. The DNAm signature associated with PEth shows greater utility on prediction hazardous alcohol drinking in comparison with self-report phenotype. These findings suggest that DNA methylation in blood is a robust biomarker for alcohol consumption."
344	Yan, Y and Song, D and Zhang, X and Hui, G and Wang, J	GEO Data Sets Analysis Identifies COX-2 and Its Related Micro RNAs as Biomarkers for Non-Ischemic Heart Failure	Frontiers in Pharmacology	11			2020	China	http://dx.doi.org/10.3389/fphar.2020.01156	article	"In this study, we enrolled 70 patients that were diagnosed with non-ischemic heart failure at Second Hospital of Jilin University, Changchun, China, From January 2018 to August 2018 [...] In addition, 77 matched control subjects without heart failure were used as control"	Case-control study	AUC (boostrap analysis)	bootstrapping	"In this study, we used the mining strategy to identify the COX-2 and it micro RNAs, which might be used as biomarkers for non-ischemic heart failure. Although the miR-4649 and miR-1297 are predicted to target the COX-2, their correlations were week. Further studies were need to confirm that their direct correlations. In addition, the sample size was small and analyzed from a single-center, larger studies are needed to confirm the current results."	"In this study, we used the mining strategy to identify the COX-2 and it micro RNAs, which might be used as biomarkers for non-ischemic heart failure. Although the miR-4649 and miR-1297 are predicted to target the COX-2, their correlations were week. Further studies were need to confirm that their direct correlations. In addition, the sample size was small and analyzed from a single-center, larger studies are needed to confirm the current results."
345	Yuan, R and Chen, S and Wang, Y	Computational Prediction of Drug Responses in Cancer Cell Lines From Cancer Omics and Detection of Drug Effectiveness Related Methylation Sites	Frontiers in Genetics	11			2020	China	http://dx.doi.org/10.3389/fgene.2020.00917	article	"GDSC provides 19,100 gene mutations in 1,001 cancer cell lines" "A dataset of 216 HNSCC patients was derived from the Cancer Genome Atlas (TCGA) with information of clinical characteristics, genetic mutation, RNA sequencing, protein expression and histopathological images. Patients were randomly assigned into training (n = 108) or validation (n = 108) sets"	Drug response prediction	AUC (5-foldCV + external validation)	cross-validation + external cohort validation	"In summary, this study indicates the important role of DNA methylation in prediction of drug response, and reveals methylation sites related to drug effectiveness. The database and literature searches on those methylation sites offers a possible mechanism of DNA methylation in regulation of drug effectiveness."	"In summary, this study indicates the important role of DNA methylation in prediction of drug response, and reveals methylation sites related to drug effectiveness. The database and literature searches on those methylation sites offers a possible mechanism of DNA methylation in regulation of drug effectiveness."
346	Zeng, H and Chen, L and Huang, Y and Luo, Y and Ma, X	Integrative Models of Histopathological Image Features and Omics Data Predict Survival in Head and Neck Squamous Cell Carcinoma	Frontiers in Cell and Developmental Biology	8			2020	China	http://dx.doi.org/10.3389/fcell.2020.553099	article	"the training set consisting of 92 healthy volunteers and 149 ischemic stroke patients (including 117 first-ever ischemic stroke patients and 32 recurrent ischemic stroke patients), and the validation set consisting of 30 healthy volunteers and 48 ischemic stroke patients (including 38 firstever ischemic stroke patients and 10 recurrent ischemic stroke patients)" "We identified 453 primary TNBCs from three publicly-available datasets and characterized each as rrTNBC, lrTNBC, or nrTNBC. We compiled primary tumor clinical and multi-omic data, including transcriptome (n=453), copy number alterations (CNAs; n=317), and mutations in 171 cancer-related genes (n=317), then calculated expression and immune signatures"	Case-control study	AUC (10-fold CV + validation)	cross-validation + test set	"The results indicated that histopathological image features had potential as significant prognostic biomarkers for overall survival in patients with HNSCC. The integrative models of genomics, transcriptomics, and proteomics along with histopathological image features may more accurately predict survival outcome than single-omics models, which might contribute to the risk stratification and personalized treatment for cancer patients."	"The results indicated that histopathological image features had potential as significant prognostic biomarkers for overall survival in patients with HNSCC. The integrative models of genomics, transcriptomics, and proteomics along with histopathological image features may more accurately predict survival outcome than single-omics models, which might contribute to the risk stratification and personalized treatment for cancer patients."
347	Zhang, L and Ma, F and Qi, A and Liu, L and Zhang, J and Xu, S and Zhong, Q and Chen, Y and Zhang, C and Cai, C	Integration of ultra-high-pressure liquid chromatography-tandem mass spectrometry with machine learning for identifying fatty acid metabolite biomarkers of ischemic stroke	Chem Commun (Camb)	56		49 6656-6659	2020	China	http://dx.doi.org/10.1039/d0cc02329a	article		Case-control study	AUC (training + test set)	training + test set	"we develop an optimal model to discriminate ischemic stroke patients from healthy persons with 100% sensitivityand 93.18% specificity. This research may facilitate understanding the roles of fatty acid metabolites in stroke occurrence, holding great potential in clinical stroke diagnosis."	"we develop an optimal model to discriminate ischemic stroke patients from healthy persons with 100% sensitivityand 93.18% specificity. This research may facilitate understanding the roles of fatty acid metabolites in stroke occurrence, holding great potential in clinical stroke diagnosis."
348	Zhang, Y and Nock, W and Wyse, M and Weber, Z and Adams, E J and Sarah, A and Stockard, S and Tallman, D and Singh, J and Bae, J and Winer, E P and Lin, N U and Jiang, Y Z and Ma, D and Wang, P and Shi, L and Huang, W and Shao, Z M and Verschraegen, C and Cherian, M and Lustberg, M B and Ramaswamy, B and Sardesai, S and VanDeusen, J and Williams, N and Robert, W and Stover, D G	Machine learning predicts rapid relapse in triple negative breast cancer	Cancer Research	80	4		2020	USA	http://dx.doi.org/10.1158/1538-7445.SABCS19-P4-05-02	meeting abstract	"We identified 453 primary TNBCs from three publicly-available datasets and characterized each as rrTNBC, lrTNBC, or nrTNBC. We compiled primary tumor clinical and multi-omic data, including transcriptome (n=453), copy number alterations (CNAs; n=317), and mutations in 171 cancer-related genes (n=317), then calculated expression and immune signatures"	Prognostic study	AUC (train + test + external validation)	external cohort validation	"We provide a new approach to define TNBCs [triple-negative breast cancer] based on timing of relapse. We identify distinct clinical and genomic features that can be incorporated into machine learning models to predict rrTNBC [rapid relapse TNBCs]." "all the identified top-ranked qualitative biomarkers and quantitative rules are correlated with the identified COVID-19-associated pathogenesis and contribute to distinguishing COVID-19-infected cases from other respiratory patients with or without virus infection, validating the efficacy and accuracy of our prediction. Therefore, the application of machine learning model may efficiently assist in the identification of potential diagnostic biomarkers and candidate drug targets and help establish a standard workflow for related analyses in such field." "We identified 8 lncRNAs as potential diagnostic biomarkers for NSCLC through integrative cross-platform data analyses. This data mining and machine learning approach would be an efficient and economical screening method for tumor biomarker discovery"	"We provide a new approach to define TNBCs [triple-negative breast cancer] based on timing of relapse. We identify distinct clinical and genomic features that can be incorporated into machine learning models to predict rrTNBC [rapid relapse TNBCs]." "all the identified top-ranked qualitative biomarkers and quantitative rules are correlated with the identified COVID-19-associated pathogenesis and contribute to distinguishing COVID-19-infected cases from other respiratory patients with or without virus infection, validating the efficacy and accuracy of our prediction. Therefore, the application of machine learning model may efficiently assist in the identification of potential diagnostic biomarkers and candidate drug targets and help establish a standard workflow for related analyses in such field." "We identified 8 lncRNAs as potential diagnostic biomarkers for NSCLC through integrative cross-platform data analyses. This data mining and machine learning approach would be an efficient and economical screening method for tumor biomarker discovery"
349	Zhang, Y H and Li, H and Zeng, T and Chen, L and Li, Z and Huang, T and Cai, Y D	Identifying Transcriptomic Signatures and Rules for SARS-CoV-2 Infection	Frontiers in Cell and Developmental Biology	8			2020	China	http://dx.doi.org/10.3389/fcell.2020.627922	article	"A total of 93 patients were infected with SARS-COV-2, 100 patients with other viruses, and 41 patients without viral infection"	Case-control study	Matthews Correlation Coefficient (10-fold CV)	cross-validation	"In this review, we survey current knowledge-guided statistical learning methods, including both supervised learning and unsupervised learning, and their applications to precision oncology, and we discuss future research directions."	"In this review, we survey current knowledge-guided statistical learning methods, including both supervised learning and unsupervised learning, and their applications to precision oncology, and we discuss future research directions."
350	Zhao, T and Khadka, V S and Deng, Y	Identification of lncRNA biomarkers for lung cancer through integrative cross-platform data analyses	Aging (Albany NY)	12	14	14506-14527	2020	USA	http://dx.doi.org/10.18632/aging.103496	article	"Lung cancer datasets were obtained from the Gene Expression Omnibus (GEO, n = 287) and The Cancer Genome Atlas (TCGA, n = 216) repositories"	Case-control study	AUC (training + external validation)	external cohort validation	"Plasma cfDNA methylome profiling identified effective biomarkers for the detection and tissue-of-origin determination of GI cancers, and outperformed mutation-based detection approach. Therefore, a liquid biopsy test capable of detecting and locating GI cancers is feasible and may serve as a valuable tool for early detection and intervention."	"Plasma cfDNA methylome profiling identified effective biomarkers for the detection and tissue-of-origin determination of GI cancers, and outperformed mutation-based detection approach. Therefore, a liquid biopsy test capable of detecting and locating GI cancers is feasible and may serve as a valuable tool for early detection and intervention."
351	Zhao, Y and Chang, C and Long, Q	Knowledge-guided statistical learning methods for analysis of high-dimensional -omics data in precision oncology	JCO Precision Oncology	3			2019	USA	http://dx.doi.org/10.1200/PO.19.00019	article	review (not applicable)	Review				
352	Zhuang, H and Chen, Y and Sheng, X and Hong, L and Gao, R and Zh, X	Searching for a signature involving 10 genes to predict the survival of patients with acute myelocytic leukemia through a combined multi-omics analysis	PeerJ	8	e9437		2020	China	http://dx.doi.org/10.7717/peerj.9437	article	"all samples (n = 229) were randomized as test set and training set, respectively"	Prognostic study	AUC, log-rank p-value (training + external test set)	external cohort validation	"Our study examined the expression patterns in AML samples from the GEO and TCGA databases. We constructed a reliable OS-related 10-gene signature that was not dependent on clinical parameters."	"Our study examined the expression patterns in AML samples from the GEO and TCGA databases. We constructed a reliable OS-related 10-gene signature that was not dependent on clinical parameters."