BMJ Open

Reporting quality of studies using machine learning models for medical diagnosis: a systematic review.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-034568
Article Type:	Original research
Date Submitted by the Author:	27-Sep-2019
Complete List of Authors:	Yusuf, Mohamed; Manchester Metropolitan University, Health Professions Atal, Ignacio; Université Paris Descartes, Centre for Research and Interdisciplinarity (CRI); Université Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Li, Jacques; Universite Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Smith, Philip; Manchester Metropolitan University, Health Professions Ravaud, Philippe; Universite Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Fergie, Martin; The University of Manchester, Imaging and Data Sciences Callaghan, Michael; Manchester Metropolitan University, Health Professions Selfe, James; Manchester Metropolitan University, Health Professions
Keywords:	Machine learning, Medical diagnosis, Clinical prediction, Reporting quality

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: Reporting quality of studies using machine learning models for medical diagnosis: a systematic review

Corresponding author:
Mohamed Yusuf
Department of Health Professions
Manchester Metropolitan University
Manchester, UK
M.Yusuf@mmu.ac.uk

Ignacio Atal
Research Fellow
Centre for Research and Interdisciplinarity (CRI),
Université Paris Descartes

Jacques Li MD
Medical Doctor/Research intern in evidence-based medicine
INSERM, Methods
Université Paris Descartes
Paris, France

Philip Smith
Senior lecturer
Department of Health Professions
Manchester Metropolitan University
Manchester, UK

Philippe Ravaud
Professor in Epidemiology
INSERM, Methods
Université Paris Descartes
Paris, France

Martin Fergie
Lecturer in Health sciences
Division of Informatics,
Imaging and Data Sciences,
University of Manchester
Manchester UK

Michael Callaghan
Professor
Department of Health Professions
Manchester Metropolitan University
Manchester, UK

James Selfe
Professor
Department of Health Professions
Manchester Metropolitan University
Manchester, U

Abstract

Aims: We conducted a systematic review assessing the reporting quality of studies validating models based on machine learning (ML) for clinical diagnosis, with a specific focus on the reporting of information concerning the participants on which the diagnostic task was evaluated on.

Method: Medline Core Clinical Journals were searched for studies published between July 2015 to July 2018. Two reviewers independently screened the retrieved articles, a third reviewer resolved any discrepancies. An extraction list was developed from the TRIPOD guideline. Two reviewers independently extracted the data from the eligible articles. Third and fourth reviewers checked, verified the extracted data as well as resolved any discrepancies between the reviewers.

Results: The search results yielded 161 papers, of which 28 conformed to the eligibility criteria. Detail of data source was reported in 86% of the papers. For all of the papers, the set of patients on which the ML-based diagnostic system was evaluated was partitioned from a larger dataset, and the method for deriving such set was always reported. Information on the diagnostic/non-diagnostic classification was reported well (82%). The least reported items were the use of reporting guideline (0%), distribution of disease severity (29%), patient flow diagram (34%) and distribution of alternative diagnosis (36%). A large proportion of studies (82%) had a delay between the conduct of the reference standard and ML tests, while 4% did not and 14% were unclear. For 54% of the studies, it was unclear whether the evaluation group corresponded to the setting in which the ML test will be applied to.

Conclusion: We found that all eligible studies in this review failed to use reporting guidelines and the studies lacked adequate detail on the participants on which the diagnostic task was evaluated on, thus making it difficult to replicate, assess and interpret study findings.

Review registration number: CRD42018099167 **word count**: 3,060

Keywords: Machine learning, Medical diagnosis, Reporting quality

Strengths and limitations of this study

- Machine learning (ML) is a rapidly developing area and due to access in large amounts of clinical data and the development of new ML techniques has led to a rise in the application of ML methods to medicine.
- Within the medical research world, there is an already established framework to guide researchers in the conduct of diagnostic accuracy studies.
- We found that studies developing or validating ML models for clinical diagnosis, failed to use reporting guidelines and lacked adequate detail on the participants on which the diagnostic task was evaluated on.
- Most studies failed to report study participant flow diagram, distribution of disease severity and distribution of alternative diagnosis.
- This review highlights the need for an evidence-based medicine framework to aid the conduct and reporting of ML methods in medical diagnosis.

Introduction

Machine learning (ML) is a rapidly developing area, characterised as the science of training computers to conduct specific tasks such as classification or prediction without explicit programming, but where decisions are taken based on patterns and relationships within large and complex datasets [1]. Over the past decade, access to large amounts of clinical data and the development of new ML techniques has led to a rise in the application of ML methods to medicine (Cleophas and Zwinderman, 2015; Topol, 2019. Due to their propensity to facilitate and promote timely and objective clinical decision-making, ML methods have been applied to gain valuable insights into clinical diagnoses. For example, ML methods have been used to diagnose skin cancer using skin lesion images [2], diagnose cerebral aneurysms using clinical notes [3] and diagnose stroke using neuroimaging data [4]- see *box 1* for an example of ML-based diagnostic system.

A key principle of ML models is that they are developed based on the automatic extraction of patterns from data, instead of relying upon explicit rules to make decisions. Generally, a researcher developing a ML model has access to a large dataset that is divided into a training set and a test set (see Box 1). The training set is used to develop a ML model that will learn the relationships between available clinical data and an outcome of interest (e.g. a diagnosis). The performance of the ML model is then evaluated by applying it to the test set.

Box 1

Machine learning is the ability to create algorithms to accomplish specific tasks without explicitly programming them, but rather take decisions based on previously seen data. Here is a summary of the steps when creating machine learning algorithms:

Model Development

Step 1: Defining the research problem. This could be broken down to either a **classification**, **regression** or a **clustering problem**.

Step 2: Identification of data sources and formats. Data could be in various formats (e.g. images, text, speech or numerical), and data could come from various sources such as hospital, insurance databases, or previous research projects.

Step 3: Training and test set derivation. Here the data could be broken down into two independent components: the **training set** and **test set**. The training set is used to create the ML algorithm, and the test set is used to evaluate the ML algorithm.

Step 4: Model development. The model is developed using the training dataset. The model could be either **supervised** or **unsupervised** (supervised models require labelled data whereas unsupervised models do not). The **loss function** and the methods for handling **outliers** and **missing data** are also described. A portion of the training set, the **model selection set**, is often withheld from model training to allow for model selection and to avoid overfitting.

Step 5: Evaluation of the model. The test set is used to evaluate the ML algorithm using a variety of metrics to compare the prediction with the gold standard outcome label (often referred to as **ground-truth**).

Model Validation

To obtain an accurate assessment of model's performance in a clinical setting, the model must be validated against data which is drawn from a clinical cohort. **Internal validation** refers to a model being evaluated on a cohort taken from the same setting as the data used to develop the model. **External validation** is where the cohort data is taken from a separate setting, which overcomes any systematic biases present in the data source used for model validation.

It is worth noting that one potential area for confusion is the differing meanings of the terms **test set** and **validation set** between the machine learning and medical research community. A medical researchers validation set is a machine learning test set.

Despite their popularity, the promising applications of ML-based diagnostic systems come with its own set of pitfalls. Studies using ML for medical diagnosis may contain systematic errors in both the design and execution [5-7]. For instance, selection bias can occur if the sample used to produce the ML-based diagnostic system is not entirely representative of the population on which the model may be used in the future [6]. Repeated evaluation of model performances against the same *test set* may result in the selected model overfitting the test set, resulting in an over-optimistic assessment of model performance [8]. These methodological biases can make it difficult to generalise conclusions from the results yielded. This could lead to erroneous yet devastating clinical decisions, i.e. recommending a medical treatment to an individual that is different from those in the population the that treatment was developed and validated on [9].

There is a parallel between what ML researchers refer to as 'test set' and the 'Population on which a diagnostic test is evaluated' within diagnostic accuracy studies. The diagnostic accuracy of an ML-based systems is reliant on demographic and clinical characteristics of the population in which it was applied on, therefore, if the cohorts are not a representative sample of the targeted population, then the generalisability of the study results may be limited. A further hindrance to

the application of ML methods for medical diagnosis (and more generally in biomedical research) is that ML researchers may not be familiar with the requirements and guidelines that biomedical research have collectively established to ensure transparent and unbiased evidence-based knowledge accumulation [10, 11].

Clinical predictions models undergo a scientifically rigorous process to establish their diagnostic accuracy, which encompasses their safety, validity, reproducibility, usability, and reliability. The importance of transparent and rigorous reporting of clinical predictions models accuracy studies. Particularly as the diagnostic prediction models of an instrument can vary greatly due to factors such as population characteristics, clinical setting, disease prevalence and severity as well as aspects of test execution and interpretation[11]. To aide with and standardise this process, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – (TRIPOD) guideline was set in place. The TRIPOD is an internationally accepted reporting guideline that was developed to improve the reliability and value of clinical prediction models through the promotion of transparent and accurate reporting [11].

In medical research, reporting guidelines are implemented to aid in the transparent evaluation, usability and reproducibility of a diagnostic instrument [12]. Luo *et al.* [13] have constructed a reporting guideline for the development and usage of ML predictive models in biomedical research. This is an important step towards a rigorous and robust approach to the usage of ML methods in medical research. Since publication in December 2016 (up to May 2019), the guideline of Luo *et al*., which is currently available on The EQUATOR Network website [14], has garnered only ~50 citations. Additionally, in 2015 a more specific and robust guideline was developed to aid the reporting of prediction models used in prognostic and diagnostic studies (TRIPOD) [11]. TRIPOD has ~1,000 citations demonstrating that is has been accepted by the community as a useful set of guidelines for diagnostic/prognostic prediction. In this work we evaluate whether ML studies make use of these guidelines.

To date, there have been no studies evaluating the reporting quality of studies using ML methods, particularly diagnostic studies. Knowing this may aid in the evaluation of reporting standards employed by ML researchers. In this review, we focus on medical research studies that used ML methods to aid clinical diagnosis. Further, we have narrowed our review to applied ML methods which are envisaged to be clinically useful, in which the end users are practitioners and research consumers.

We aimed to produce a systematic review assessing the reporting quality of studies developing or validating ML models for clinical diagnosis, with a specific focus on the reporting of information concerning the participants on which the diagnostic task was evaluated.

Methods

This review was registered with International prospective register of systematic reviews (PROSPERO) on 30/07/2018 (reference: CRD42018099167). The framework used for this

methodological systematic review is Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guideline for Systematic reviews [15].

Literature search

On July 2018, two authors (MY & JL) independently searched through the Medline Core Clinical Journals for articles developing or validating ML models for clinical diagnosis. Core Clinical Journals, also known as Abridged Index Medicus (AIM), is a filter option within Medline that limits to clinically useful journals. This is a selection of 119 English-language journals that focus on clinical studies and that are considered to be of immediate interest to practising physicians [16]. Using this filter excludes journals in bioinformatics or computational biology, which are highly likely to include articles explaining the development of ML-based diagnosis systems. However, these journals might not target clinicians. In addition to this, due to the ever-expanding ML literature, we have narrowed our review to studies published between July 2015 to 1 July 2018. See supplementary file 1 for the search strategy.

Subsequent to the literature search, the two reviewers (MY & JL) screened the title and abstracts of the search results. Once the eligible papers were identified and retrieved, both the first reviewer (MY) and second reviewer (JL) independently screened the full articles for eligibility. Discrepancies between the two reviewers were discussed and resolved by a third reviewer (IA).

Inclusion and exclusion

Studies were included if they used ML for clinical diagnosis, for example if they used statistical techniques to conduct classification, regression or clustering based on clinical data for disease diagnosis without being explicitly programmed. Other inclusions were primary study designs that evaluated the accuracy of such ML-based systems for diagnostic tasks, and articles in the English language. Studies were excluded if they did not report original research, if they were systematic reviews, had no abstract or did not specify the type of ML model adopted.

Extraction list

An extraction list based on the TRIPOD guideline was developed. The focus of the extraction list was to extract information about the participants upon which the diagnostic task was evaluated on, namely selection method and population characteristics. We additionally extracted general information concerning the diagnostic tasks, namely the target condition and the target population. The extraction list was tested and validated by two reviewers (MY & JL), by applying it on a random sample of the eligible papers.

Data extraction

Two reviewers (JL & PS) independently extracted the data from the eligible articles based on the items listed in *Table 1*. For each of the items, reviewers declared if the item was clearly reported (yes/no/unclear) and justify the declaration by citing the manuscript verbatim, as well as providing a written explanation if the reporting was considered unclear. A third and fourth reviewer (MY &

IA) checked and verified the extracted data and resolved any disagreements between the reviewers through discussion.

Data analysis

Findings from the included studies demonstrating study characteristics, reporting quality and presence of bias were presented in descriptive statistics and figures.

Patient and public involvement

There was no patient or public involvement in any phase of this study, this included the development of the research question, the analysis and the conclusions.

Table 1: Item list used to extract eligible papers.

Item groups	Item list	Detailed items
General characteristics	Diagnostic task	What is the target condition?
	Study objective	Is the study aiming at the development of a diagnostic method, evaluation of a diagnostic method, or both?
	Target population	What is the population targeted by the diagnostic test?
Methods	Data sources	Where and when potentially eligible participants were identified (setting, location and dates)
	Data split	Method for partitioning the evaluation set from the training data. To assess whether participants formed a consecutive, random or convenience series.
	Test dataset eligibility criteria	On what basis potentially eligible participants were identified within the test dataset (such as symptoms, results from previous tests, inclusion in registry).
Results	Baseline characteristics	Baseline demographic and clinical characteristics of participants
	Diagnosis/non-diagnosis classification	Classification of the diagnosed and non-diagnosed patients within the test set.
	Flow diagram	Flow of participants, using a diagram.
	Severity	Distribution of severity of disease in those with the target condition.
	Alternative diagnosis	Distribution of alternative diagnoses in those without the target condition.
	Difference between reference test and ML test.	Is there a time interval between index test and reference standard?
	Applicability	Does the evaluation population correspond to the setting in which the diagnosis test will be applied?

Results

The search yielded 161 papers, of which 28 conformed to the eligibility criteria, see *figure 1*. During the screening of the title and abstract, most papers were excluded due to the search term 'CAD' being analogous to both 'Computer-aided Detection' and Coronary Artery Disease'. During the full text review, eleven papers were excluded because they did not use ML methods for medical diagnosis, and three papers were excluded because they did not use ML method but were captured in the search because they studied Coronary Artery Disease (CAD).

Figure 1: PRISMA Flow diagram

Study characteristics

The study characteristics of the all eligible studies are presented in Table 2 (see supplementary file 1 for list of studies). From the papers extracted, majority of the studies were published in 2017 (43%) and mostly in the Radiology journal (29%). Oncology was the most researched domain (47%), followed by Neurology (18%). The majority of studies focused on model development (97%), with only one study looking at model validation

Table 2: Study characteristics

Items	Total n (%)
Year	
2015	4 (14)
2016	9 (32)
2017	12 (43)
2018	3 (11)
Journals	
Radiology	8 (29)
Jama	2 (7)
Brain	2 (7)
American Journal of Roentgenology	2 (7)
Neurology	1 (3)
Medicine	1 (3)
Surgery	1 (3)
Chest	1 (3)
Gastroenterology	1 (3)
Journal of the American College of Cardiology	1 (3)
Journal of Allergy and Clinical Immunology	1 (3)
American journal of clinical pathology	1 (3)
American journal of ophthalmology	1 (3)
The Journal of infectious diseases	1 (3)
Digestive diseases and sciences	1 (3)
The British journal of radiology	1 (3)
The Journal of pediatrics	1 (3)
Clinical Specialty	
Oncology	13 (47)
Neurology	5 (18)
Immunology	2 (7)
Ophthalmology	2 (7)
Others specialties[a]	6 (21)
Task	
Development and Evaluation	27 (97)
Evaluation	1 (3)

[a] Other clinical specialities include Cardiology, Gastroenterology, Infectious disease, Psychiatry, Endocrinology and Various.

Reporting quality

Detail of the data source was reported in 86% of the papers, with all studies providing information on the separation method for deriving the evaluation set from the larger dataset. 82% of studies reported eligibility criteria for both evaluation set. Information on the diagnostic/non-diagnostic classification evaluation metric used was included in 82% of all papers. The least reported items were use of reporting guideline (0%), distribution of disease severity (29%), patient flow diagram (34%), distribution of alternative diagnosis (36%) and baseline characteristic (64%). See table 3 for a full breakdown of reporting quality.

Table 3: Reporting quality

Items	Reported n (%)	Not reported n (%)	Unclear n (%)
Methods			
Data source	24 (86)	0 (0)	4 (14)
Data split methods	28 (100)	0 (0)	0 (0)
Test set eligibility criteria (evaluation set)	23 (82)	5 (18)	0 (0)
Results			
Baseline characteristic	17 (61)	11 (39)	0 (0)
Diagnosis/non-diagnosis classification	23 (82)	4 (14)	1 (4)
Flow diagram	10 (36)	18 (64)	0 (0)
Disease severity	8 (29)	18 (64)	2 (7)
Alternative diagnosis	10 (36)	18 (64)	0 (0)
Use of reporting guideline	0 (0)	28 (100)	0 (0)

N = number

Presence of bias

Within the eligible studies, 71% did not report a time interval between the conduct of the reference standard and ML test (table 4). Within 54% of studies, it was unclear whether the study populations corresponded to the setting in which the diagnostic test will be applied to. However, in 29% of studies, the clinical setting of the gathered evaluation dataset did not correspond to the clinical setting in which study authors hoped it would be applied.

Table 4: Presence of bias

Items	Yes n (%)	No n (%)	Unclear n (%)
Is there a time interval between reference standard ML test?	23 (82)	1 (4)	4 (14)
Does the test population correspond to the population/setting in which the diagnosis test will be applied?	5 (18)	8 (29)	15 (54)

Discussion

This review found that studies developing or validating ML-based systems for clinical diagnosis failed to use reporting guidelines and lacked adequate detail for assessment, interpretation and reproducibility. With nearly all studies providing detail on data sources, eligibility criteria and diagnosis classification, only a few studies reported study participant flow diagram, distribution of disease severity, distribution of alternative diagnosis. Our findings are in line with those of Faes et al. recent systematic reviews [17] in which they found poor reporting and potential biases arising from study design in studies using ML methods for classifying diseases from medical imaging. Similarly, in another systematic review, Christodoulou et al' [18] found studies comparing the performance of logistic regression models with ML models for clinical prediction to have poor methodology and reporting quality.

A high number of studies reviewed had a time difference between the conduct of the reference test and that of ML-based diagnostic systems, suggesting the potential for incorporation bias [19, 20]. This is largely an issue in ML-based diagnostic systems where labelling is the gold-standard, but patient data is labelled retrospectively. This may happen several years after initial data collection and in a different setting.

Though unclear in majority of studies, there was some evidence suggesting the study test populations did not correspond to the populations in which tests were hoped to be applied to, further limiting their generalisability. In addition to this, studies utilising ML-based diagnostics systems fail to report baseline characteristics. This could be problematic; within diagnostic accuracy studies it is imperative to report sample characteristics as this aid researchers, research consumers and practitioners in determining the relevancy and applicability of study findings to a wider setting.

Another vital element in diagnostic accuracy studies is the use of different methods to derive the evaluation sample from the wider population; this could lead to more or less accurate estimation of the diagnostic performance. The ideal method for sampling should be based on probability and not convenience, as this allows for a representative sample to be selected from a sampling frame whereby all eligible individuals have an equal chance of being selected. In addition to this, ML-based diagnostic systems that are evaluated using internal validation, where the evaluation set is partitioned from the same cohort as the training set, risk learning the systematic biases in the data of the particular centre from which the cohort was drawn. Such methods only address the systems internal validity, and model performance may deteriorate when deployed on an cohort drawn from a different centre [21]. External validation, where the ML-based diagnostic systems are evaluated on a cohort that has played no role in model development, is an important step to verify and asses the whether the system is reliable and deployable on potential populations for clinical use [22-24]. This is further highlighted in a recent systematic review evaluating the performance of ML algorithms for the diagnostic analysis of medical images; within this review, Kim et al. [25] found only 6% (31 out of 516 studies) had externally validated their algorithms.

Low-quality clinical research that is reported inadequately or that offers invalid data and distorted outcomes are deemed wasteful; such research is non-replicable and unusable [26, 27]. One way to increase the value and reusability of these novel and promising ML-based systems is through complete, accurate and transparent reporting [14, 27].. Some of the methodological and reporting issues facing the studies reviewed in this systematic review can be mitigated through the use of reporting guidelines such as TRIPOD guideline. More specifically, there has been a recent initiative to develop an extension of the TRIPOD statement which is specific to ML studies (TRIPOD-ML)[28]. Such guidelines aid researchers developing ML-based diagnostic systems in addressing the important aspects of design, execution and complete and reliable reporting of studies. However, no reporting guideline can salvage a poorly designed and executed study. To prevent flawed design and execution of ML methods, informatic and biomedical researchers looking to develop ML-based diagnostic systems should consult with a methodologist, epidemiologist or a statistician. Having input from such experts will aid in research that is methodological robust in design and execution - resulting in research that it is reliable, reproducible and that adds scientific value [29].

Strengths and limitations of study

To our knowledge, this is the first systematic review evaluating the reporting quality of studies developing and/or validating ML methods for medical diagnosis within the medical literature. However, it is worth noting that we have not included all medical journals and therefore our findings may not be applicable to all journals. Despite this, we have included studies published within the Medline Core Clinical Journals, these journals cover all areas of clinical and public health.

This review did not evaluate the statistical methodology and conduct of studies using ML diagnostic systems. This could be considered a limitation as a transparent reporting does not guarantee a quality study. Nevertheless, this review shows that these studies do not comply with TRIPOD guideline on the reporting this considerably affects the trust we have in the estimates they are giving concerning the efficacy of their diagnostic methods.

Conclusion

We found that all eligible studies in this review failed to use reporting guidelines and the studies lacked adequate detail on the participants on which the diagnostic task was evaluated on, thus making it difficult to replicate, assess and interpret study findings.

Diagnostic studies using ML methods have great potential to improve clinical decision-making and take the load off health systems. However, studies with poor reporting can be more problematic than of help. Within biomedical research, there is an already established framework and guidelines in which ML researchers can utilise to aid the execution and reporting of ML methods for clinical diagnosis, with the TRIPOD guideline being the most robust and widely used.

Abbreviation: ML: Machine learning, CAD: Computer Aided Diagnosis, TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Legends: * Other clinical specialities include Cardiology, Gastroenterology, Infectious disease, Psychiatry, Endocrinology and Various.
Figure 1 is the PRISMA 2009 Flow diagram.

Contributors: JS is a guarantor of this review. All authors have made substantive intellectual contributions to the development of this review. MY and IA were involved in conceptualising the review. MY, IA and PR developed the protocol. MY, JL & IA did the literature search, MY, JL, PS & IA carried out the study selection and data extraction. MY, IA, JS, PS, MF and MC were involved in the writing and editing of the manuscript.

Transparency: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and registered) have been explained.

Funding: There is no funding attached to this systematic review. This review was conducted independently by the research team.

Ethical approval: Not required as this is a review

Data sharing: All data are freely available within the appendices. No additional available.

Competing interest: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

References

1. Paliwal M, Kumar UA: **Neural networks and statistical techniques: A review of applications**. *Expert systems with applications* 2009, **36**(1):2-17.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature* 2017, **542**(7639):115.
3. Shin H, Kim KH, Song C, Lee I, Lee K, Kang J, Kang YK: **Electrodiagnosis support system for localizing neural injury in an upper limb**. *Journal of the American Medical Informatics Association* 2010, **17**(3):345-347.

4. Rehme AK, Volz LJ, Feis D-L, Bomilcar-Focke I, Liebig T, Eickhoff SB, Fink GR, Grefkes C: **Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques**. *Cerebral cortex* 2014, **25**(9):3046-3056.
5. Park SH, Han K: **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction**. *Radiology* 2018, **286**(3):800-809.
6. Bone D, Goodwin MS, Black MP, Lee C-C, Audhkhasi K, Narayanan S: **Applying machine learning to facilitate autism diagnostics: pitfalls and promises**. *Journal of autism and developmental disorders* 2015, **45**(5):1121-1136.
7. Foster KR, Koprowski R, Skufca JD: **Machine learning, medical diagnosis, and biomedical engineering research-commentary**. *Biomedical engineering online* 2014, **13**(1):94.
8. Subramanian J, Simon R: **Overfitting in prediction models–is it a problem only in high dimensions?** *Contemporary clinical trials* 2013, **36**(2):636-641.
9. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W: **Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users**. *Statistics in medicine* 2008, **27**(11):1801-1813.
10. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HCW *et al*: **STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration**. *BMJ Open* 2016, **6**(11):e012799.
11. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS: **Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration**. *Annals of internal medicine* 2015, **162**(1):W1-W73.
12. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG: **Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network**. *BMC medicine* 2010, **8**(1):24.
13. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB: **Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view**. *Journal of medical Internet research* 2016, **18**(12).
14. Simera I, Moher D, Hoey J, Schulz KF, Altman DGJM: **The EQUATOR Network and reporting guidelines: Helping to achieve high standards in reporting health research studies**. 2009, **63**(1):4-6.
15. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D: **The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration**. *BMJ* 2009, **339**.
16. **Committed Selection: Abridged Index Medicus**. *New England Journal of Medicine* 1970, **282**(4):220-221.
17. Faes L, Liu X, Kale A, Bruynseels A, Shamdas M, Moraes G, Fu DJ, Wagner SK, Kern C, Ledsam JR: **Deep Learning Under Scrutiny: Performance Against Health Care Professionals in Detecting Diseases from Medical Imaging-Systematic Review and Meta-Analysis**. 2019.

18. Jie M, Collins GS, Steyerberg EW, Verbakel JY, van Calster B: **A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models**. *Journal of clinical epidemiology* 2019.
19. Kramer MS, Roberts-Bräuer R, Williams RL: **Bias andovercall'in interpreting chest radiographs in young febrile children**. *Pediatrics* 1992, **90**(1):11-13.
20. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J: **Sources of variation and bias in studies of diagnostic accuracy. A systematic review**. *Annals of internal medicine* 2004, **140**(3):189-202.
21. Steyerberg EW: **Clinical prediction models: a practical approach to development, validation, and updating**: Springer Science & Business Media; 2008.
22. Steyerberg E: **Overfitting and optimism in prediction models**. In: *Clinical Prediction Models.* edn.: Springer; 2009: 83-100.
23. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, Collins GS: **External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges**. *bmj* 2016, **353**:i3140.
24. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF: **Internal validation of predictive models: efficiency of some procedures for logistic regression analysis**. *Journal of clinical epidemiology* 2001, **54**(8):774-781.
25. Kim DW, Jang HY, Kim KW, Shin Y, Park SH: **Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers**. *Korean journal of radiology* 2019, **20**(3):405-410.
26. Chan A-W, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, Krumholz HM, Ghersi D, Van Der Worp HBJTL: **Increasing value and reducing waste: addressing inaccessible research**. 2014, **383**(9913):257-266.
27. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E: **Reducing waste from incomplete or unusable reports of biomedical research**. *The Lancet* 2014, **383**(9913):267-276.
28. Collins GS, Moons KG: **Reporting of artificial intelligence prediction models**. *The Lancet* 2019, **393**(10181):1577-1579.
29. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R: **Increasing value and reducing waste in research design, conduct, and analysis**. *The Lancet* 2014, **383**(9912):166-175.

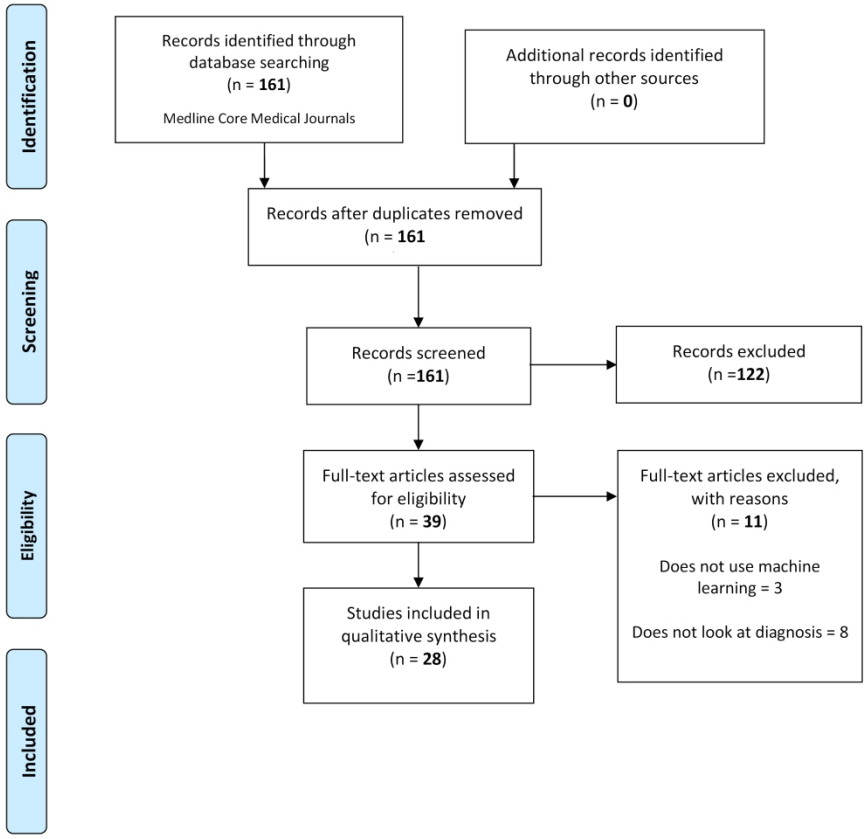

Figure 1: PRISMA 2009 Flow diagram

487x545mm (300 x 300 DPI)

Search strategy using Medline (restricted to Core Clinical Journals)

1) machine learning,
2) supervised learning
3) unsupervised learning
4) deep learning
5) artificial Intelligence
6) decision trees
7) Artificial
8) Neural Network
9) CNN
10) ANN
11) Convolutional Neural Network
12) random forest,
13) reinforcement learning
14) gradient boosting
15) computer aided diagnosis
16) CAD
17) computer assisted diagnosis
18) computational analysis

19) OR/ 1-18

20) Diagnosis

21) 19 AND 20

List of eligible studies

Asaoka R, Hirasawa K, Iwase A, Fujino Y, Murata H, Shoji N, Araie M. Validating the usefulness of the "random forests" classifier to diagnose early glaucoma with optical coherence tomography. American journal of ophthalmology. 2017 Feb 1;174:95-103.

Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. Radiology. 2017 Oct 17;286(3):810-8.

Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. The British journal of radiology. 2018 Jan;91(xxxx):20170576.

Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hermsen M, Manson QF, Balkenhol M, Geessink O. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama. 2017 Dec 12;318(22):2199-210.

Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS. Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology. 2018 Feb 1;154(3):568-75.

Demertzi A, Antonopoulos G, Heine L, Voss HU, Crone JS, de Los Angeles C, Bahri MA, Di Perri C, Vanhaudenhuyse A, Charland-Verville V, Kronbichler M. Intrinsic functional connectivity differentiates minimally conscious from unresponsive patients. Brain. 2015 Jun 27;138(9):2619-31.

Dinh AH, Melodelima C, Souchon R, Moldovan PC, Bratan F, Pagnoux G, Mège-Lechevallier F, Ruffion A, Crouzet S, Colombel M, Rouvière O. Characterization of prostate cancer with Gleason score of at least 7 by using quantitative multiparametric MR imaging: validation of a computer-aided diagnosis system in patients referred for prostate biopsy. Radiology. 2018 Jan 22;287(2):525-33.

Eshaghi A, Wottschel V, Cortese R, Calabrese M, Sahraian MA, Thompson AJ, Alexander DC, Ciccarelli O. Gray matter MRI differentiates neuromyelitis optica from multiple sclerosis using random forest. Neurology. 2016 Dec 6;87(23):2463-70.

Gallego-Ortiz C, Martel AL. Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions. Radiology. 2015 Sep 18;278(3):679-88.

Hao S, Jin B, Tan Z, Li Z, Ji J, Hu G, Wang Y, Deng X, Kanegaye JT, Tremoulet AH, Burns JC. A classification tool for differentiation of Kawasaki disease from other febrile illnesses. The Journal of pediatrics. 2016 Sep 1;176:114-20.

Harper L, Fumagalli GG, Barkhof F, Scheltens P, O'Brien JT, Bouwman F, Burton EJ, Rohrer JD, Fox NC, Ridgway GR, Schott JM. MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. Brain. 2016 Mar 1;139(4):1211-25.

Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, Rust KC. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Digestive diseases and sciences. 2017 Oct 1;62(10):2719-27.

Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, Hussien A, Rathmell J, Thomas B, Chen C, Hales R. Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. Radiology. 2017 Sep 5;286(1):286-95.

Keller MD, Pandey R, Li D, Glessner J, Tian L, Henrickson SE, Chinn IK, Monaco-Shawver L, Heimall J, Hou C, Otieno FG. Mutation in IRF2BP2 is responsible for a familial form of common variable immunodeficiency disorder. Journal of Allergy and Clinical Immunology. 2016 Aug 1;138(2):544-50.

Lee Y, Kim JK, Shim W, Sung YS, Cho KS, Shin JH, Kim MH. Does computer-aided diagnosis permit differentiation of angiomyolipoma without visible fat from renal cell carcinoma on MDCT?. American Journal of Roentgenology. 2015 Sep;205(3):W305-12.

Li M, Narayan V, Gill RR, Jagannathan JP, Barile MF, Gao F, Bueno R, Jayender J. Computer-aided diagnosis of ground-glass opacity nodules using open-source software for quantifying tumor heterogeneity. American Journal of Roentgenology. 2017 Dec;209(6):1216-27.

Lu X, Yang Y, Wu F, Gao M, Xu Y, Zhang Y, Yao Y, Du X, Li C, Wu L, Zhong X. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. Medicine. 2016 Jul;95(30).

Möller C, Pijnenburg YA, van der Flier WM, Versteeg A, Tijms B, de Munck JC, Hafkemeijer A, Rombouts SA, van der Grond J, van Swieten J, Dopper E. Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis. Radiology. 2015 Dec 11;279(3):838-48.

Narula S, Shameer K, Omar AM, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. Journal of the American College of Cardiology. 2016 Nov 29;68(21):2287-95.

Ng DP, Wu D, Wood BL, Fromm JR. Computer-aided detection of rare tumor populations in flow cytometry: An example with classic Hodgkin lymphoma. American journal of clinical pathology. 2015 Sep 1;144(3):517-24.

Silterra J, Gillette MA, Lanaspa M, Pellé KG, Valim C, Ahmad R, Acácio S, Almendinger KD, Tan Y, Madrid L, Alonso PL. Transcriptional categorization of the etiology of pneumonia syndrome in pediatric patients in malaria-endemic areas. The Journal of infectious diseases. 2016 Nov 10;215(2):312-20.

Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, Schneider DF. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. Surgery. 2017 Apr 1;161(4):1113-21.

Sun H, Chen Y, Huang Q, Lui S, Huang X, Shi Y, Xu X, Sweeney JA, Gong Q. Psychoradiologic utility of MR imaging for diagnosis of attention deficit hyperactivity disorder: a radiomics analysis. Radiology. 2017 Nov 22;287(2):620-30.

Ta CN, Kono Y, Eghtedari M, Oh YT, Robbin ML, Barr RG, Kummel AC, Mattrey RF. Focal Liver Lesions: Computer-aided Diagnosis by Using Contrast-enhanced US Cine Recordings. Radiology. 2017 Oct 25;286(3):1062-71.

Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EY. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama. 2017 Dec 12;318(22):2211-23.

Tomlinson GS, Thomas N, Chain BM, Best K, Simpson N, Hardavella G, Brown J, Bhowmik A, Navani N, Janes SM, Miller RF. Transcriptional profiling of endobronchial ultrasound-guided lymph node samples aids diagnosis of mediastinal lymphadenopathy. Chest. 2016 Feb 1;149(2):535-44.

Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology. 2017 Oct 23;286(3):887-96.

Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid–enhanced hepatobiliary phase MR images. Radiology. 2017 Dec 14;287(1):146-55.

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1-2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2-4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	5
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	8-9
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	9
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	9
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	9

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Reporting quality of studies using machine learning models for medical diagnosis: a systematic review.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-034568.R1
Article Type:	Original research
Date Submitted by the Author:	02-Dec-2019
Complete List of Authors:	Yusuf, Mohamed; Manchester Metropolitan University, Health Professions Atal, Ignacio; Université Paris Descartes, Centre for Research and Interdisciplinarity (CRI); Université Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Li, Jacques; Universite Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Smith, Philip; Manchester Metropolitan University, Health Professions Ravaud, Philippe; Universite Paris Descartes, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM Fergie, Martin; The University of Manchester, Imaging and Data Sciences Callaghan, Michael; Manchester Metropolitan University, Health Professions Selfe, James; Manchester Metropolitan University, Health Professions
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Diagnostics, Epidemiology, Evidence based practice
Keywords:	Machine learning, Medical diagnosis, Clinical prediction, Reporting quality

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: Reporting quality of studies using machine learning models for medical diagnosis: a systematic review

Corresponding author:
Mohamed Yusuf
Department of Health Professions
Manchester Metropolitan University
Manchester, UK
M.Yusuf@mmu.ac.uk

Ignacio Atal
Research Fellow
Centre for Research and Interdisciplinarity (CRI),
Université Paris Descartes

Jacques Li MD
Medical Doctor/Research intern in evidence-based medicine
INSERM, Methods
Université Paris Descartes
Paris, France

Philip Smith
Senior lecturer
Department of Health Professions
Manchester Metropolitan University
Manchester, UK

Philippe Ravaud
Professor in Epidemiology
INSERM, Methods
Université Paris Descartes
Paris, France

Martin Fergie
Lecturer in Health sciences
Division of Informatics,
Imaging and Data Sciences,
University of Manchester
Manchester UK

Michael Callaghan
Professor
Department of Health Professions
Manchester Metropolitan University
Manchester, UK

James Selfe
Professor
Department of Health Professions
Manchester Metropolitan University
Manchester, U

Abstract

Aims: We conducted a systematic review assessing the reporting quality of studies validating models based on machine learning (ML) for clinical diagnosis, with a specific focus on the reporting of information concerning the participants on which the diagnostic task was evaluated on.

Method: Medline Core Clinical Journals were searched for studies published between July 2015 to July 2018. Two reviewers independently screened the retrieved articles, a third reviewer resolved any discrepancies. An extraction list was developed from the TRIPOD guideline. Two reviewers independently extracted the data from the eligible articles. Third and fourth reviewers checked, verified the extracted data as well as resolved any discrepancies between the reviewers.

Results: The search results yielded 161 papers, of which 28 conformed to the eligibility criteria. Detail of data source was reported in 24 of the 28 papers. For all of the papers, the set of patients on which the ML-based diagnostic system was evaluated was partitioned from a larger dataset, and the method for deriving such set was always reported. Information on the diagnostic/non-diagnostic classification was reported well (23/28). The least reported items were the use of reporting guideline (0/28), distribution of disease severity (8/28 patient flow diagram (10/28) and distribution of alternative diagnosis (10/28). A large proportion of studies (23/28) had a delay between the conduct of the reference standard and ML tests, while one study did not and four studies were unclear. For 15 studies, it was unclear whether the evaluation group corresponded to the setting in which the ML test will be applied to.

Conclusion: All studies in this review failed to use reporting guidelines, and a large proportion of them lacked adequate detail on participants, making it difficult to replicate, assess and interpret study findings.

Review registration number: CRD42018099167 **word count**: 3,060

Keywords: Machine learning, Medical diagnosis, Reporting quality

Strengths and limitations of this study

- This is the first systematic review evaluating the reporting quality of studies developing and/or validating machine learning (ML) methods for medical diagnosis within the medical literature.
- Using a systematic approach, this review included studies published within the Medline Core Clinical Journals, these journals cover all areas of clinical and public health
- The review used TRIPOD to help extract information concerning the participants within the reviewed studies.
- This review only focused on the reporting quality and therefore did not evaluate the statistical methodology and conduct of studies using ML diagnostic systems.
- Although a risk of bias assessment is not essential for research on research, the following review did not use risk of bias assessment tool.

Introduction

Machine learning (ML) is a rapidly developing area, characterised as the science of training computers to conduct specific tasks such as classification or prediction without explicit programming, but where decisions are taken based on patterns and relationships within large and complex datasets [1]. Over the past decade, access to large amounts of clinical data and the development of new ML techniques has led to a rise in the application of ML methods to medicine [2 3]. Due to their propensity to facilitate and promote timely and objective clinical decision-making, ML methods have been applied to gain valuable insights into clinical diagnoses. For example, ML methods have been used to diagnose skin cancer using skin lesion images [4], diagnose cerebral aneurysms using clinical notes [5] and diagnose stroke using neuroimaging data [6]- see *box 1* for an example of ML-based diagnostic system.

While there is no consensus on the definition, a key principle of ML models is that they are developed based on the automatic extraction of patterns from data [7]. In contrast to traditional statistics, whereby models are explicitly programmed based on statistical theory and assumptions, ML models learn from examples without the need for explicit rules to make decisions [8] . Generally, a researcher developing a ML model has access to a large dataset that is divided into a training set and a test set (see Box 1). The training set is used to develop a ML model that will learn the relationships between available clinical data and an outcome of interest (e.g. a diagnosis). The performance of the ML model is then evaluated by applying it to the test set. As ML models are

only as good as the data used to train them, it is vital to emphasise the importance of data quality [9].

Box 1

Machine learning is the ability to create algorithms to accomplish specific tasks without explicitly programming them, but rather take decisions based on previously seen data. Here is a summary of the steps when creating machine learning algorithms:

Model Development

Step 1: Defining the research problem. This could be broken down to either a **classification**, **regression** or a **clustering problem**.

Step 2: Identification of data sources and formats. Data could be in various formats (e.g. images, text, speech or numerical), and data could come from various sources such as hospital, insurance databases, or previous research projects.

Step 3: Training and test set derivation. Here the data could be broken down into two independent components: the **training set** and **test set**. The training set is used to create the ML algorithm, and the test set is used to evaluate the ML algorithm.

Step 4: Model development. The model is developed using the training dataset. The model could be either **supervised** or **unsupervised** (supervised models require labelled data whereas unsupervised models do not). The **loss function** and the methods for handling **outliers** and **missing data** are also described. A portion of the training set, the **model selection set**, is often withheld from model training to allow for model selection and to avoid overfitting.

Step 5: Evaluation of the model. The test set is used to evaluate the ML algorithm using a variety of metrics to compare the prediction with the gold standard outcome label (often referred to as **ground-truth**).

Model Validation

To obtain an accurate assessment of model's performance in a clinical setting, the model must be validated against data which is drawn from a clinical cohort. **Internal validation** refers to a model being evaluated on a cohort taken from the same setting as the data used to develop the model. **External validation** is where the cohort data is taken from a separate setting, which overcomes any systematic biases present in the data source used for model validation.

It is worth noting that one potential area for confusion is the differing meanings of the terms **test set** and **validation set** between the machine learning and medical research community. A medical researchers validation set is a machine learning test set.

Despite their popularity, the promising applications of ML-based diagnostic systems come with its own set of pitfalls. Studies using ML for medical diagnosis may contain systematic errors in both the design and execution [10-12]. For instance, selection bias can occur if the sample used to produce the ML-based diagnostic system is not entirely representative of the population on which the model may be used in the future [11]. Repeated evaluation of model performances against the same *test set* may result in the selected model overfitting the test set, resulting in an over-optimistic assessment of model performance [13]. These methodological biases can make it difficult to generalise conclusions from the results yielded. This could lead to erroneous yet devastating clinical decisions, i.e. recommending a medical treatment to an individual that is different from those in the population the that treatment was developed and validated on [14].

There is a parallel between what ML researchers refer to as 'test set' and the 'Population on which a diagnostic test is evaluated' within diagnostic accuracy studies. The diagnostic accuracy of an ML-based systems is reliant on demographic and clinical characteristics of the population in which it was applied on, therefore, if the cohorts are not a representative sample of the targeted population, then the generalisability of the study results may be limited. A further hindrance to the application of ML methods for medical diagnosis (and more generally in biomedical research) is that ML researchers may not be familiar with the requirements and guidelines that biomedical research have collectively established to ensure transparent and unbiased evidence-based knowledge accumulation [15 16].

Clinical predictions models undergo a scientifically rigorous process to establish their diagnostic accuracy, which encompasses their safety, validity, reproducibility, usability, and reliability. Highlighting the importance of transparent and rigorous reporting of clinical predictions models accuracy studies, particularly as the diagnostic prediction models of an instrument can vary greatly due to factors such as population characteristics, clinical setting, disease prevalence and severity as well as aspects of test execution and interpretation [16]. To aide with and standardise this process, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – (TRIPOD) guideline was set in place. The TRIPOD is an internationally accepted reporting guideline that was developed to improve the reliability and value of clinical prediction models through the promotion of transparent and accurate reporting [16].

In medical research, reporting guidelines are implemented to aid in the transparent evaluation, usability and reproducibility of a diagnostic instrument [17]. Luo *et al.* [18] have constructed a reporting guideline for the development and usage of ML predictive models in biomedical research. This is an important step towards a rigorous and robust approach to the usage of ML methods in medical research. Since publication in December 2016 (up to May 2019), the guideline of Luo *et al.*, which is currently available on The EQUATOR Network website [19], has garnered only ~50 citations. Additionally, in 2015 a more specific and robust guideline was developed to aid the reporting of prediction models used in prognostic and diagnostic studies (TRIPOD) [16]. TRIPOD has ~1,000 citations demonstrating that is has been accepted by the community as a useful set of guidelines for diagnostic/prognostic prediction. In this work we evaluate whether ML studies make use of these guidelines.

To date, there have been no studies evaluating the reporting quality of studies using ML methods, particularly diagnostic studies. Knowing this may aid in the evaluation of reporting standards employed by ML researchers. In this review, we focus on medical research studies that used ML methods to aid clinical diagnosis. Further, we have narrowed our review to applied ML methods which are envisaged to be clinically useful, in which the end users are practitioners and research consumers.

We aimed to produce a systematic review assessing the reporting quality of studies developing or validating ML models for clinical diagnosis, with a specific focus on the reporting of information concerning the participants on which the diagnostic task was evaluated.

Methods

This review was registered with International prospective register of systematic reviews (PROSPERO) on 30/07/2018 (reference: CRD42018099167). The framework used for this methodological systematic review is Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guideline for Systematic reviews [20].

Literature search

On July 2018, two authors (MY & JL) independently searched through the Medline Core Clinical Journals for articles developing or validating ML models for clinical diagnosis. Core Clinical Journals, also known as Abridged Index Medicus (AIM), is a filter option within Medline that limits to clinically useful journals. This is a selection of 119 English-language journals that focus on clinical studies and that are considered to be of immediate interest to practising physicians [21]. Using this filter excludes journals in bioinformatics or computational biology, which are highly likely to include articles explaining the development of ML-based diagnosis systems. However, these journals might not target clinicians. In addition to this, due to the ever-expanding ML literature, we have narrowed our review to studies published between July 2015 to 1 July 2018. See supplementary file 1 for the search strategy.

Subsequent to the literature search, the two reviewers (MY & JL) screened the title and abstracts of the search results. Once the eligible papers were identified and retrieved, both the first reviewer (MY) and second reviewer (JL) independently screened the full articles for eligibility. Discrepancies between the two reviewers were discussed and resolved by a third reviewer (IA).

Inclusion and exclusion

Studies were included if they used ML for clinical diagnosis, for example if they used statistical techniques to conduct classification, regression or clustering based on clinical data for disease diagnosis without being explicitly programmed. Other inclusions were primary study designs that evaluated the accuracy of such ML-based systems for diagnostic tasks, and articles in the English language. Studies were excluded if they did not report original research, if they were systematic reviews, had no abstract or did not specify the type of ML model adopted.

Extraction list

For studies developing, evaluating or updating clinical prediction models (this includes diagnosis), the TRIPOD provides guidance on reporting the key items. As it stands, TRIPOD is the most rigorous and relevant guideline for evaluating the use of ML methods for medical diagnosis. As such, an extraction list based on the TRIPOD checklist was developed. The focus of the extraction list was to extract information about the participants upon which the diagnostic task was evaluated on,

namely selection method and population characteristics. We additionally extracted general information concerning the diagnostic tasks, namely the target condition and the target population. The extraction list was tested and validated by two reviewers (MY & JL), by applying it on a random sample of the eligible papers.

Data extraction

Two reviewers (JL & PS) independently extracted the data from the eligible articles based on the items listed in *Table 1*. For each of the items, reviewers declared if the item was clearly reported (yes/no/unclear) and justify the declaration by citing the manuscript verbatim, as well as providing a written explanation if the reporting was considered unclear. A third and fourth reviewer (MY & IA) checked and verified the extracted data and resolved any disagreements between the reviewers through discussion.

Data analysis

Findings from the included studies demonstrating study characteristics, reporting quality and presence of bias were presented in descriptive statistics and figures.

Patient and public involvement

There was no patient or public involvement in any phase of this study, this included the development of the research question, the analysis and the conclusions.

Table 1: Item list used to extract eligible papers.

Item groups	Item list	Detailed items
General characteristics	Diagnostic task	What is the target condition?
	Study objective	Is the study aiming at the development of a diagnostic method, evaluation of a diagnostic method, or both?
	Target population	What is the population targeted by the diagnostic test?
Methods	Data sources	Where and when potentially eligible participants were identified (setting, location and dates)
	Data split	Method for partitioning the evaluation set from the training data. To assess whether participants formed a consecutive, random or convenience series.
	Test dataset eligibility criteria	On what basis potentially eligible participants were identified within the test dataset (such as symptoms, results from previous tests, inclusion in registry).
Results	Baseline characteristics	Baseline demographic and clinical characteristics of participants
	Diagnosis/non-diagnosis classification	Classification of the diagnosed and non-diagnosed patients within the test set.
	Flow diagram	Flow of participants, using a diagram.
	Severity	Distribution of severity of disease in those with the target condition.
	Alternative diagnosis	Distribution of alternative diagnoses in those without the target condition.
	Difference between reference test and ML test.	Is there a time interval between index test and reference standard?

	Applicability	Does the evaluation population correspond to the setting in which the diagnosis test will be applied?

Results

The search yielded 161 papers, of which 28 conformed to the eligibility criteria, see *figure 1*. During the screening of the title and abstract, most papers were excluded due to the search term 'CAD' being analogous to both 'Computer-aided Detection' and Coronary Artery Disease'. During the full text review, eleven papers were excluded because they did not use ML methods for medical diagnosis, and three papers were excluded because they did not use ML method but were captured in the search because they studied Coronary Artery Disease (CAD).

Figure 1: PRISMA Flow diagram

Study characteristics

The study characteristics of the all eligible studies are presented in Table 2 (see supplementary file 1 for list of studies). From the papers extracted, majority of the studies were published in 2017 (43%) and mostly in the Radiology journal (29%). Oncology was the most researched domain (47%), followed by Neurology (18%). The majority of studies focused on model development (97%), with only one study looking at model validation

Table 2: Study characteristics

Items	Total n (%)
Year	
2015	4 (14)
2016	9 (32)
2017	12 (43)
2018	3 (11)
Journals	
Radiology	8 (29)
Jama	2 (7)
Brain	2 (7)
American Journal of Roentgenology	2 (7)
Neurology	1 (3)
Medicine	1 (3)
Surgery	1 (3)
Chest	1 (3)
Gastroenterology	1 (3)
Journal of the American College of Cardiology	1 (3)
Journal of Allergy and Clinical Immunology	1 (3)
American journal of clinical pathology	1 (3)
American journal of ophthalmology	1 (3)
The Journal of infectious diseases	1 (3)
Digestive diseases and sciences	1 (3)
The British journal of radiology	1 (3)

The Journal of pediatrics	1 (3)
Clinical Specialty	
Oncology	13 (47)
Neurology	5 (18)
Immunology	2 (7)
Ophthalmology	2 (7)
Others specialties[a]	6 (21)
Task	
Development and Evaluation	27 (97)
Evaluation	1 (3)

[a] Other clinical specialities include Cardiology, Gastroenterology, Infectious disease, Psychiatry, Endocrinology and Various.

Reporting quality

Detail of the data source was reported in 86% of the papers, with all studies providing information on the separation method for deriving the evaluation set from the larger dataset. 82% of studies reported eligibility criteria for both evaluation set. Information on the diagnostic/non-diagnostic classification evaluation metric used was included in 82% of all papers. The least reported items were use of reporting guideline (0%), distribution of disease severity (29%), patient flow diagram (34%), distribution of alternative diagnosis (36%) and baseline characteristic (64%). See table 3 for a full breakdown of reporting quality.

Table 3: Reporting quality

Items	Reported n (%)	Not reported n (%)	Unclear n (%)
Methods			
Data source	24 (86)	0 (0)	4 (14)
Data split methods	28 (100)	0 (0)	0 (0)
Test set eligibility criteria (evaluation set)	23 (82)	5 (18)	0 (0)
Results			
Baseline characteristic	17 (61)	11 (39)	0 (0)
Diagnosis/non-diagnosis classification	23 (82)	4 (14)	1 (4)
Flow diagram	10 (36)	18 (64)	0 (0)
Disease severity	8 (29)	18 (64)	2 (7)
Alternative diagnosis	10 (36)	18 (64)	0 (0)
Use of reporting guideline	0 (0)	28 (100)	0 (0)

N = number

Presence of bias

Within the eligible studies, 71% did not report a time interval between the conduct of the reference standard and ML test (table 4). Within 54% of studies, it was unclear whether the study populations corresponded to the setting in which the diagnostic test will be applied to. However,

in 29% of studies, the clinical setting of the gathered evaluation dataset did not correspond to the clinical setting in which study authors hoped it would be applied.

Table 4: Presence of bias

Items	Yes n (%)	No n (%)	Unclear n (%)
Is there a time interval between reference standard ML test?	23 (82)	1 (4)	4 (14)
Does the test population correspond to the population/setting in which the diagnosis test will be applied?	5 (18)	8 (29)	15 (54)

Discussion

This review found that studies developing or validating ML-based systems for clinical diagnosis failed to use reporting guidelines and lacked adequate detail for assessment, interpretation and reproducibility. With nearly all studies providing detail on data sources, eligibility criteria and diagnosis classification, only a few studies reported study participant flow diagram, distribution of disease severity, distribution of alternative diagnosis. Our findings are in line with those of Faes et al. recent systematic reviews [22] in which they found poor reporting and potential biases arising from study design in studies using ML methods for classifying diseases from medical imaging. Similarly, in another systematic review, Christodoulou et al' [23] found studies comparing the performance of logistic regression models with ML models for clinical prediction to have poor methodology and reporting quality.

A high number of studies reviewed had a time difference between the conduct of the reference test and that of ML-based diagnostic systems, suggesting the potential for incorporation bias [24 25]. This is largely an issue in ML-based diagnostic systems where labelling is the gold-standard, but patient data is labelled retrospectively. This may happen several years after initial data collection and in a different setting.

In more than half of the studies, it was unclear whether the study population corresponded to the setting in which the ML diagnostic system will be used in. However, in a third of the reviewed studies, the test populations did not correspond to the populations in which tests were hoped to be applied to, further limiting their generalisability. In addition to this, studies utilising ML-based diagnostics systems fail to report baseline characteristics. This could be problematic; within diagnostic accuracy studies it is imperative to report sample characteristics as this aid researchers, research consumers and practitioners in determining the relevancy and applicability of study findings to a wider setting.

Information on data source was unclear in four studies; this is vital in evaluating the source and methods used to derive study samples. In diagnostic studies the use of different methods to derive the evaluation sample from the wider population could lead to more or less accurate estimation of the diagnostic performance. The ideal method for sampling should be based on probability and not convenience, as this allows for a representative sample to be selected from a sampling frame whereby all eligible individuals have an equal chance of being selected. In addition to this, ML-based diagnostic systems that are evaluated using internal validation, where the evaluation set is partitioned from the same cohort as the training set, risk learning the systematic biases in the data of the particular centre from which the cohort was drawn. Such methods only address the systems internal validity, and model performance may deteriorate when deployed on an cohort drawn from a different centre [8]. External validation, where the ML-based diagnostic systems are evaluated on a cohort that has played no role in model development, is an important step to verify and asses the whether the system is reliable and deployable on potential populations for clinical use [26-28]. This is further highlighted in a recent systematic review evaluating the performance of ML algorithms for the diagnostic analysis of medical images; within this review, Kim et al. [29] found only 6% (31 out of 516 studies) had externally validated their algorithms.

Low-quality clinical research that is reported inadequately or that offers invalid data and distorted outcomes are deemed wasteful; such research is non-replicable and unusable [30 31]. One way to increase the value and reusability of these novel and promising ML-based systems is through complete, accurate and transparent reporting [19 31].. Some of the methodological and reporting issues facing the studies reviewed in this systematic review can be mitigated through the use of reporting guidelines such as TRIPOD guideline. More specifically, there has been a recent initiative to develop an extension of the TRIPOD statement which is specific to ML studies (TRIPOD-ML)[32]. Such guidelines aid researchers developing ML-based diagnostic systems in addressing the important aspects of design, execution and complete and reliable reporting of studies. However, no reporting guideline can salvage a poorly designed and executed study. To prevent flawed design and execution of ML methods, informatic and biomedical researchers looking to develop ML-based diagnostic systems should consult with a methodologist, epidemiologist or a statistician. Having input from such experts will aid in research that is methodological robust in design and execution - resulting in research that it is reliable, reproducible and that adds scientific value [33].

Strengths and limitations of study

To our knowledge, this is the first systematic review evaluating the reporting quality of studies developing and/or validating ML methods for medical diagnosis within the medical literature. A possible limitation within this review, is that we have not included all medical journals and therefore our findings may not be applicable to all journals. Despite this, we have included studies published within the Medline Core Clinical Journals, these journals cover all areas of clinical and public health.

This review did not evaluate the statistical methodology and conduct of studies using ML diagnostic systems. This could be considered a limitation as a transparent reporting does not

guarantee a quality study. Nevertheless, this review shows that these studies do not comply with TRIPOD guideline on the reporting this considerably affects the trust we have in the estimates they are giving concerning the efficacy of their diagnostic methods. Another potential limitation is that the following review did not utilise risk of assessment tool, however, this is not an essential component for this type of review, as the main objective is to determine the reporting quality of studies and not synthesis research evidence .

Conclusion

We found that all eligible studies in this review failed to use reporting guidelines and the studies lacked adequate detail on the participants on which the diagnostic task was evaluated on, thus making it difficult to replicate, assess and interpret study findings.

Diagnostic studies using ML methods have great potential to improve clinical decision-making and take the load off health systems. However, studies with poor reporting can be more problematic than of help. Within biomedical research, there is an already established framework and guidelines in which ML researchers can utilise to aid the execution and reporting of ML methods for clinical diagnosis, with the TRIPOD guideline being the most robust and widely used.

Abbreviation: ML: Machine learning, CAD: Computer Aided Diagnosis, TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Legends: * Other clinical specialities include Cardiology, Gastroenterology, Infectious disease, Psychiatry, Endocrinology and Various.
Figure 1 is the PRISMA 2009 Flow diagram.

Contributors: JS is a guarantor of this review. All authors have made substantive intellectual contributions to the development of this review. MY and IA were involved in conceptualising the review. MY, IA and PR developed the protocol. MY, JL & IA did the literature search, MY, JL, PS & IA carried out the study selection and data extraction. MY, IA, JS, PS, MF and MC were involved in the writing and editing of the manuscript.

Transparency: The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and registered) have been explained.

Funding: There is no funding attached to this systematic review. This review was conducted independently by the research team.

Ethical approval: Not required as this is a review

Data sharing: All data are freely available within the appendices. No additional available.

Competing interest: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

References

1. Paliwal M, Kumar UA. Neural networks and statistical techniques: A review of applications. *Expert systems with applications* 2009;36(1):2-17.
2. Cleophas TJ, Zwinderman AH. Machine Learning in Medicine - a Complete Overview: Springer International Publishing 2015.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 2019;25(1):44.
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115.
5. Shin H, Kim KH, Song C, et al. Electrodiagnosis support system for localizing neural injury in an upper limb. *Journal of the American Medical Informatics Association* 2010;17(3):345-47.
6. Rehme AK, Volz LJ, Feis D-L, et al. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. *Cerebral cortex* 2014;25(9):3046-56.
7. Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 2005;27(2):83-85.
8. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating: Springer Science & Business Media 2008.
9. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine* 2016;375(13):1216-19. doi: 10.1056/NEJMp1606181
10. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286(3):800-09.
11. Bone D, Goodwin MS, Black MP, et al. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders* 2015;45(5):1121-36.
12. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical engineering online* 2014;13(1):94.
13. Subramanian J, Simon R. Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary clinical trials* 2013;36(2):636-41.
14. Greenhouse JB, Kaizar EE, Kelleher K, et al. Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users. *Statistics in medicine* 2008;27(11):1801-13.
15. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6(11):e012799. doi: 10.1136/bmjopen-2016-012799

16. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015;162(1):W1-W73.
17. Simera I, Moher D, Hirst A, et al. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC medicine* 2010;8(1):24.
18. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research* 2016;18(12)
19. Simera I, Moher D, Hoey J, et al. The EQUATOR Network and reporting guidelines: Helping to achieve high standards in reporting health research studies. 2009;63(1):4-6.
20. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339 doi: 10.1136/bmj.b2700
21. Committed Selection: Abridged Index Medicus. *New England Journal of Medicine* 1970;282(4):220-21. doi: 10.1056/nejm197001222820410
22. Faes L, Liu X, Kale A, et al. Deep Learning Under Scrutiny: Performance Against Health Care Professionals in Detecting Diseases from Medical Imaging-Systematic Review and Meta-Analysis. 2019
23. Jie M, Collins GS, Steyerberg EW, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* 2019
24. Kramer MS, Roberts-Bräuer R, Williams RL. Bias andovercall'in interpreting chest radiographs in young febrile children. *Pediatrics* 1992;90(1):11-13.
25. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy. A systematic review. *Annals of internal medicine* 2004;140(3):189-202.
26. Steyerberg E. Overfitting and optimism in prediction models. Clinical Prediction Models: Springer 2009:83-100.
27. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj* 2016;353:i3140.
28. Steyerberg EW, Harrell Jr FE, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology* 2001;54(8):774-81.
29. Kim DW, Jang HY, Kim KW, et al. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean journal of radiology* 2019;20(3):405-10.
30. Chan A-W, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. 2014;383(9913):257-66.
31. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet* 2014;383(9913):267-76.
32. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. *The Lancet* 2019;393(10181):1577-79.
33. Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet* 2014;383(9912):166-75.

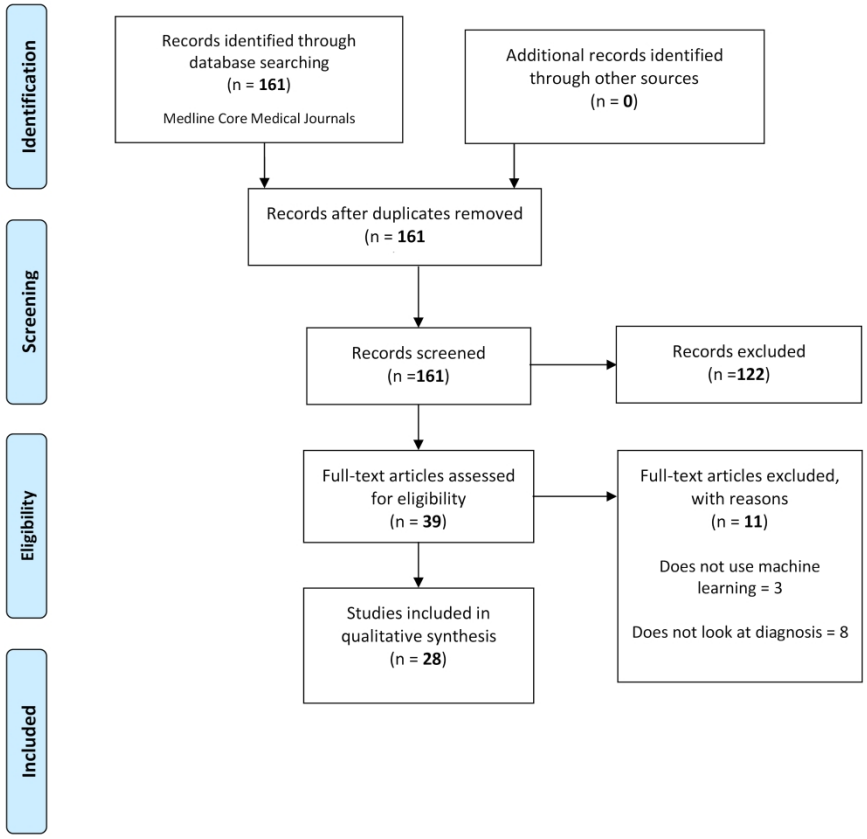

Figure 1: PRISMA 2009 Flow diagram

487x545mm (300 x 300 DPI)

Search strategy using Medline (restricted to Core Clinical Journals)

1) machine learning,
2) supervised learning
3) unsupervised learning
4) deep learning
5) artificial Intelligence
6) decision trees
7) Artificial
8) Neural Network
9) CNN
10) ANN
11) Convolutional Neural Network
12) random forest,
13) reinforcement learning
14) gradient boosting
15) computer aided diagnosis
16) CAD
17) computer assisted diagnosis
18) computational analysis

19) OR/ 1-18

20) Diagnosis

21) 19 AND 20

List of eligible studies

Author	Journal	Specialty	Title
Asaoka et al., 2017	American journal of ophthalmology	Ophthalmology	Validating the usefulness of the "random forests" classifier to diagnose early glaucoma with optical coherence tomography.
Bahl et al., 2017	Radiology	Oncology	High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision.
Becker et al., 2018	The British journal of radiology	Oncology	Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study.
Bejnordi et al., 2017	JAMA	Oncology	Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer.
Chen et al., 2018	Gastroenterology	Gastroenterology	Accurate classification of diminutive colorectal polyps using computer-aided analysis.
Demertzi et al., 2015	Brain	Neurology	Intrinsic functional connectivity differentiates minimally conscious from unresponsive patients.
Dinh et al., 2018	Radiology	Oncology	Characterization of prostate cancer with Gleason score of at least 7 by using quantitative multiparametric MR imaging: validation of a computer-aided diagnosis system in patients referred for prostate biopsy.
Eshagi et al., 2016	Neurology	Neurology	Gray matter MRI differentiates neuromyelitis optica from multiple sclerosis using random forest.
Gallego-Ortiz et al., 2015	Radiology	Oncology	Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions.
Hao et al., 2016	The Journal of pediatrics	Immunology	A classification tool for differentiation of Kawasaki disease from other febrile illnesses.
Harper et al., 2016	Brain	Neurology	MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases.
Hornbrook et al., 2017	Digestive diseases and sciences	Oncology	Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data.
Huang et al., 2017	Radiology	Oncology	Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study.
Keller et al., 2016	Journal of Allergy and Clinical Immunology	Immunology	Mutation in IRF2BP2 is responsible for a familial form of common variable immunodeficiency disorder.
Lee et al., 2015	American Journal of Roentgenology	Oncology	Does computer-aided diagnosis permit differentiation of angiomyolipoma without visible fat from renal cell carcinoma on MDCT?
Li et al., 2017	American Journal of Roentgenology	Oncology	Computer-aided diagnosis of ground-glass opacity nodules using open-source software for quantifying tumor heterogeneity.

Lu et al., 2016	Medicine	Psychiatry	Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images.
Moller et al., 2015	Radiology	Neurology	Alzheimer disease and behavioral variant frontotemporal dementia: automatic classification based on cortical atrophy for single-subject diagnosis.
Narula et al., 2016	Cardiology	Cardiology	Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography.
Ng et al., 2015	American journal of clinical pathology	Oncology	Computer-aided detection of rare tumor populations in flow cytometry: An example with classic Hodgkin lymphoma.
Silterra et al., 2016	The Journal of infectious diseases	Infectious diseases	Transcriptional categorization of the etiology of pneumonia syndrome in pediatric patients in malaria-endemic areas.
Somnay et al., 2017	Surgery	Endocrinology	Improving diagnostic recognition of primary hyperparathyroidism with machine learning
Sun et al., 2017	Radiology	Neurology	Psychoradiologic utility of MR imaging for diagnosis of attention deficit hyperactivity disorder: a radiomics analysis
Ta et al., 2017	Radiology	Hepatology	Focal Liver Lesions: Computer-aided Diagnosis by Using Contrast-enhanced US Cine Recording.
Ting et al., 2017	JAMA	Ophthalmology	Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes.
Tomlinson et al., 2016	Chest	Oncology	Transcriptional profiling of endobronchial ultrasound-guided lymph node samples aids diagnosis of mediastinal lymphadenopathy.
Yasaka et al., 2017a	Radiology	Oncology	Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study.
Yasaka et al., 2017b	Radiology	Oncology	Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid–enhanced hepatobiliary phase MR images.

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1-2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2-4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	5
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	7-8
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	8-9
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	9
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	9
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	9

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.