

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Predicting the risk of cancer in adults using supervised machine learning: a scoping review

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-047755
Article Type:	Original research
Date Submitted by the Author:	20-Dec-2020
Complete List of Authors:	Alfayez, Asma; University College London; King Abdullah International Medical Research Center Kunz, Holger; University College London
Keywords:	Information technology < BIOTECHNOLOGY & BIOINFORMATICS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, ONCOLOGY

SCHOLARONE™
Manuscripts



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.

1
2
3 **1 Predicting the risk of cancer in adults using supervised machine learning: a scoping review**
4

5
6 **2 Asma Abdullah Alfayez^{1,2,3}, Holger Kunz¹**
7

8 **3 ¹Institute of Health Informatics, University College London, London, UK**

9 **4 ²King Abdullah International Medical Research Center, Riyadh, Saudi Arabia**

10 **5 ³King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia**

11 **6 *Corresponding author: Asma Abdullah Alfayez (asma.alfayez.17@ucl.ac.uk)**
12
13

14
15 **7 Word count: 4375**
16

17 **8 Keywords: Machine learning; cancer; prediction; scoping review; artificial intelligence**
18

19 **9 Competing interests: None**
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

10 ABSTRACT

11 **Objectives:** The purpose of this scoping review is to identify existing supervised machine learning
12 (ML) approaches on the prediction of cancer in asymptomatic adults, to compare the performance of
13 ML models with each other, and to identify potential gaps in research.

14 **Design:** Scoping review using the population, concept, and context approach.

15 **Search strategy:** The literature was searched according to the following inclusion criteria: (i) a
16 general adult (≥ 18 years) population, either sex, asymptomatic (*population*); (ii) any study using ML
17 techniques to derive predictive models for future cancer risk using clinical and/or demographic and/or
18 basic laboratory data (*concept*); and (iii) original research articles conducted in all settings in any
19 region of the world (*context*).

20 **Results:** The search returned 627 unique articles, of which 580 articles were excluded because they
21 did not meet the inclusion criteria, were duplicates, or were related to benign neoplasm. Full-text
22 reviews were conducted for 47 articles and a final set of 10 articles were included in this scoping
23 review. These 10 very heterogeneous studies used ML to predict future cancer risk in asymptomatic
24 individuals. Nine out of 10 ML models reported either excellent or good performance.

25 **Conclusions:** Research gaps that must be addressed in order to deliver validated ML-based models
26 to assist clinical decision-making include: (i) establishing model generalisability through validation in
27 independent cohorts, including those from low- and middle-income countries; (ii) establishing models
28 for all cancer types; (iii) thorough comparisons of ML models with best available clinical tools to
29 ensure transparency of their potential clinical utility; and (iv) comparisons of different methods on the
30 same cohort to reveal important information about model generalisability and performance.

32 ARTICLE SUMMARY

33 Strengths and limitations of this study

- 34 • This study used the population, concept, and context scoping review approach to explore the
35 machine learning techniques used to derive predictive models for future cancer risk using
36 basic clinical and/or demographic and/or laboratory data (*concept*) in asymptomatic adults
37 ≥ 18 years (*population*) in all settings in any region of the world (*context*).

- 1
2
3 38 • Although the ML methodologies were heterogeneous, the standard use of the area under the
4 receiver operating characteristics curve (AUC) metric to evaluate model performance allowed
5 39 comparisons of different ML techniques with each other.
6
7 40
8
9 41 • This scoping review is limited to papers published in English between 2011 and 2020.
10 42
11
12 43

14 44 INTRODUCTION

16
17 45 Cancer remains a leading cause of morbidity and mortality, with an estimated 1.8 million new cases
18 and 0.6 million deaths in the US in 2019 and approximately 367,000 new cases and 165,000 cancer
19 46 deaths in the UK each year between 2015 and 2017.^{1,2} Annual death rates only modestly decreased
20 47 (1.4% and 1.8% in women and men, respectively) between 2012 and 2016, despite significant
21 48 research.¹ Cancer cases also continue to increase, not least due to increased life expectancy, which
22 49 increases the risk of developing cancer.³
23 50

24 51 Early cancer diagnosis is associated with significantly higher survival rate and lower mortality and
25 52 associated costs. Early-stage cancers require less complex treatment regimens and reduced hospital
26 53 utilization, resulting in reduced healthcare costs, whereas late-stage cancers require complex
27 54 multimodal management, several rounds of extremely expensive drugs over significant periods of
28 55 time, and the treatment of recurrences, equating to a staggering economic burden. Therefore, the
29 56 importance of early diagnosis cannot be overestimated.⁴⁻⁶
30
31
32
33
34
35
36
37
38
39

40 57 Survival rates significantly improve if cancer is diagnosed at stage I or II compared with later stages
41 58 (stage III and IV),^{7,8} as once the cancer has metastasised, it becomes difficult to treat with
42 59 radiotherapy or surgery, leading to treatment failure and death. For example, five-year survival rates
43 60 for women diagnosed with localised breast or ovarian cancer are 99% and 92% compared to 27%
44 61 and 29% for metastatic disease, respectively.¹ A report by Cancer Research UK indicated that, in the
45 62 UK, the ten-year survival proportions of patients with eight cancers (combined) were around 80% for
46 63 stage I and stage II detection (breast, bladder, ovarian, colorectal, uterine, testicular, and cervical
47 64 cancer and malignant melanoma) but only 26% for cancers detected at later stages, notably lung
48 65 cancer (stage III and IV).⁹
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 66 Globally, treatment for early-stage cancer confer significant cost-saving benefits. In the US, during the
4
5 67 first 24 months after diagnosis, there is an increase in cancer treatment costs with stage: US\$72,000
6
7 68 for stage 0, US\$97,000 for stage I/II, US\$159,000 for stage III, and \$182,000 for stage IV.¹⁰ An
8
9 69 estimate of the cost savings from early cancer diagnosis is 26 billion US dollars per annum in the US
10
11 70 alone.¹¹ Similarly, in the UK, early diagnosis of colorectal, ovarian, and lung cancer in England alone
12
13 71 could provide savings of over £44 million and benefit nearly 11,000 patients.¹²

15 72 **Current approaches to diagnose incident cancer**

17 73 One approach to the early detection of cancer is population-wide screening, which aims to find
18
19 74 asymptomatic individuals so that they can be promptly referred for treatment. Examples include
20
21 75 mammography for breast cancer, cervical screening for cervical cancer, and faecal occult blood
22
23 76 testing or sigmoidoscopy for colorectal cancer.¹³ There are three examples of national screening
24
25 77 programs in UK (bowel, breast, and cervical cancer screening programs¹⁴) and two in the US: the
26
27 78 Colorectal Cancer Control Program (CRCCP) and the National Breast and Cervical Cancer Early
28
29 79 Detection Program (NBCCEDP).¹⁵ However, significant proportions of individuals eligible for these
30
31 80 programs do not participate (for example through fear or not prioritizing time to attend for screening),¹⁶
32
33 81 and comprehensive screening programs are costly to implement, especially in resource-poor settings
34
35 82 or low- and middle-income countries. Other approaches include public health campaigns to
36
37 83 encourage individuals experiencing particular symptoms such as weight loss, anorexia, and fatigue to
38
39 84 visit their family doctors.¹⁷ However, patient help-seeking around cancer is complex, multi-staged, and
40
41 85 often leads to long delays of weeks or even months.¹⁸ Patients find it hard to interpret and recognise
42
43 86 symptoms, with fears of embarrassment and having a potentially fatal or painful condition contributing
44
45 87 to long and avoidable delays in help-seeking from health professionals.^{18 19} Patients often do not seek
46
47 88 help from health professionals for early cancer symptoms, notably from general family physicians, for
48
49 89 many reasons including a complex mix of fear, worry, and of 'wasting' health professionals' time¹⁹ or
50
51 90 due to the high costs of medical care, a lack of health insurance, or time constraints.²⁰

52 91 **Detecting future risk of cancer by modelling data**

55 92 Screening approaches represent a patient identification (or "phenotyping" problem) that aims to detect
56
57 93 whether the individual has cancer at a particular point in time. However, the ultimate goal of cancer
58
59 94 prediction is to determine whether an individual will develop cancer at some point in the future. A
60

1
2
3 95 simple approach is to stratify populations according to the presence and absence of risk factors,
4
5 96 which have been extensively characterised for most cancer types through epidemiological studies
6
7 97 over many decades. For example, age, gender, ethnicity, family history, and lifestyle factors are well-
8
9 98 established risk factors for many types of cancer.²¹ The cancer prediction problem can either be
10
11 99 regarded as a supervised learning problem where the input variables are clinical-demographic
12
13 100 variables and the output variable is the probability of developing cancer at some point in the future or
14
15 101 as a binary classification problem to determine whether or not a patient will develop cancer at a
16
17 102 specific point in time.

19 103 **Big data and machine learning for medical prediction models**

21 104 Advances in digital medicine and computational science have altered the landscape of data available
22
23 105 for cancer risk prediction models. For example, in the data-driven healthcare era, there is an
24
25 106 increasing amount of “big” medical data, as most individuals have had interactions with the healthcare
26
27 107 system where data is collected in the form of electronic health records (EHRs), which are systematic
28
29 108 collections of longitudinal patient health data collected in real time.^{22 23} Such large datasets provide
30
31 109 powerful new opportunities to develop and refine predictive models and to explore potentially
32
33 110 unknown predictor variables.²² Leveraging often massive amounts of data generated from large
34
35 111 populations, much of which may be unstructured, and building optimal models requires the
36
37 112 exploitation of advanced computational tools and supporting infrastructure. Machine learning (ML) is a
38
39 113 branch of artificial intelligence (AI) and an extension of traditional statistical techniques that uses
40
41 114 computational resources to detect underlying patterns in high-dimensional data, and it is increasingly
42
43 115 being used in different areas of medicine requiring predictions.²⁴ For example, ML has successfully
44
45 116 been used with EHR data to predict incident hypertension²⁵ and incident chronic kidney disease,²⁶
46
47 117 and wider popular uses of ML in medicine include the automatic interpretation of medical images such
48
49 118 as in radiology²⁷ and histopathology²⁸ images.

51 119 **A brief description of machine learning**

53 120 A comprehensive description of ML models is beyond the scope of this scoping review. However,
54
55 121 relevant ML techniques relate to the problem of *learning* from data samples (e.g., EHR data) rather
56
57 122 than being pre-programmed with existing knowledge or rules. ML models can either be supervised
58
59 123 (i.e., where the data are labelled and the algorithm uses these data to learn to predict the output) or
60

1
2
3 124 unsupervised (i.e., where the data are unlabelled and the algorithm learns a structure inherent in the
4
5 125 data).²⁹ The cancer prediction problem is therefore a supervised problem; examples are provided as
6
7 126 inputs (or features) such as cancer risk factors like age, history, ethnicity, or blood count parameters
8
9 127 and outputs (or labels) such as whether or not the individual subsequently develops cancer. A variety
10
11 128 of available algorithms learn the best way to map the features to the labels by learning from the
12
13 129 observations.^{30 31} The resulting model, ideally, will then be able to generalise the information so that it
14
15 130 can be applied with high precision to new and unseen data.^{30 31}

16
17 131 Some of the main supervised ML models used in medical applications include decision trees (DTs;
18
19 132 and their adaptation, random forests (RFs)), support vector machines (SVMs), and artificial neural
20
21 133 networks (ANNs).^{30 31} DTs produce an output similar to a flow chart formed from feature nodes (risk
22
23 134 variables) that best discriminate between different labels (future cancer occurrence) to split the tree.³⁰

24
25 135 ³¹ In this way, new cases can be assessed by traversing the tree based on the feature values to
26
27 136 determine the output for that example.^{30 31} Decision trees are easy to interpret, since users are usually
28
29 137 able to visualise the steps leading to a particular classification, which may be useful in a clinical
30
31 138 setting where experts might wish to see how a particular decision was made.^{30 31} In RFs, several trees
32
33 139 are built using subsets of data and features, with predictions decided based on majority voting after
34
35 140 the example is assessed with respect to all the constructed trees.^{30 31}

36
37 141 In SVMs, each feature (risk factor) is mapped into a higher-dimensional space and the hyperplane
38
39 142 that optimally separates the output (future cancer occurrence) modelled.^{30 31} SVMs tend to generalise
40
41 143 well to unseen data and work well with complex (multidimensional) data but can be hard to interpret.³⁰

42 144 ³¹
43
44
45 145 ANNs are inspired by the neural connections in the human brain and are developed by creating
46
47 146 nodes (neurons) that weight certain features and produce an output value.^{30 31} By layering nodes in
48
49 147 between the input layer (features; cancer risk factors) and output layer (label; future cancer
50
51 148 occurrence) and modifying the weights during learning through a process called back-propagation,
52
53 149 the resulting model forms a prediction for unseen data when one of the nodes in the output layer is
54
55 150 positive.^{30 31} The terms “deep neural network” and “deep learning” are applied to ANNs with large
56
57 151 numbers of layers.^{30 31} While proving extremely powerful across a range of applications, ANNs can be
58
59 152 computationally very expensive and the way in which they classify (i.e., the intermediate “hidden”
60

1
2
3 153 layers) is opaque, making it difficult to determine exactly how they performed the classification
4
5 154 problem.^{30 31}
6
7

8 155 **Rationale for performing a scoping review**

9
10 156 This was a scoping review of studies using supervised ML techniques to predict the future risk of
11
12 157 developing cancer or specific cancers within a general asymptomatic adult (≥ 18 years) population
13
14 158 using clinical and/or demographic and/or basic laboratory data (e.g., complete blood counts) that are
15
16 159 likely to be readily available within the primary care setting. This approach therefore allowed to: (i)
17
18 160 identify the types of evidence available; (ii) clarify key concepts and definitions; (iii) examine how
19
20 161 research is currently being conducted; and (iv) to identify knowledge gaps.³²
21

22 162 **OBJECTIVES**

23
24 163 The objective of this study was to perform a scoping review and to synthesize knowledge of the
25
26 164 nature and effects of current ML techniques for early cancer detection in asymptomatic adults. The
27
28 165 scoping review was guided by the following research questions:

- 29
30 166 (i) Which, if any, ML methods are being developed for cancer risk prediction in asymptomatic
31
32 167 individuals in the community?
33
34
35 168 (ii) How do these models perform compare to each other?
36
37 169 (iii) Which research or knowledge gaps need to be addressed in order to advance the field?
38

39
40 170
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

171 **METHODS**

172 **Inclusion and exclusion criteria**

173 Therefore, using the population, concept, and context approach,³³ the inclusion criteria were: (i)
174 general adult (≥ 18 years) population, either sex, asymptomatic (*population*); (ii) any study using ML
175 techniques to derive predictive models for future cancer risk using clinical and/or demographic and/or
176 basic laboratory data carried out prior to August 7, 2020 (*concept*); and (iii) original research articles
177 conducted in all settings in any region of the world (*context*).

178 For the purposes of this study, and recognizing that 'machine learning' algorithms fall along a
179 continuum with statistical techniques,³⁴ all modelling approaches were included were defined as
180 machine learning in the respective papers (such as logistic regression).

181 Exclusion criteria were any ML model used to predict future events in patients with pre-existing or
182 symptoms of cancer; ML models developed using specialised tests such as genetic profiling or
183 imaging tests not generally available in the community; unsupervised ML models; and studies not
184 written in English.

185 **Literature search**

186 To identify relevant studies, PubMed was searched using the search string:

187 ("Cancer" Or "Cancers" OR "Oncology") AND ("Machine Learning" OR "ML" OR "Data Mining" OR
188 "Decision Support System" OR "Clinical Support System" OR "Classification" OR "Regression" OR
189 "Support vector machines" OR "Gaussian process" OR "Neural networks" OR "Logical learning" OR
190 "Bayesian network" OR "linear model") AND ("prognosis" OR "prognostic estimate" OR "predictor" OR
191 "prediction" OR "model" OR "diagnosis" OR "diagnostic"). This search was supplemented with manual
192 searching of the references and citations of previously published studies. All abstracts identified by
193 the initial search were screened for inclusion and checked for accuracy. For the included studies, data
194 were extracted from full papers. In instances where more information was required to determine
195 inclusion, the full text of the article was retrieved and assessed against the eligibility criteria.

196 **Assessment metric**

197 The strength of the predictive ability of the included models was assessed using AUC (area under the
198 receiver operating characteristics curve) data, a valid measure for evaluating classification

1
2
3 199 algorithms,³⁵ where an AUC of 0.90-1 = excellent, 0.80-0.89 = good, 0.70-0.79 = fair, 0.60-0.69 =
4
5 200 poor, and 0.50-0.59 = fail to describe model performance.³⁶
6
7
8 201
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

202 RESULTS

203 Main Findings

204 *Identified risk models*

205 Using the search strategy, 627 initial studies were identified where 10 studies met the inclusion
206 criteria (Table 1; Figure 1).^{31 37-46} The most common reasons for exclusion of studies were: (i) models
207 were derived to predict prognosis or responses to therapy in patients with pre-existing cancer; and/or
208 (ii) the studies used features other than clinical and/or demographic and/or basic laboratory data,
209 such as genetic biomarkers. All studies were retrospective cohort or case-control studies conducted
210 between 2011 and 2020, with 8 out of 10 studies completed in the last two years. Eight studies were
211 conducted in the USA and two in Taiwan. One model was built for breast cancer, three for colorectal
212 cancer, one for lung cancer, one for melanoma, two for non-melanoma skin cancer, one for
213 pancreatic cancer, and one a general cancer prediction model. Two studies performed external
214 validations of a previously developed colorectal cancer prediction model (Table 1).^{38 42}

215 *Development of the risk models*

216 The models developed in the studies employed a wide range of ML techniques. Two studies
217 compared different modelling approaches on the same dataset,^{40 43} while the other eight developed a
218 model using a single approach. The following ML approaches were used: ANNs (8 out of 10 studies),
219 logistic regression (2/10 studies), Gaussian naïve Bayes (1 out of 10 studies), Bayesian network
220 inference (1/10 studies), DTs (1/10 studies) and RFs (2/10 studies), linear discriminant analysis (1/10
221 studies), and SVMs (1/10 studies) (Table 1). Data for training and testing were from medical
222 insurance databases (4/10 studies), EHR data repositories (3/10 studies), surveys (2/10 studies), or
223 represented a retrospective analysis of prospectively collected data from a clinical trial (1/10 studies).

224 As a result of the diverse cancer types being modelled, study aims, and the available data, a range of
225 different predictors, features, and/or risk factors were included the developed predictive models,
226 which can be grouped into the following categories: (1) patient demographic data: e.g., age, gender,
227 ethnicity, family history; (2) social and lifestyle data: e.g., cigarette smoking and intensity of exercise;
228 (3) comorbidities: e.g., diabetes mellitus, hypertension, congestive heart failure, and chronic
229 obstructive pulmonary disease; (4) clinical and practice data: e.g., Anatomical Therapeutic Chemical
230 (WHO-ATC) prescription codes and clinical encounters; and (5) laboratory tests: e.g., complete blood

1
2
3 231 count (Table 1). The models that automatically extracted features from EHR records used features
4
5 232 that were not always explicitly defined in the respective articles.
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

233 **Table 1.** Summary of studies investigating ML approaches for early cancer detection.

Type of cancer	Reference	Year	Country	Method	Sample	Input	Validation	Performance	Model performance	Notes
Breast	Stark ⁴³	2019	USA	Logistic regression, Gaussian naive Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network	1343 breast cancer and 63,396 non-breast cancer cases (PLCO dataset)	Age, age at menarche, age at first live birth, number of first-degree relatives who have had breast cancer, ethnicity, age at menopause, an indicator of current hormone usage, number of years of hormone usage, BMI, pack years of cigarettes smoked, years of birth control usage, number of live births, and an indicator of personal prior history of cancer	20% testing data (269 breast cancer and 12,679 non-breast cancer cases)	AUC 0.61 (0.58-0.65); NB 0.59 (0.56-0.62); DT 0.61 (0.50-0.52); LDA 0.61 (0.58-0.65); SVM 0.52 (0.48-0.55); NN 0.61 (0.57-0.64)	Fail - poor	At an 0.05 level, the logistic regression, linear discriminant analysis, and neural network with the broader set of inputs were all significantly stronger than the BCRAT
Colorectal cancer	Hornbrook ³⁸	2017	USA	ColonFlag ML model	17,095 US community-based insured adults (16,195 controls, 900 cases) (insurance data)	Age, gender, and blood count panel parameters	Study was a validation of a previously derived model ⁴⁷	AUC 0.80 (0.79-0.82)	Good	
Colorectal	Wang ⁴⁵	2019	Taiwan	Convolutional neural network	10,185 with CRC, 47,967 controls	ICD-9-CM diagnostic codes, World Health Organization-	5-fold cross-validation	AUC 0.92	Excellent	

					(insurance data)	Anatomical Therapeutic Chemical (WHO-ATC) prescription codes				
Colorectal cancer	Schneider ⁴²	2020	USA	ColonFlag ML model	308,721 insurance health plan members (insurance data)	Age, gender, and blood count panel parameters	Study was a validation of a previously derived model ⁴⁷	AUC 0.78 (95% CI 0.77-0.78)	Good	The algorithm's accuracy decreased with the time interval between blood test result and CRC diagnosis
General	Miotto ³⁹	2016	USA	Deep neural network and random forests	Model training on 704,587, testing on 76,214 (EHR data)	Features extracted from EHR records	Testing on 76,214	Colorectal cancer AUC 0.89, liver cancer 0.89, prostate cancer 0.86	Good	Outperformed RawFeat and PCA
Lung	Hart ³⁷	2018	USA	Artificial neural network	1997-2015 National Health Interview Survey adult data; 648 cancer and 488,418 non-cancer cases (survey data)	Gender, age, BMI, diabetes, smoking status, emphysema, asthma, ethnicity, Hispanic ethnicity, hypertension, heart diseases, vigorous exercise habits, and history of stroke	30% of data; 195 lung cancer cases and 146,524 never cancer cases	AUC 0.86 (training; 95% CI 0.85-0.88) and 0.86 (validation; 95% CI 0.84-0.89)	Good	Random forests and SVM also applied which trained well (RF AUC of 1.00 (95% CI 1.00- 1.00) and SVM AUC of 0.96 (95% CI 0.95-0.97). However, not generalisable: AUC SVM 0.55 (95% CI 0.51-0.58); AUC RF 0.81 (95% CI 0.78-0.84).

bmjopen-2020-047755 on 14 September 2021. Downloaded from <http://bmjopen.bmj.com/> on April 20, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

bmjopen-2020-047715
Downloaded from <http://bmjopen.bmj.com/> on April 20, 2024 by guest. Protected by copyright.

Melanoma	Richter ⁴⁰	2019	USA	LR, RF, XGBoost	4,061,172 patients, 10,129 with melanoma (EHR data)	Features extracted from EHR records	5-fold cross-validation	AUC LR 0.76; AUC RF 0.69; AUC XGBoost 0.80	Poor - Good	Smaller amounts of data improved the AUCs
Non-melanoma skin cancer	Roffman ⁴¹	2018	USA	Artificial neural network	1997–2015 NHIS adult survey data, 2,056 NMSC and 460,574 non-cancer cases (survey data)	Gender, age, BMI, diabetes, smoking status, emphysema, asthma, ethnicity, Hispanic ethnicity, hypertension, heart diseases, vigorous exercise habits, and history of stroke	30% for validation (752 NMSC cases and 138,172 never cancer cases)	AUC values of 0.81 (training, 95% CI 0.80–0.82) and 0.81 (validation, 95% CI 0.79–0.82)	Good	
Non-melanoma skin cancer	Wang ⁴⁴	2019	Taiwan	Convolutional neural network	1829 patients with nonmelanoma skin cancer as their first diagnosed cancer and 7665 random controls (insurance data)	Age, sex, ICD-9-CM diagnostic codes, World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) prescription codes, and the total numbers of clinical encounters	5-fold cross-validation	AUC 0.89 (0.87–0.91)	Good	
Pancreatic	Zhao ⁴⁴	2011	USA	Bayesian network inference	98 cases and 14,971 controls (EHR data)	Demographics, lifestyle, symptoms, co-morbidities, and lab test results (20 variables)	Null	0.91 (0.87–0.95)	Excellent	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

1
2
3 234 **Abbreviations:** AUC, area under the curve; BMI, body mass index; LR, logistic regression; NB, Gaussian naive Bayes; DT, decision tree; LDA, linear
4
5 235 discriminant analysis; SVM, support vector machine; ANN artificial neural network; RF, random forest; NMSC, non-melanoma skin cancer; ML, machine
6
7 236 learning.
8

9 237
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

For peer review only

238 *Discrimination of the risk models*

239 All studies provided AUC values as an assessment of model performance. The majority of models
240 (7/10)^{31 37-42 44} showed “good” performance, two had “excellent” performance,^{45 46} and one “failed”.⁴³
241 The two models showing excellent performance were the Bayesian network inference model
242 developed by Zhao et al.,⁴⁶ which used 20 demographic, lifestyle, symptom, co-morbidity, and lab test
243 results to predict the risk of pancreatic cancer with an AUC of 0.91, and the CRC predictive model
244 developed by Wang et al.,⁴⁵ which used a CNN learning on 1929 features (1099 ICD-9 codes and 830
245 ATC codes). The models that “failed” were the range of models (logistic regression, Gaussian naive
246 Bayes, DT, LDA, SVM, and feed-forward ANN) developed by Stark et al.⁴³; however, as discussed
247 below, although these models only had AUCs between 0.51 and 0.61, two of the models compared
248 favourably with the BRCAT clinical risk tool.

249 *Comparison of the risk models with existing predictive algorithms*

250 Stark et al.⁴³ compared their ML models with an existing clinical prediction tool, the Breast Cancer
251 Risk Prediction Tool (BCRAT; <https://bcrisktool.cancer.gov/>). The BCRAT tool is an implementation
252 of the Gail model,⁴⁸ which is a statistical model that estimates five-year breast cancer risk in women
253 without a personal history of breast cancer and without known mutations in high-risk breast cancer
254 genes such as *BRCA1* and *BRCA2*. In the Gail model, patients self-report their current age, age at
255 menarche, age at first live birth, number of first-degree relatives who have had breast cancer,
256 ethnicity, and number of previous breast biopsies, variables which are weighted within the model by
257 logistic regression.⁴⁸ In addition, BCRAT uses data on a personal history of atypical hyperplasia,
258 where available. Although the AUC values for the models (logistic regression (LR), naïve Bayes, DTs,
259 linear discriminant analysis (LDA), SVM, and an ANN) tested using a broader set of features than
260 BCRAT were only between 0.51 (DT) and 0.61 (LR, LDA, and ANN), four of the six models (LR, NB,
261 LDA, and ANN) outperformed BCRAT (AUC 0.56). Other metrics were also used to assess model
262 performance (sensitivity, specificity, and precision), which were comparable between the ML
263 algorithms and the BCRAT, and both BCRAT and the ML models had low precision (~2%).
264 Furthermore, when comparing the different ML models, LR and LDA produced higher AUCs than the
265 ANN model, despite the potential for ANNs to better model noisy data and complex non-linear
266 functions.⁴⁹ The authors suggested that this might have been due to the limited amount of available
267 training data or the selection of hyperparameters.⁴³ It was observed that (i) the derived ML models

1
2
3 268 using an extended and set of features available in primary care can deliver improvements on current
4
5 269 clinical algorithms; (ii) that adding additional features has a greater impact on improving model
6
7 270 performance (i.e., higher AUC) rather than simply using more complex models; and (iii) that AUC
8
9 271 values must be interpreted in the context of existing methods, such as existing, clinically-used risk
10
11 272 prediction models such as the BCRAT or Gail model, rather than in isolation.

12
13 273 In a systematic review of 52 colorectal cancer models predicting future risk of disease in
14
15 274 asymptomatic individuals,⁵⁰ 37 models reported AUC values, which ranged from 0.65 and 0.75. These
16
17 275 included five models that used routine data exclusively and did not include questionnaires or genetic
18
19 276 biomarkers. In comparison, the AUC values for ColonFlag,^{38 42} an ML model that uses age, gender,
20
21 277 and complete blood count (CBC) features to predict the future occurrence of colorectal cancer up to
22
23 278 12 months prior to diagnosis, were 0.78-0.82.

24
25 279 In another systematic review involving 25 risk prediction models for lung cancer that used only
26
27 280 epidemiological parameters as input (i.e., no laboratory parameters),⁵¹ AUCs ranged between 0.57
28
29 281 and 0.86, which compares to an AUC of 0.86 (in both training and validation cohorts) for the ANN
30
31 282 model developed by Hart et al.³⁷ In their systematic review of 25 melanoma risk prediction models,⁵²
32
33 283 Usher-Smith et al. showed in a summary ROC curve that most models had similar discrimination of
34
35 284 0.76, which compares to the highest AUC of 0.80 achieved using XGBoost ML by Richter et al.⁴⁰

36
37 285

38 39 286 **DISCUSSION**

40
41
42 287

43 44 288 **Strengths and limitations of existing ML approaches**

45
46
47 289 The studies reviewed highlight that several different techniques have successfully been used to
48
49 290 develop models with generally very good discriminative performance. Other strengths of the studies
50
51 291 are the demonstration of how ML can be applied to large-scale insurance and EHR data containing
52
53 292 hundreds or thousands of features in order to build predictive models.

54
55 293 However, the above survey also highlights a number of gaps in the application of ML to predicting the
56
57 294 risk of future cancer in asymptomatic individuals. These can be divided into those relating to: (i) study
58
59 295 populations; (ii) model types and comparisons; and (iii) model validation and comparisons.

1
2
3 296 *Study populations*
4

5 297 To date, ML techniques have only been applied to or validated in datasets from developed countries,
6
7 298 representing a fraction of the overall global population and their dietary and lifestyle factors. Given
8
9 299 that the aetiology of cancer, risk factors, and genetics differ in different populations,⁵³ models
10
11 300 developed in populations in high-income countries may not be generalisable to those from low- and
12
13 301 middle-income countries (LMICs). The development and validation of models in LMICs could have
14
15 302 two advantages: first, it would determine the generalizability (and therefore utility) of that model in
16
17 303 other populations, better serving the needs of individuals in LMICs; second, disparities between
18
19 304 models developed in different geographical settings could provide valuable new information about
20
21 305 factors contributing to cancer risk. Generalizing risk prediction models is likely to be challenging, since
22
23 306 resource-poor countries often do not have the necessary infrastructure nor the epidemiological
24
25 307 research capabilities of institutions in high-income countries.

26
27 308 Furthermore, current ML models predict the risk of a limited number of cancer types. Although breast,
28
29 309 colorectal, and lung cancer are the three most common cancers and therefore account for a large
30
31 310 proportion of overall cancer burden, it is still important to detect all cancers early. This is especially
32
33 311 true for those cancers that are usually silent (asymptomatic) for long periods of time, present late with
34
35 312 advanced-stage disease, and for which there are currently no screening programs in place, such as
36
37 313 ovarian and pancreatic cancer. Predicting future risk of these cancers could allow closer monitoring of
38
39 314 at-risk individuals.

40
41 315 *Model types and comparisons*
42

43 316 A wide variety of ML methodologies have been applied and, despite being applied to the same
44
45 317 research problem, this scoping review has not identified a single 'best' method. Two issues arose in
46
47 318 studies that compared different ML approaches on the same datasets. First, although different models
48
49 319 had similar AUCs during training, not all models generalised well to validation datasets; robust model
50
51 320 validation is therefore important to ensure model validity (see below). Second, although in general it is
52
53 321 assumed that larger amounts of training data improve model performance,⁵⁴ Richter et al.⁴⁰ found that
54
55 322 equivalent or even better model performance was achievable using reduced datasets (hundreds of
56
57 323 thousands vs. millions of datapoints). This might be due to high levels of homogeneity in the "no
58
59 324 cancer" class, resulting in fewer instances being required to produce a generalisable model, or as a
60

1
2
3 325 result of overfitting. The requirement for less data for the cancer prediction problem could make ML
4
5 326 techniques more accessible to researchers without extensive computing infrastructure and allow
6
7 327 smaller datasets to be leveraged for model construction.
8

9 328 *Model validation and comparisons*

10
11 329 With the exception of the two studies evaluating a previously defined algorithm for colorectal cancer,
12
13 330 no other study used external validation datasets to assess model generalizability, instead opting for
14
15 331 either a single holdout validation sample or 5-fold cross-validation. While useful for assessing
16
17 332 overfitting,⁵⁵ these approaches do not account for population bias in the training dataset nor
18
19 333 differences in other target populations. Studies seeking to develop ML models should seek to
20
21 334 validate models in independent populations, recognizing that an advantage of an 'ungeneralisable'
22
23 335 model might be insights into cancer risk in other populations. Furthermore, since physicians may code
24
25 336 diseases in EHRs differently over time (for instance, due to altered management or incentives), even
26
27 337 initially generalisable models may need re-validation over time^{23 56}.
28

29 338

30 31 32 339 **Implications for clinical practice**

33
34 340 The ML models described in this scoping review generally show very good performance. So, are any
35
36 341 of these models ready for clinical use? The ColonFlag model^{38 42} is an example has recently been
37
38 342 implemented at Barts Health NHS Trust⁵⁷ to identify patients at particularly high risk of CRC,
39
40 343 particularly as clinicians struggle to prioritise patients in the backlog created by the COVID-19
41
42 344 pandemic. The ColonFlag model is the only model identified in this scoping review that has
43
44 345 undergone extensive external validation in independent datasets.

45
46 346 New ML models need to be contextualised with currently available best clinical practice in order to
47
48 347 fully evaluate their potential clinical value. Comparing the relatively poor AUC values of the Stark et
49
50 348 al.⁴³ models with BCRAT revealed that they in fact outperformed it in many cases. In their comparison
51
52 349 of their ANN with screening methods for lung cancer such as low-dose CT, chest X-ray, and sputum
53
54 350 cytology, Hart et al.³⁷ noted that (according to sensitivity and specificity) it outperformed most of the
55
56 351 other available non-invasive methods. Thorough side-by-side comparisons of newly developed
57
58 352 models with other prediction tools would be helpful in establishing future clinical utility.
59
60

353

354 Unanswered questions and future research

355 Although the few models that are currently available are methodologically diverse, rarely validated in
356 independent datasets to ensure generalisability, and do not cover all cancer types. Even if ML
357 techniques offer only small improvements in cancer detection rates, these improvements are likely to
358 be of high clinical significance given the large size of the global population with or at high risk of
359 cancer and the high mortality and costs associated with late cancer diagnoses.

360 However, the scoping review identifies a number of research gaps that will need to be addressed in
361 order to deliver validated ML-based models to assist clinical decision-making. Firstly, future studies
362 must take steps to establish model generalisability through validation in independent cohorts,
363 including those from LMICs. Although the latter may be challenging, it could be argued that even
364 negative generalisability studies might provide an opportunity to learn more about cancer risk factors
365 in different populations. Secondly, the scoping review fails to establish which ML approach best suits
366 the cancer prediction problem but does show that, where possible, side-by-side comparisons of
367 different methods can reveal important information about generalisability as well as performance and
368 that these comparisons are desirable whenever possible. Thirdly, many important cancer types,
369 particularly 'silent killers' like ovarian cancer, have currently not been the subject of ML modelling
370 approaches; ML could provide an important, low-cost, non-invasive method to identify individuals at
371 high risk of clinically silent cancers that require closer monitoring. Furthermore, it might not
372 necessarily be true that more data equals improved model performance, which might broaden
373 accessibility of model development to a wider range of clinicians and epidemiologists. Finally, ML
374 models need to be compared to best available clinical tools so that their potential clinical utility is
375 transparent.

376

377 CONCLUSIONS

378 In conclusion, this scoping review highlights that the application of ML to cancer prediction is a
379 nascent field, with the majority of the few available studies published in the last two years.

380 Nevertheless, most of ML model performance appears to be good which makes them reliable

1
2
3 381 approach. We hope that the identified research gaps focus future research efforts to deliver validated
4
5 382 ML-based models to assist and improve clinical decision-making.
6

7 383

8
9 384 **Acknowledgment**

10
11 385 The authors would like to thank Dr. Alvina Lai, for her feedback in this scoping review.
12
13 386

14
15 387 **Contributorship statement**

16
17 388 Author AA defined the research question of the scoping review, conducted the literature search, and
18
19 389 summarised the findings. Author HK provided academic guidance as domain expert in machine
20
21 390 learning for healthcare and revised the draft of the manuscript.
22

23 391

24
25 392 **Competing interests**

26
27 393 Non competing interests
28

29 394

30
31 395 **Funding**

32
33 396 No funding received
34

35 397

36
37 398 **Data sharing statement**

38
39 399 Only public published papers were used. No confidential data
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

400 REFERENCES

- 401 1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;69(1):7-34. doi:
402 10.3322/caac.21551
- 403 2. Cancer Research UK. Cancer Statistics for the UK 2020 [Available from:
404 <https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk>.
405 3. Cancer Research UK. Why are cancer rates increasing? 2014 [Available from:
406 <https://scienceblog.cancerresearchuk.org/2015/02/04/why-are-cancer-rates-increasing/>.
407 4. World Health Organization. Early detection of cancer 2016 [Available from:
408 <https://www.who.int/cancer/detection/en/>.
409 5. World Health Organization. Early cancer diagnosis saves lives, cuts treatment costs 2017
410 [Available from: [https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-](https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-saves-lives-cuts-treatment-costs)
411 [saves-lives-cuts-treatment-costs](https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-saves-lives-cuts-treatment-costs).
412 6. Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor
413 DNA. *Sci Transl Med* 2017;9(403) doi: 10.1126/scitranslmed.aan2415
414 7. Bannister N, Broggio J. Cancer survival by stage at diagnosis for England (experimental statistics):
415 adults diagnosed 2012, 2013 and 2014 and followed up to 2015. *Produced in collaboration*
416 *with Public Health England* 2016
417 8. Canary Foundation. Early Detection Facts and Figures Early Detection Works. California 2019.
418 9. Cancer Research UK. Why is early diagnosis important? 2018 [Available from:
419 [https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-](https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important)
420 [important](https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important).
421 10. Blumen H, Fitch K, Polkus V. Comparison of Treatment Costs for Breast Cancer, by Tumor Stage
422 and Type of Service. *Am Health Drug Benefits* 2016;9(1):23-32.
423 11. Kakushadze Z, Raghubanshi R, Yu W. Estimating cost savings from early cancer diagnosis. *Data*
424 2017;2(3):30.
425 12. Cancer Research UK. Saving lives, averting costs 2014 [Available from:
426 https://www.cancerresearchuk.org/sites/default/files/saving_lives_averting_costs.pdf.
427 13. Weller DP, Campbell C. Uptake in cancer screening programmes: a priority in cancer control. *Br J*
428 *Cancer* 2009;101 Suppl 2:S55-9. doi: 10.1038/sj.bjc.6605391

- 1
2
3 429 14. Cancer Research UK. About cancer screening 2020 [Available from:
4
5 430 <https://www.cancerresearchuk.org/about-cancer/screening>.
6
7 431 15. Smith RA, Andrews KS, Brooks D, et al. Cancer screening in the United States, 2019: A review of
8
9 432 current American Cancer Society guidelines and current issues in cancer screening. *CA*
10
11 433 *Cancer J Clin* 2019;69(3):184-210. doi: 10.3322/caac.21557
12
13 434 16. Joseph DA, King JB, Dowling NF, et al. Vital signs: colorectal cancer screening test use—United
14
15 435 States, 2018. *Morbidity and Mortality Weekly Report* 2020;69(10):253.
16
17 436 17. Power E, Wardle J. Change in public awareness of symptoms and perceived barriers to seeing a
18
19 437 doctor following Be Clear on Cancer campaigns in England. *British Journal of Cancer*
20
21 438 2015;112(1):S22-S26.
22
23 439 18. Smith LK, Pope C, Botha JL. Patients' help-seeking experiences and delay in cancer
24
25 440 presentation: a qualitative synthesis. *Lancet* 2005;366(9488):825-31. doi: 10.1016/S0140-
26
27 441 6736(05)67030-4
28
29 442 19. Balasooriya-Smeekens C, Walter FM, Scott S. The role of emotions in time to presentation for
30
31 443 symptoms suggestive of cancer: a systematic literature review of quantitative studies.
32
33 444 *Psychooncology* 2015;24(12):1594-604. doi: 10.1002/pon.3833
34
35 445 20. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? A qualitative study using
36
37 446 national data. *J Gen Intern Med* 2015;30(3):290-7. doi: 10.1007/s11606-014-3089-1
38
39 447 21. American Cancer Society. Lifetime risk of developing or dying from cancer. 2014
40
41 448 22. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk
42
43 449 prediction models with electronic health records data: a systematic review. *J Am Med Inform*
44
45 450 *Assoc* 2017;24(1):198-208. doi: 10.1093/jamia/ocw042
46
47 451 23. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw Open*
48
49 452 2018;1(4):e181404. doi: 10.1001/jamanetworkopen.2018.1404
50
51 453 24. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380(14):1347-
52
53 454 58. doi: 10.1056/NEJMra1814259
54
55 455 25. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective
56
57 456 Study Using Statewide Electronic Health Records and Machine Learning. *J Med Internet Res*
58
59 457 2018;20(1):e22. doi: 10.2196/jmir.9268
60

- 1
2
3 458 26. Hao S, Fu T, Wu Q, et al. Estimating One-Year Risk of Incident Chronic Kidney Disease:
4
5 459 Retrospective Development and Validation Study Using Electronic Medical Record Data From
6
7 460 the State of Maine. *JMIR Med Inform* 2017;5(3):e21. doi: 10.2196/medinform.7954
8
9 461 27. Martin Noguero T, Paulano-Godino F, Martin-Valdivia MT, et al. Strengths, Weaknesses,
10
11 462 Opportunities, and Threats Analysis of Artificial Intelligence and Machine Learning
12
13 463 Applications in Radiology. *J Am Coll Radiol* 2019;16(9 Pt B):1239-47. doi:
14
15 464 10.1016/j.jacr.2019.05.047
16
17 465 28. Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology - new tools for
18
19 466 diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019;16(11):703-15. doi:
20
21 467 10.1038/s41571-019-0252-y
22
23 468 29. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference,
24
25 469 and prediction: Springer Science & Business Media 2009.
26
27 470 30. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis
28
29 471 and prediction. *Comput Struct Biotechnol J* 2015;13:8-17. doi: 10.1016/j.csbj.2014.11.005
30
31 472 31. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling
32
33 473 cancer risk using structured clinical data. *Artif Intell Med* 2018;90:1-14. doi:
34
35 474 10.1016/j.artmed.2018.06.002
36
37 475 32. Munn Z, Peters MDJ, Stern C, et al. Systematic review or scoping review? Guidance for authors
38
39 476 when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*
40
41 477 2018;18(1):143. doi: 10.1186/s12874-018-0611-x
42
43 478 33. Peters MD. In no uncertain terms: the importance of a defined objective in scoping reviews. *JBIR*
44
45 479 *Database System Rev Implement Rep* 2016;14(2):1-4. doi: 10.11124/jbisrir-2016-2838
46
47 480 34. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018;319(13):1317-
48
49 481 18. doi: 10.1001/jama.2017.18391
50
51 482 35. Waegeman W, De Baets B, Boullart L. ROC analysis in ordinal regression learning. *Pattern*
52
53 483 *Recognition Letters* 2008;29(1):1-9.
54
55 484 36. Tape GE. The Area Under an ROC Curve 2019 [Available from:
56
57 485 <http://gim.unmc.edu/dxtests/roc3.htm>.
58
59 486 37. Hart GR, Roffman DA, Decker R, et al. A multi-parameterized artificial neural network for lung
60
487 cancer risk prediction. *PLoS One* 2018;13(10):e0205264. doi: 10.1371/journal.pone.0205264

- 1
2
3 488 38. Hornbrook MC, Goshen R, Choman E, et al. Early Colorectal Cancer Detected by Machine
4
5 489 Learning Model Using Gender, Age, and Complete Blood Count Data. *Dig Dis Sci*
6
7 490 2017;62(10):2719-27. doi: 10.1007/s10620-017-4722-8
8
9 491 39. Miotto R, Li L, Kidd BA, et al. Deep Patient: An Unsupervised Representation to Predict the Future
10
11 492 of Patients from the Electronic Health Records. *Sci Rep* 2016;6:26094. doi:
12
13 493 10.1038/srep26094
14
15 494 40. Richter AN, Khoshgoftaar TM. Efficient learning from big data for cancer risk modeling: A case
16
17 495 study with melanoma. *Comput Biol Med* 2019;110:29-39. doi:
18
19 496 10.1016/j.combiomed.2019.04.039
20
21 497 41. Roffman D, Hart G, Girardi M, et al. Predicting non-melanoma skin cancer via a multi-
22
23 498 parameterized artificial neural network. *Sci Rep* 2018;8(1):1701. doi: 10.1038/s41598-018-
24
25 499 19907-9
26
27 500 42. Schneider J, Layefsky E, Udaltsova N, et al. Validation of an Algorithm to Identify Patients at Risk
28
29 501 for Colorectal Cancer Based on Laboratory Test and Demographic Data in Diverse,
30
31 502 Community-Based Population. *Clin Gastroenterol Hepatol* 2020 doi:
32
33 503 10.1016/j.cgh.2020.04.054
34
35 504 43. Stark GF, Hart GR, Nartowt BJ, et al. Predicting breast cancer risk using personal health data and
36
37 505 machine learning models. *PLoS One* 2019;14(12):e0226765. doi:
38
39 506 10.1371/journal.pone.0226765
40
41 507 44. Wang HH, Wang YH, Liang CW, et al. Assessment of Deep Learning Using Nonimaging
42
43 508 Information and Sequential Medical Records to Develop a Prediction Model for
44
45 509 Nonmelanoma Skin Cancer. *JAMA Dermatol* 2019 doi: 10.1001/jamadermatol.2019.2335
46
47 510 45. Wang YH, Nguyen PA, Islam MM, et al. Development of Deep Learning Algorithm for Detection of
48
49 511 Colorectal Cancer in EHR Data. *Stud Health Technol Inform* 2019;264:438-41. doi:
50
51 512 10.3233/SHTI190259
52
53 513 46. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian
54
55 514 network for pancreatic cancer prediction. *J Biomed Inform* 2011;44(5):859-68. doi:
56
57 515 10.1016/j.jbi.2011.05.004
58
59 516 47. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for
60
517 detection of colorectal cancer in primary care by analysis of complete blood counts: a

- 1
2
3 518 binational retrospective study. *J Am Med Inform Assoc* 2016;23(5):879-90. doi:
4
5 519 10.1093/jamia/ocv195
6
7 520 48. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast
8
9 521 cancer for white females who are being examined annually. *J Natl Cancer Inst*
10
11 522 1989;81(24):1879-86. doi: 10.1093/jnci/81.24.1879
12
13 523 49. Lorena AC, Jacintho LF, Siqueira MF, et al. Comparing machine learning classifiers in potential
14
15 524 distribution modelling. *Expert Systems with Applications* 2011;38(5):5268-75.
16
17 525 50. Usher-Smith JA, Walter FM, Emery JD, et al. Risk Prediction Models for Colorectal Cancer: A
18
19 526 Systematic Review. *Cancer Prev Res (Phila)* 2016;9(1):13-26. doi: 10.1158/1940-
20
21 527 6207.CAPR-15-0274
22
23 528 51. Gray EP, Teare MD, Stevens J, et al. Risk Prediction Models for Lung Cancer: A Systematic
24
25 529 Review. *Clin Lung Cancer* 2016;17(2):95-106. doi: 10.1016/j.clcc.2015.11.007
26
27 530 52. Usher-Smith JA, Emery J, Kassianos AP, et al. Risk prediction models for melanoma: a
28
29 531 systematic review. *Cancer Epidemiol Biomarkers Prev* 2014;23(8):1450-63. doi:
30
31 532 10.1158/1055-9965.EPI-14-0295
32
33 533 53. Rastogi T, Hildesheim A, Sinha R. Opportunities for cancer epidemiology in developing countries.
34
35 534 *Nat Rev Cancer* 2004;4(11):909-17. doi: 10.1038/nrc1475
36
37 535 54. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a
38
39 536 simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
40
41 537 doi: 10.1186/1471-2288-14-137
42
43 538 55. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995.
44
45 539 Montreal, Canada.
46
47 540 56. Bergquist SL, Brooks GA, Keating NL, et al. Classifying Lung Cancer Severity with Ensemble
48
49 541 Machine Learning in Health Care Claims Data. *Proc Mach Learn Res* 2017;68:25-38.
50
51 542 57. Downing M. Barts Health using AI to prioritise care for colon cancer patients 2020 [Available from:
52
53 543 [https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-](https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-colon-cancer-patients-8867)
54
55 544 [colon-cancer-patients-8867](https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-colon-cancer-patients-8867)].
56
57 545

546 FIGURE LEGEND

547 **Figure 1.** PRISMA flowchart depicting the search strategy.

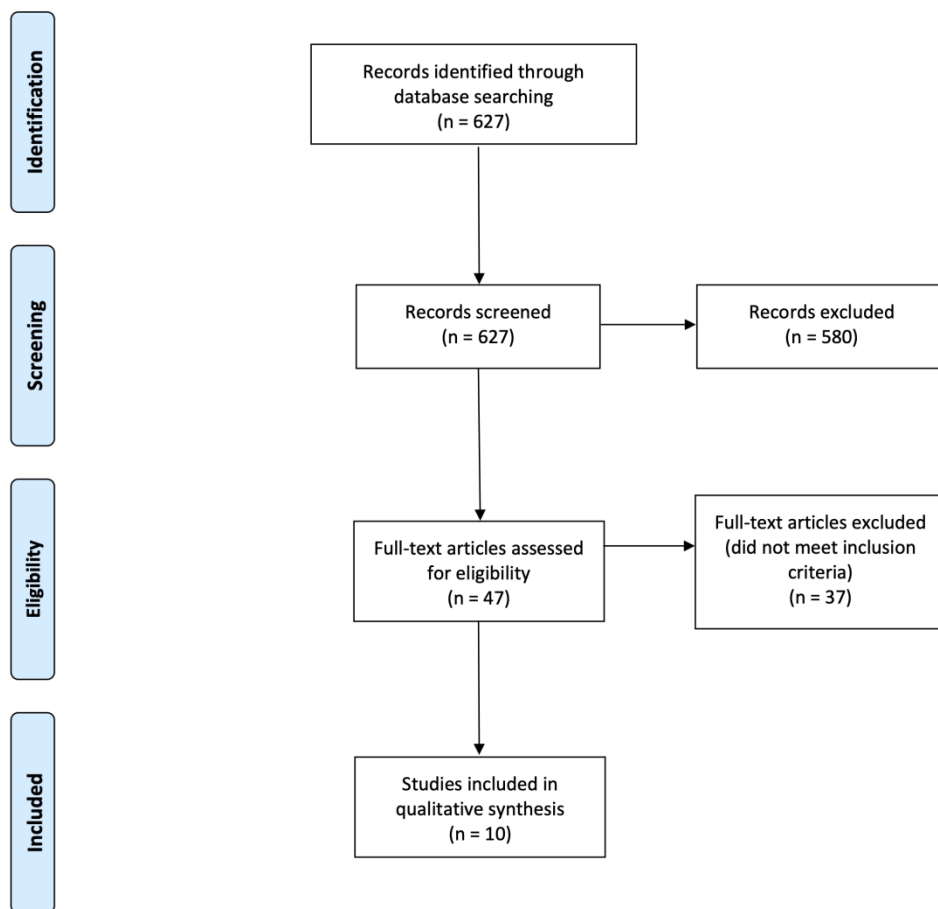


Figure 1. PRISMA flowchart depicting the search strategy.

190x181mm (300 x 300 DPI)

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

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	7
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	7
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	NA
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	8
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	8
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	8
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	8 and 10
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	NA
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	12, 13, and 14
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	NA



SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	8 and 9
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	10
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	12, 13, and 14
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	NA
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	12, 13, and 14
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	10, 11, 16, and 17
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	17-20
Limitations	20	Discuss the limitations of the scoping review process.	17
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	20-21
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	21

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* 2018;169:467–473. doi: 10.7326/M18-0850.



BMJ Open

Predicting the risk of cancer in adults using supervised machine learning: a scoping review

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2020-047755.R1
Article Type:	Original research
Date Submitted by the Author:	22-Aug-2021
Complete List of Authors:	Alfayez, Asma; University College London; King Abdullah International Medical Research Center Kunz, Holger; University College London Lai, Alvina; University College London
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Oncology
Keywords:	Information technology < BIOTECHNOLOGY & BIOINFORMATICS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, ONCOLOGY

SCHOLARONE™
Manuscripts



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.

1
2
3 **Predicting the risk of cancer in adults using supervised machine learning: a scoping review**
4

5 Asma Abdullah Alfayez^{1,2,3}, Holger Kunz¹, Alvina Grace Lai¹
6
7

8 ¹Institute of Health Informatics, University College London, London, UK

9 ²King Abdullah International Medical Research Center, Riyadh, Saudi Arabia

10 ³King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

11 ***Corresponding author:** Asma Abdullah Alfayez (asma.alfayez.17@ucl.ac.uk); Alvina Grace Lai
12
13 (alvina.lai@ucl.ac.uk)
14
15

16 **Word count:** 5158
17

18 **Keywords:** Machine learning; cancer; prediction; scoping review; artificial intelligence
19
20

21 **Competing interests:** None
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

1 **ABSTRACT**

2 **Objectives:** The purpose of this scoping review is to: (i) identify existing supervised machine learning
3 (ML) approaches on the prediction of cancer in asymptomatic adults; (ii) to compare the performance
4 of ML models with each other, and (iii) to identify potential gaps in research.

5 **Design:** Scoping review using the population, concept, and context approach.

6 **Search strategy:** PubMed search engine was used from inception through to November 10, 2020 to
7 identify literature meeting following inclusion criteria: (i) a general adult (≥ 18 years) population, either
8 sex, asymptomatic (*population*); (ii) any study using ML techniques to derive predictive models for
9 future cancer risk using clinical and/or demographic and/or basic laboratory data (*concept*); and (iii)
10 original research articles conducted in all settings in any region of the world (*context*).

11 **Results:** The search returned 627 unique articles, of which 580 articles were excluded because they
12 did not meet the inclusion criteria, were duplicates, or were related to benign neoplasm. Full-text
13 reviews were conducted for 47 articles and a final set of 10 articles were included in this scoping
14 review. These 10 very heterogeneous studies used ML to predict future cancer risk in asymptomatic
15 individuals. All studies reported area under the receiver operating characteristics curve (AUC) values
16 as metrics of model performance, but no study reported measures of model calibration.

17 **Conclusions:** Research gaps that must be addressed in order to deliver validated ML-based models
18 to assist clinical decision-making include: (i) establishing model generalisability through validation in
19 independent cohorts, including those from low- and middle-income countries; (ii) establishing models
20 for all cancer types; (iii) thorough comparisons of ML models with best available clinical tools to
21 ensure transparency of their potential clinical utility; (iv) reporting of model calibration performance;
22 and (v) comparisons of different methods on the same cohort to reveal important information about
23 model generalisability and performance.

24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60

26 **ARTICLE SUMMARY**

27 **Strengths and limitations of this study**

- 1
2
3 28 • This study used a recognised scoping review approach (population, concept, and context) to
4
5 29 explore the machine learning techniques used to derive predictive models for future cancer
6
7 30 risk.
8
9 31 • Identified studies were not subjected to comprehensive qualitative assessments.
10
11 32 • Only ten studies were identified, making it difficult to draw firm conclusions about their relative
12
13 33 performance.
14
15 34 • AUC values alone do not allow for meaningful comparisons of models as they have been
16
17 35 trained and evaluated on different datasets under different circumstances and conditions.
18
19 36 • This scoping review is limited to papers published in English until 2020 and only the PubMed
20
21 37 search engine was used.
22
23 38
24
25 39

40 INTRODUCTION

41 Cancer remains a leading cause of morbidity and mortality, with an estimated 1.8 million new cases
42 and 0.6 million deaths in the US in 2019 and approximately 367,000 new cases and 165,000 cancer
43 deaths in the UK each year between 2015 and 2017.^{1,2} Annual death rates only modestly decreased
44 (1.4% and 1.8% in women and men, respectively) between 2012 and 2016, despite significant
45 research.¹ Cancer cases also continue to increase, not least due to increased life expectancy, which
46 increases the risk of developing cancer.³

47 Early cancer diagnosis is associated with significantly higher survival rate and lower mortality and
48 associated costs. Early-stage cancers require less complex treatment regimens and reduced hospital
49 utilization, resulting in reduced healthcare costs, whereas late-stage cancers require complex
50 multimodal management, several rounds of extremely expensive drugs over significant periods of
51 time, and the treatment of recurrences, equating to a staggering economic burden. Therefore, the
52 importance of early diagnosis cannot be overestimated.⁴⁻⁶ Treating cancer early has significant cost-
53 saving benefits. In the US, during the first 24 months after diagnosis, there is an increase in cancer
54 treatment costs with stage: US\$72,000 for stage 0, US\$97,000 for stage I/II, US\$159,000 for stage III,
55 and \$182,000 for stage IV.⁷ An estimate of the cost savings from early cancer diagnosis is 26 billion
56 US dollars per annum in the US alone.⁸ Similarly, in the UK, early diagnosis of colorectal, ovarian,

1
2
3 57 and lung cancer in England alone could provide savings of over £44 million and benefit nearly 11,000
4
5 58 patients.⁹
6

7 59 Survival rates significantly improve if cancer is diagnosed at stage I or II compared with later stages
8
9 60 (stage III and IV),^{10 11} as once the cancer has metastasised, it becomes difficult to treat with
10
11 61 radiotherapy or surgery, leading to treatment failure and death. For example, five-year survival rates
12
13 62 for women diagnosed with localised breast or ovarian cancer are 99% and 92% compared to 27%
14
15 63 and 29% for metastatic disease, respectively.¹ A report by Cancer Research UK indicated that, in the
16
17 64 UK, the ten-year survival proportions of patients with eight cancers (combined) were around 80% for
18
19 65 stage I and stage II detection (breast, bladder, ovarian, colorectal, uterine, testicular, and cervical
20
21 66 cancer and malignant melanoma) but only 26% for cancers detected at later stages, notably lung
22
23 67 cancer (stage III and IV).¹²
24

25 68 **Current approaches to diagnose incident cancer**

26
27
28 69 One approach to the early detection of cancer is population-wide screening, which aims to find
29
30 70 asymptomatic individuals so that they can be promptly referred for treatment. Examples include
31
32 71 mammography for breast cancer, cervical screening for cervical cancer, and faecal occult blood
33
34 72 testing or sigmoidoscopy for colorectal cancer.¹³ There are three examples of national screening
35
36 73 programs in UK (bowel, breast, and cervical cancer screening programs¹⁴) and two in the US: the
37
38 74 Colorectal Cancer Control Program (CRCCP) and the National Breast and Cervical Cancer Early
39
40 75 Detection Program (NBCCEDP).¹⁵ However, significant proportions of individuals eligible for these
41
42 76 programs do not participate (for example through fear or not prioritizing time to attend for screening),¹⁶
43
44 77 and comprehensive screening programs are costly to implement, especially in resource-poor settings
45
46 78 or low- and middle-income countries. Other approaches include public health campaigns to
47
48 79 encourage individuals experiencing particular symptoms such as weight loss, anorexia, and fatigue to
49
50 80 visit their family doctors.¹⁷ However, patient help-seeking around cancer is complex, multi-staged, and
51
52 81 often leads to long delays of weeks or even months.¹⁸ Patients find it hard to interpret and recognise
53
54 82 symptoms, with fears of embarrassment and having a potentially fatal or painful condition contributing
55
56 83 to long and avoidable delays in help-seeking from health professionals.^{18 19} Patients often do not seek
57
58 84 help from health professionals for early cancer symptoms, notably from general family physicians, for
59
60

1
2
3 85 many reasons including a complex mix of fear, worry, and of 'wasting' health professionals' time¹⁹ or
4
5 86 due to the high costs of medical care, a lack of health insurance, or time constraints.²⁰
6
7

8 87 **Detecting future risk of cancer by modelling data**

9
10 88 Screening approaches represent a patient identification (or "phenotyping" problem) that aims to detect
11
12 89 whether the individual has cancer at a particular point in time. However, the ultimate goal of cancer
13
14 90 prediction is to determine whether an individual will develop cancer at some point in the future. A
15
16 91 simple approach is to stratify populations according to the presence and absence of risk factors,
17
18 92 which have been extensively characterised for most cancer types through epidemiological studies
19
20 93 over many decades. For example, age, gender, ethnicity, family history, and lifestyle factors are well-
21
22 94 established risk factors for many types of cancer.²¹ The cancer prediction problem can either be
23
24 95 regarded as a regression problem, where the input variables are clinical-demographic variables and
25
26 96 the output variable is the probability of developing cancer at some point in the future, or as a binary
27
28 97 classification problem to determine whether or not a patient will develop cancer at a specific point in
29
30 98 time.

31 99 **Big data and machine learning for medical prediction models**

32
33
34 100 Advances in digital medicine and computational science have altered the landscape of data available
35
36 101 for cancer risk prediction models. For example, in the data-driven healthcare era, there is an
37
38 102 increasing amount of "big" medical data, as most individuals have had interactions with the healthcare
39
40 103 system where data is collected in the form of electronic health records (EHRs), which are systematic
41
42 104 collections of longitudinal patient health data collected in real time.^{22 23} Such large datasets provide
43
44 105 powerful new opportunities to develop and refine predictive models and to explore potentially
45
46 106 unknown predictor variables.²² Leveraging often massive amounts of data generated from large
47
48 107 populations, much of which may be unstructured, and building optimal models requires the
49
50 108 exploitation of advanced computational tools and supporting infrastructure. Machine learning (ML) is a
51
52 109 branch of artificial intelligence (AI) and an extension of traditional statistical techniques that uses
53
54 110 computational resources to detect underlying patterns in high-dimensional data, and it is increasingly
55
56 111 being used in different areas of medicine requiring predictions.²⁴ For example, ML has successfully
57
58 112 been used with EHR data to predict incident hypertension²⁵ and incident chronic kidney disease,²⁶
59
60

1
2
3 113 and wider popular uses of ML in medicine include the automatic interpretation of medical images such
4
5 114 as in radiology²⁷ and histopathology²⁸ images.

7 115 **A brief description of machine learning**

9
10 116 A comprehensive description of ML models is beyond the scope of this scoping review. However,
11
12 117 relevant ML techniques relate to the problem of *learning* from data samples (e.g., EHR data) rather
13
14 118 than being pre-programmed with existing knowledge or rules. ML models can either be supervised
15
16 119 (i.e., where the data are labelled and the algorithm uses these data to learn to predict the output) or
17
18 120 unsupervised (i.e., where the data are unlabelled and the algorithm learns a structure inherent in the
19
20 121 data).²⁹ The cancer prediction problem is therefore a supervised problem; examples are provided as
21
22 122 inputs (or features) such as cancer risk factors like age, history, ethnicity, or blood count parameters
23
24 123 and outputs (or labels) such as whether or not the individual subsequently develops cancer. A variety
25
26 124 of available algorithms learn the best way to map the features to the labels by learning from the
27
28 125 observations.^{30 31} The resulting model, ideally, will then be able to generalise the information so that it
29
30 126 can be applied with high precision to new and unseen data.^{30 31}

31
32 127 Some of the main supervised ML models used in medical applications include decision trees (DTs;
33
34 128 and their adaptation, random forests (RFs)), support vector machines (SVMs), and artificial neural
35
36 129 networks (ANNs).^{30 31} DTs produce an output similar to a flow chart formed from feature nodes (risk
37
38 130 variables) that best discriminate between different labels (future cancer occurrence) to split the tree.³⁰
39
40 131 ³¹ In this way, new cases can be assessed by traversing the tree based on the feature values to
41
42 132 determine the output for that example.^{30 31} Decision trees are easy to interpret, since users are usually
43
44 133 able to visualise the steps leading to a particular classification, which may be useful in a clinical
45
46 134 setting where experts might wish to see how a particular decision was made.^{30 31} In RFs, several trees
47
48 135 are built using subsets of data and features, with predictions decided based on majority voting after
49
50 136 the example is assessed with respect to all the constructed trees.^{30 31}

51
52 137 In SVMs, each feature (risk factor) is mapped into a higher-dimensional space and the hyperplane
53
54 138 that optimally separates the output (future cancer occurrence) modelled.^{30 31} SVMs tend to generalise
55
56 139 well to unseen data and work well with complex (multidimensional) data but can be hard to interpret.³⁰

57 140 ³¹

1
2
3 141 ANNs are inspired by the neural connections in the human brain and are developed by creating
4
5 142 nodes (neurons) that weight certain features and produce an output value.^{30 31} By layering nodes in
6
7 143 between the input layer (features; cancer risk factors) and output layer (label; future cancer
8
9 144 occurrence) and modifying the weights during learning through a process called back-propagation,
10
11 145 the resulting model forms a prediction for unseen data when one of the nodes in the output layer is
12
13 146 positive.^{30 31} The terms “deep neural network” and “deep learning” are applied to ANNs with large
14
15 147 numbers of layers.^{30 31} While proving extremely powerful across a range of applications, ANNs can be
16
17 148 computationally very expensive and the way in which they classify (i.e., the intermediate “hidden”
18
19 149 layers) is opaque, making it difficult to determine exactly how they performed the classification
20
21 150 problem.^{30 31}

22 23 151 **Rationale for performing a scoping review**

24
25 152 Machine learning remains a relatively recent field, so it is unclear exactly to what extent advances
26
27 153 have impacted specific healthcare domains. There are currently no extended systematic reviews or
28
29 154 scoping reviews on the application of ML to cancer risk prediction in asymptomatic individuals. This
30
31 155 prompted us to perform a scoping review of studies using supervised ML techniques to predict the
32
33 156 future risk of developing cancer or specific cancers within a general asymptomatic adult (≥ 18 years)
34
35 157 population using clinical and/or demographic and/or basic laboratory data (e.g., complete blood
36
37 158 counts) that are likely to be readily available within the primary care setting. This approach therefore
38
39 159 allowed to: (i) identify the types of evidence available; (ii) clarify key concepts and definitions; (iii)
40
41 160 examine how research is currently being conducted; and (iv) to identify knowledge gaps.³²

42 43 161 **OBJECTIVES**

44
45 162 The objective of this study was to perform a scoping review and to synthesize knowledge of the
46
47 163 nature and effects of current ML techniques for early cancer detection in asymptomatic adults. The
48
49 164 scoping review was guided by the following research questions:

- 50
51 165 (i) Which, if any, ML methods are being developed for cancer risk prediction in asymptomatic
52
53 166 individuals in the community?
54
55
56 167 (ii) How do these models perform compare to each other?
57
58 168 (iii) Which research or knowledge gaps need to be addressed in order to advance the field?
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

169

For peer review only

170 **METHODS**

171 **Inclusion and exclusion criteria**

172 We used the population, concept, and context approach³³ with the following inclusion criteria: (i)
173 general adult (≥ 18 years) population, either sex, asymptomatic (*population*); (ii) any study using ML
174 techniques to derive predictive models for future cancer risk using clinical and/or demographic and/or
175 basic laboratory data carried out prior to August 7, 2020 (*concept*); and (iii) original research articles
176 conducted in all settings in any region of the world (*context*).

177 For the purposes of this study, and recognizing that 'machine learning' algorithms fall along a
178 continuum with statistical techniques,³⁴ all modelling approaches were included were defined as
179 machine learning in the respective papers (such as logistic regression).

180 Exclusion criteria were any ML model used to predict future events in patients with pre-existing or
181 symptoms of cancer; ML models developed using specialised tests such as genetic profiling or
182 imaging tests not generally available in the community; unsupervised ML models; and studies not
183 written in English.

184 **Literature search**

185 To identify relevant studies, the PubMed database was searched from inception through to November
186 10, 2020 using the search string: ("Cancer" Or "Cancers" OR "Oncology") AND ("Machine Learning"
187 OR "ML" OR "Data Mining" OR "Decision Support System" OR "Clinical Support System" OR
188 "Classification" OR "Regression" OR "Support vector machines" OR "Gaussian process" OR "Neural
189 networks" OR "Logical learning" OR "Bayesian network" OR "linear model") AND ("prognosis" OR
190 "prognostic estimate" OR "predictor" OR "prediction" OR "model" OR "diagnosis" OR "diagnostic").

191 This search was supplemented with manual searching of the references and citations of previously
192 published studies. All abstracts identified by the initial search were screened for inclusion and
193 checked for accuracy. For the included studies, data were extracted from full papers. In instances
194 where more information was required to determine inclusion, the full text of the article was retrieved
195 and assessed against the eligibility criteria.

196 **Study assessment**

1
2
3 197 The quality of the included studies was assessed using the Newcastle Ottawa Scale (NOS) for
4
5 198 observational studies included in the review.³⁵ The strength of the predictive ability of the included
6
7 199 models was assessed using AUC (area under the receiver operating characteristics curve) data, a
8
9 200 valid measure for evaluating classification algorithms and one that has been used to compare
10
11 201 different algorithms in other meta-analyses.^{36 37}
12

13 202 **Patient and public involvement**

14
15 203 This study was not explicitly informed by patient priorities, experiences, and preferences, although the
16
17 204 application of predictive models to assess cancer risk would have a direct bearing on identifying those
18
19 205 most at risk and implementing investigations in a timely manner. No patients were involved in the
20
21 206 design or conduct of the study and since this was a scoping review of the literature, there were no
22
23 207 study participants.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

208 RESULTS

209 Main Findings

210 *Identified risk models*

211 Using the search strategy, 627 initial studies were identified where 10 studies met the inclusion
212 criteria (Table 1; Figure 1).^{31 38-47} The most common reasons for exclusion of studies were: (i) models
213 were derived to predict outcomes or responses to therapy in patients with pre-existing cancer; and/or
214 (ii) the studies used features other than clinical and/or demographic and/or basic laboratory data,
215 such as genetic biomarkers. All studies were retrospective cohort or case-control studies conducted
216 between 2011 and 2020, with 8 out of 10 studies completed in the last two years. Eight studies were
217 conducted in the USA and two in Taiwan. One model was built for breast cancer, three for colorectal
218 cancer, one for lung cancer, one for melanoma, two for non-melanoma skin cancer, one for
219 pancreatic cancer, and one a general cancer prediction model. Two studies performed external
220 validations of a previously developed colorectal cancer prediction model (Table 1).^{39 43} In terms of
221 quality assessment, four studies were graded as “good” quality by the NOS,^{39 43 44 46} while six studies
222 were graded as “poor”, in all cases due to comparability of cohorts on the basis of the design or
223 analysis adequately controlling for confounders.^{31 38 40 42 45 47}

224 *Development of the risk models*

225 The models developed in the studies employed a wide range of ML techniques. Two studies
226 compared different modelling approaches on the same dataset,^{41 44} while the other eight developed a
227 model using a single approach. The following ML approaches were used: ANNs (8 out of 10 studies),
228 logistic regression (2/10 studies), Gaussian naïve Bayes (1 out of 10 studies), Bayesian network
229 inference (1/10 studies), DTs (1/10 studies) and RFs (2/10 studies), linear discriminant analysis (1/10
230 studies), and SVMs (1/10 studies) (Table 1). Data for training and testing were from medical
231 insurance databases (4/10 studies), EHR data repositories (3/10 studies), surveys (2/10 studies), or
232 represented a retrospective analysis of prospectively collected data from a clinical trial (1/10 studies).
233 As a result of the diverse cancer types being modelled, study aims, and the available data, a range of
234 different predictors, features, and/or risk factors were included the developed predictive models,
235 which can be grouped into the following categories: (1) patient demographic data: e.g., age, gender,
236 ethnicity, family history; (2) social and lifestyle data: e.g., cigarette smoking and intensity of exercise;

1
2
3 237 (3) comorbidities: e.g., diabetes mellitus, hypertension, congestive heart failure, and chronic
4
5 238 obstructive pulmonary disease; (4) clinical and practice data: e.g., Anatomical Therapeutic Chemical
6
7 239 (WHO-ATC) prescription codes and clinical encounters; and (5) laboratory tests: e.g., complete blood
8
9 240 count (Table 1). The models that automatically extracted features from EHR records used features
10
11 241 that were not always explicitly defined in the respective articles.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

242 **Table 1.** Summary of studies investigating ML approaches for early cancer detection.

Type of cancer	Reference	Year	Country	Method	Sample	Input	Validation	Performance	NOS	Notes
Breast	Stark ⁴⁴	2019	USA	Logistic regression, Gaussian naive Bayes, decision tree, linear discriminant analysis, support vector machine, and feed-forward artificial neural network	1343 breast cancer and 63,396 non-breast cancer cases (PLCO dataset)	Age, age at menarche, age at first live birth, number of first-degree relatives who have had breast cancer, ethnicity, age at menopause, an indicator of current hormone usage, number of years of hormone usage, BMI, pack years of cigarettes smoked, years of birth control usage, number of live births, and an indicator of personal prior history of cancer	20% testing data (269 breast cancer and 12,679 non-breast cancer cases)	AUC 0.61 (0.58-0.65); NB 0.59 (0.56-0.62); DT 0.61 (0.50-0.52); LDA 0.61 (0.58-0.65); SVM 0.52 (0.48-0.55); NN 0.61 (0.57-0.64)	9 (Good)	At an 0.05 level, the logistic regression, linear discriminant analysis, and neural network with the broader set of inputs were all significantly stronger than the BCRAT
Colorectal cancer	Hornbrook ³⁹	2017	USA	ColonFlag ML model	17,095 US community-based insured adults (16,195 controls, 900 cases) (insurance data)	Age, gender, and blood count panel parameters	Study was a validation of a previously derived model ⁴⁸	AUC 0.80 (0.79-0.82)	7 (Good)	
Colorectal	Wang ⁴⁶	2019	Taiwan	Convolutional neural network	10,185 with CRC, 47,967 controls	ICD-9-CM diagnostic codes, World Health Organization-	5-fold cross-validation	AUC 0.92	7 (Good)	

bmjopen-2020-047755 on 1 September 2024 by guest. Protected by copyright.

					(insurance data)	Anatomical Therapeutic Chemical (WHO-ATC) prescription codes				
Colorectal cancer	Schneider ⁴³	2020	USA	ColonFlag ML model	308,721 insurance health plan members (insurance data)	Age, gender, and blood count panel parameters	Study was a validation of a previously derived model ⁴⁸	AUC 0.78 (95% CI 0.77-0.78)	8 (Good)	The algorithm's accuracy decreased with the time interval between blood test result and CRC diagnosis
General	Miotto ⁴⁰	2016	USA	Deep neural network and random forests	Model training on 704,587, testing on 76,214 (EHR data)	Features extracted from EHR records	Testing on 76,214	Colorectal cancer AUC 0.89, liver cancer 0.89, prostate cancer 0.86	6 (Poor)	Outperformed RawFeat and PCA
Lung	Hart ³⁸	2018	USA	Artificial neural network	1997-2015 National Health Interview Survey adult data; 648 cancer and 488,418 non-cancer cases (survey data)	Gender, age, BMI, diabetes, smoking status, emphysema, asthma, ethnicity, Hispanic ethnicity, hypertension, heart diseases, vigorous exercise habits, and history of stroke	30% of data; 195 lung cancer cases and 146,524 never cancer cases	AUC 0.86 (training; 95% CI 0.85-0.88) and 0.86 (validation; 95% CI 0.84-0.89)	6 (Poor)	Random forests and SVM also applied which trained well (RF AUC of 1.00 (95% CI 1.00- 1.00) and SVM AUC of 0.96 (95% CI 0.95-0.97). However, not generalisable: AUC SVM 0.55 (95% CI 0.51-0.58); AUC RF 0.81 (95% CI 0.78-0.84).

Melanoma	Richter ⁴¹	2019	USA	LR, RF, XGBoost	4,061,172 patients, 10,129 with melanoma (EHR data)	Features extracted from EHR records	5-fold cross-validation	AUC LR 0.76; AUC RF 0.69; AUC XGBoost 0.80	7 (Poor)	Smaller amounts of data improved the AUCs
Non-melanoma skin cancer	Roffman ⁴²	2018	USA	Artificial neural network	1997–2015 NHIS adult survey data, 2,056 NMSC and 460,574 non-cancer cases (survey data)	Gender, age, BMI, diabetes, smoking status, emphysema, asthma, ethnicity, Hispanic ethnicity, hypertension, heart diseases, vigorous exercise habits, and history of stroke	30% for validation (752 NMSC cases and 138,172 never cancer cases)	AUC values of 0.81 (training, 95% CI 0.80–0.82) and 0.81 (validation, 95% CI 0.79–0.82)	6 (Poor)	
Non-melanoma skin cancer	Wang ⁴⁵	2019	Taiwan	Convolutional neural network	1829 patients with nonmelanoma skin cancer as their first diagnosed cancer and 7665 random controls (insurance data)	Age, sex, ICD-9-CM diagnostic codes, World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) prescription codes, and the total numbers of clinical encounters	5-fold cross-validation	AUC 0.89 (0.87–0.91)	6 (Poor)	
Pancreatic	Zhao ⁴⁷	2011	USA	Bayesian network inference	98 cases and 14,971 controls (EHR data)	Demographics, lifestyle, symptoms, co-morbidities, and lab test results (20 variables)	Null	0.91 (0.87–0.95)	4 (Poor)	

bmjopen-2020-047715
 Protected from <http://bmjopen.bmj.com/> on April 20, 2024 by guest. Protected by copyright.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

1
2
3 243 **Abbreviations:** ANN artificial neural network; AUC, area under the curve; BMI, body mass index; LR, logistic regression; NB, Gaussian naive Bayes; DT,
4
5 244 decision tree; LDA, linear discriminant analysis; ML, machine learning; NMSC, non-melanoma skin cancer; NOS, Newcastle Ottawa Scale; RF, random
6
7 245 forest; SVM, support vector machine.
8

9 246
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

For peer review only

1
2
3 247 *Discrimination and calibration of the risk models*

4
5 248 All studies provided AUC values as an assessment of model performance. Calibration (i.e., whether
6
7 249 the risk estimates were accurate), was not assessed in any study. Two models with particularly high
8
9 250 AUC values were the Bayesian network inference model developed by Zhao et al.,⁴⁷ which used 20
10
11 251 demographic, lifestyle, symptom, co-morbidity, and lab test results to predict the risk of pancreatic
12
13 252 cancer with an AUC of 0.91, and the CRC predictive model developed by Wang et al.,⁴⁶ which used a
14
15 253 CNN learning on 1929 features (1099 ICD-9 codes and 830 ATC codes). Models with particularly low
16
17 254 AUC values were the range of models (logistic regression, Gaussian naive Bayes, DT, LDA, SVM,
18
19 255 and feed-forward ANN) developed by Stark et al.⁴⁴; however, as discussed below, although these
20
21 256 models only had AUCs between 0.51 and 0.61, two of the models compared favourably with the
22
23 257 BRCAT clinical risk tool.

24
25 258 *Comparison of the risk models with existing predictive algorithms*

26
27 259 Hundreds of risk prediction models have been published in the literature for every cancer type, and
28
29 260 some of these are already used in clinical practice. It is therefore important to understand whether the
30
31 261 performance of the newer ML-based cancer risk models is comparable to that of existing predictive
32
33 262 algorithms. We therefore specifically examined whether the studies compared their ML algorithms
34
35 263 with existing algorithms or, if not, how model performance as described by AUCs compared with other
36
37 264 published data, despite the limitations of this approach (see below).

38
39 265 Stark et al.⁴⁴ compared their ML models with an existing clinical prediction tool, the Breast Cancer
40
41 266 Risk Prediction Tool (BCRAT; <https://bcrisktool.cancer.gov/>). The BCRAT tool is an implementation
42
43 267 of the Gail model,⁴⁹ which is a statistical model that estimates five-year breast cancer risk in women
44
45 268 without a personal history of breast cancer and without known mutations in high-risk breast cancer
46
47 269 genes such as *BRCA1* and *BRCA2*. In the Gail model, patients self-report their current age, age at
48
49 270 menarche, age at first live birth, number of first-degree relatives who have had breast cancer,
50
51 271 ethnicity, and number of previous breast biopsies, variables which are weighted within the model by
52
53 272 logistic regression.⁴⁹ In addition, BCRAT uses data on a personal history of atypical hyperplasia,
54
55 273 where available. Although the AUC values for the models (logistic regression (LR), naïve Bayes, DTs,
56
57 274 linear discriminant analysis (LDA), SVM, and an ANN) tested using a broader set of features than
58
59 275 BCRAT were only between 0.51 (DT) and 0.61 (LR, LDA, and ANN), four of the six models (LR, NB,
60
276 LDA, and ANN) outperformed BCRAT (AUC 0.56). Other metrics were also used to assess model

1
2
3 277 performance (sensitivity, specificity, and precision), which were comparable between the ML
4
5 278 algorithms and the BCRAT, and both BCRAT and the ML models had low precision (~2%).
6
7 279 Furthermore, when comparing the different ML models, LR and LDA produced higher AUCs than the
8
9 280 ANN model, despite the potential for ANNs to better model noisy data and complex non-linear
10
11 281 functions.⁵⁰ The authors suggested that this might have been due to the limited amount of available
12
13 282 training data or the selection of hyperparameters.⁴⁴ It was observed that (i) the derived ML models
14
15 283 using an extended and set of features available in primary care can deliver improvements on current
16
17 284 clinical algorithms; (ii) that adding additional features has a greater impact on improving model
18
19 285 performance (i.e., higher AUC) rather than simply using more complex models; and (iii) that AUC
20
21 286 values must be interpreted in the context of existing methods, such as existing, clinically-used risk
22
23 287 prediction models such as the BCRAT or Gail model, rather than in isolation.

24
25 288 In a systematic review of 52 colorectal cancer models predicting future risk of disease in
26
27 289 asymptomatic individuals,⁵¹ 37 models reported AUC values, which ranged from 0.65 and 0.75. These
28
29 290 included five models that used routine data exclusively and did not include questionnaires or genetic
30
31 291 biomarkers. In comparison, the AUC values for ColonFlag,^{39 43} an ML model that uses age, gender,
32
33 292 and complete blood count (CBC) features to predict the future occurrence of colorectal cancer up to
34
35 293 12 months prior to diagnosis, were 0.78-0.82.

36
37 294 In another systematic review involving 25 risk prediction models for lung cancer that used only
38
39 295 epidemiological parameters as input (i.e., no laboratory parameters),⁵² AUCs ranged between 0.57
40
41 296 and 0.86, which compares to an AUC of 0.86 (in both training and validation cohorts) for the ANN
42
43 297 model developed by Hart et al.³⁸ In their systematic review of 25 melanoma risk prediction models,⁵³
44
45 298 Usher-Smith et al. showed in a summary ROC curve that most models had similar discrimination of
46
47 299 0.76, which compares to the highest AUC of 0.80 achieved using XGBoost ML by Richter et al.⁴¹

48
49 300

50 51 301 **DISCUSSION**

52 53 302 **Strengths and limitations of existing ML approaches**

54
55
56 303 The reviewed studies reviewed highlight that several different techniques have successfully been
57
58 304 used to develop models and that ML can be applied to large-scale insurance and EHR data
59
60 305 containing hundreds or thousands of features in order to build predictive models. However, the survey

1
2
3 306 also highlights a number of gaps in the application of ML to predicting the risk of future cancer in
4
5 307 asymptomatic individuals. These can be divided into those relating to: (i) study populations; (ii) model
6
7 308 types and comparisons; and (iii) model validation, comparisons, and calibration.

8
9 309 *Study populations*

10
11 310 To date, ML techniques have only been applied to or validated in datasets from developed countries,
12
13 311 representing a fraction of the overall global population and their dietary and lifestyle factors. Given
14
15 312 that the aetiology of cancer, risk factors, and genetics differ in different populations,⁵⁴ models
16
17 313 developed in populations in high-income countries may not be generalisable to those from low- and
18
19 314 middle-income countries (LMICs). The development and validation of models in LMICs could have
20
21 315 two advantages: first, it would determine the generalizability (and therefore utility) of that model in
22
23 316 other populations, better serving the needs of individuals in LMICs; second, disparities between
24
25 317 models developed in different geographical settings could provide valuable new information about
26
27 318 factors contributing to cancer risk. Generalizing risk prediction models is likely to be challenging, since
28
29 319 resource-poor countries often do not have the necessary infrastructure nor the epidemiological
30
31 320 research capabilities of institutions in high-income countries.

32
33 321 Furthermore, current ML models predict the risk of a limited number of cancer types. Although breast,
34
35 322 colorectal, and lung cancer are the three most common cancers and therefore account for a large
36
37 323 proportion of overall cancer burden, it is still important to detect all cancers early. This is especially
38
39 324 true for those cancers that are usually silent (asymptomatic) for long periods of time, present late with
40
41 325 advanced-stage disease, and for which there are currently no screening programs in place, such as
42
43 326 ovarian and pancreatic cancer. Predicting future risk of these cancers could allow closer monitoring of
44
45 327 at-risk individuals.

46
47 328 *Model types and comparisons*

48
49 329 A wide variety of ML methodologies have been applied and, despite being applied to the same
50
51 330 research problem, this scoping review has not identified a single 'best' method. Two issues arose in
52
53 331 studies that compared different ML approaches on the same datasets. First, although different models
54
55 332 had similar AUCs during training, not all models generalised well to validation datasets; robust model
56
57 333 validation is therefore important to ensure model validity (see below). Second, although in general it is
58
59 334 assumed that larger amounts of training data improve model performance,⁵⁵ Richter et al.⁴¹ found that

1
2
3 335 equivalent or even better model performance was achievable using reduced datasets (hundreds of
4
5 336 thousands vs. millions of datapoints). This might be due to high levels of homogeneity in the “no
6
7 337 cancer” class, resulting in fewer instances being required to produce a generalisable model, or as a
8
9 338 result of overfitting. Although the requirement for less data for the cancer prediction problem could
10
11 339 make ML techniques more accessible to researchers without extensive computing infrastructure and
12
13 340 allow smaller datasets to be leveraged for model construction, ML requires over ten-times the amount
14
15 341 of data per variable for stable discrimination compared with traditional approaches such as logistic
16
17 342 regression.⁵⁵ Instead of regarding data requirements as “too high” or “too low”, it might be better to
18
19 343 consider how much data is required for a particular predictive context. Riley et al.⁵⁶ recently provided
20
21 344 an implementation of how to calculate the sample size required to develop specific clinical prediction
22
23 345 models, which will help researchers prospectively plan their *in silico* experiments and avoid using
24
25 346 datasets that are too small for the total number of participants or outcome events.

26
27 34728
29 348 *Model validation, comparisons, and performance evaluation*

30
31 349 With the exception of the two studies evaluating a previously defined algorithm for colorectal cancer,
32
33 350 no other study used external validation datasets to assess model generalizability, instead opting for
34
35 351 either a single holdout validation sample or 5-fold cross-validation. While useful for assessing
36
37 352 overfitting,⁵⁷ these approaches do not account for population bias in the training dataset nor
38
39 353 differences in other target populations. Studies seeking to develop ML models should seek to validate
40
41 354 models in independent populations, recognizing that an advantage of an ‘ungeneralisable’ model
42
43 355 might be insights into cancer risk in other populations. Furthermore, since physicians may code
44
45 356 diseases in EHRs differently over time (for instance, due to altered management or incentives), even
46
47 357 initially generalisable models may need re-validation over time.^{23 58}

48
49 358 Discrimination (i.e., the ability to distinguish a patient with a high(er) risk of developing cancer from
50
51 359 one with a low(er) risk of developing cancer) was measured in every study using the AUC, as is
52
53 360 common in the field. However, discrimination is not the only metric of model performance.⁵⁹ Another
54
55 361 important measure of model performance, particularly for the clinical setting, is calibration; that is,
56
57 362 establishing that the risk estimates are accurate.⁶⁰ In this setting, this means that the model should
58
59 363 not unduly over- or underestimate the risk that a patient will develop cancer; to do so would mean that
60

1
2
3 364 a patient might be subjected to investigations and the associated worry of their likelihood of
4
5 365 developing cancer (overestimated risk), or, conversely, under-investigated and falsely reassured in
6
7 366 the case of underestimated risk. Therefore, a highly discriminatory but poorly calibrated model is likely
8
9 367 to have poor clinical utility.

10
11 368 None of the studies reviewed here performed calibration analysis, which is not uncommon in this field.
12
13 369 Indeed, in their systematic review of 71 studies using ML for clinical prediction for a wide variety of
14
15 370 clinical purposes, Christodoulou et al. reported that 79% of studies failed to address the calibration
16
17 371 problem.³⁷ Therefore, caution must be applied when interpreting and comparing the performance of
18
19 372 current ML models based on AUC alone, since is an incomplete measure of performance that must
20
21 373 be considered together with methodological aspects such overfitting, measurement error, and
22
23 374 population heterogeneity that might influence the estimation of predictive performance.^{37 60}

24
25 375

26 27 376 **Implications for clinical practice**

28
29
30 377 The ML models described in this scoping review generally show high AUC values. So, are any of
31
32 378 these models ready for clinical use? The ColonFlag model^{39 43} is an example has recently been
33
34 379 implemented at Barts Health NHS Trust⁶¹ to identify patients at particularly high risk of CRC,
35
36 380 particularly as clinicians struggle to prioritise patients in the backlog created by the COVID-19
37
38 381 pandemic. The ColonFlag model is the only model identified in this scoping review that has
39
40 382 undergone extensive external validation in independent datasets.

41
42 383 New ML models need to be contextualised with currently available best clinical practice in order to
43
44 384 fully evaluate their potential clinical value. Comparing the relatively poor AUC values of the Stark et
45
46 385 al.⁴⁴ models with BCRAT revealed that they in fact outperformed it in many cases. In their comparison
47
48 386 of their ANN with screening methods for lung cancer such as low-dose CT, chest X-ray, and sputum
49
50 387 cytology, Hart et al.³⁸ noted that (according to sensitivity and specificity) it outperformed most of the
51
52 388 other available non-invasive methods. Thorough side-by-side comparisons of newly developed
53
54 389 models with other prediction tools would be helpful in establishing future clinical utility.

55
56 390 Finally, this scoping review highlights that model performance should not be evaluated solely on the
57
58 391 basis of AUC values but also in terms of other importance performance metrics such as calibration,

1
2
3 392 without which a model might inaccurately assess risk and therefore prompt inappropriate
4
5 393 management.

6
7 394

8
9
10 395 **Unanswered questions and future research**

11
12 396 The few models that are currently available are methodologically diverse, rarely validated in
13
14 397 independent datasets to ensure generalisability, and do not cover all cancer types. Even if ML
15
16 398 techniques offer only small improvements in cancer detection rates, these improvements are likely to
17
18 399 be of high clinical significance given the large size of the global population with or at high risk of
19
20 400 cancer and the high mortality and costs associated with late cancer diagnoses.

21
22 401 However, the scoping review identifies a number of research gaps that will need to be addressed in
23
24 402 order to deliver validated ML-based models to assist clinical decision-making. Firstly, future studies
25
26 403 must take steps to establish model generalisability through validation in independent cohorts,
27
28 404 including those from LMICs. Although the latter may be challenging, it could be argued that even
29
30 405 negative generalisability studies might provide an opportunity to learn more about cancer risk factors
31
32 406 in different populations. Secondly, the scoping review fails to establish which ML approach best suits
33
34 407 the cancer prediction problem but does show that, where possible, side-by-side comparisons of
35
36 408 different methods can reveal important information about generalisability as well as performance and
37
38 409 that these comparisons are desirable whenever possible. Thirdly, many important cancer types,
39
40 410 particularly 'silent killers' like ovarian cancer, have currently not been the subject of ML modelling
41
42 411 approaches; ML could provide an important, low-cost, non-invasive method to identify individuals at
43
44 412 high risk of clinically silent cancers that require closer monitoring. Fourthly, progress has been made
45
46 413 in defining approaches to tailor sample sizes to the specific setting of interest to minimise overfitting
47
48 414 and targeting precise estimates of key parameters, and these principles must be applied when testing
49
50 415 and validating models to ensure robust model performance. Finally, ML models need to be compared
51
52 416 to best available clinical tools so that their potential clinical utility is transparent.

53 417

54
55
56 418 **Limitations of this study**

1
2
3 419 Our study has a number of limitations. First, despite recognising the need for a scoping review due to
4
5 420 the paucity of literature on the topic, we were only able to identify ten papers meeting the inclusion
6
7 421 criteria. It is therefore difficult to draw definitive conclusions about the performance of these models.
8
9 422 Furthermore, although AUC values provide an indication of how discriminative the models are, they
10
11 423 do not allow for meaningful comparisons of models trained and evaluated on different datasets. Six
12
13 424 out of ten studies were defined as poor quality due to a lack of controlling for confounders in the study
14
15 425 design, which may have introduced significant bias. Finally, we only search the PubMed database
16
17 426 and articles published in English, so some papers in other languages or in databases for non-medical
18
19 427 disciplines may have been missed.
20
21 428

428

429 **CONCLUSIONS**

23
24
25
26 430 This scoping review highlights that applying ML to cancer prediction is a promising field provided that
27
28 431 the identified issues such as generalisability, validation and clinical applicability, model calibration,
29
30 432 and dataset selection are addressed in future studies. We hope that the identified research gaps
31
32 433 focus future research efforts to deliver validated ML-based models to assist and improve clinical
33
34 434 decision-making
35
36 435
37
38 436
39
40 437
41
42 438

435

436

437

438

439 **Contributorship statement**

44
45 440 AAA defined the research question of the scoping review, conducted the literature search, and
46
47 441 summarised the findings. HK and AGL supervised the research. All authors drafted and revised the
48
49 442 manuscript.
50
51 443

443

53 444 **Competing interests**

54
55 445 No competing interests.
56
57 446

446

58 447 **Funding**

447

1
2
3 448 AGL is supported by funding from the Wellcome Trust (204841/Z/16/Z), National Institute for Health
4
5 449 Research (NIHR) University College London Hospitals Biomedical Research Centre
6
7 450 (BRC714/HI/RW/101440), NIHR Great Ormond Street Hospital Biomedical Research Centre
8
9 451 (19RX02).

10
11 452 **Data sharing statement**

12
13 453 Only public published papers were used.

14
15 454 **Ethics Approval**

16
17 455 Not Applicable. This study does not involve human participants nor animal subjects
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 456 **REFERENCES**

- 457 1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;69(1):7-34. doi:
458 10.3322/caac.21551
- 459 2. Cancer Research UK. Cancer Statistics for the UK 2020 [Available from:
460 <https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk>].
- 461 3. Cancer Research UK. Why are cancer rates increasing? 2014 [Available from:
462 <https://scienceblog.cancerresearchuk.org/2015/02/04/why-are-cancer-rates-increasing/>].
- 463 4. World Health Organization. Early detection of cancer 2016 [Available from:
464 <https://www.who.int/cancer/detection/en/>].
- 465 5. World Health Organization. Early cancer diagnosis saves lives, cuts treatment costs 2017
466 [Available from: [https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-](https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-saves-lives-cuts-treatment-costs)
467 [saves-lives-cuts-treatment-costs](https://www.who.int/news-room/detail/03-02-2017-early-cancer-diagnosis-saves-lives-cuts-treatment-costs)].
- 468 6. Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor
469 DNA. *Sci Transl Med* 2017;9(403) doi: 10.1126/scitranslmed.aan2415
- 470 7. Blumen H, Fitch K, Polkus V. Comparison of Treatment Costs for Breast Cancer, by Tumor Stage
471 and Type of Service. *Am Health Drug Benefits* 2016;9(1):23-32.
- 472 8. Kakushadze Z, Raghubanshi R, Yu W. Estimating cost savings from early cancer diagnosis. *Data*
473 2017;2(3):30.
- 474 9. Cancer Research UK. Saving lives, averting costs 2014 [Available from:
475 https://www.cancerresearchuk.org/sites/default/files/saving_lives_averting_costs.pdf].
- 476 10. Bannister N, Broggio J. Cancer survival by stage at diagnosis for England (experimental
477 statistics): adults diagnosed 2012, 2013 and 2014 and followed up to 2015. *Produced in*
478 *collaboration with Public Health England* 2016
- 479 11. Canary Foundation. Early Detection Facts and Figures Early Detection Works. California 2019.
- 480 12. Cancer Research UK. Why is early diagnosis important? 2018 [Available from:
481 [https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-](https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important)
482 [important](https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important)].
- 483 13. Weller DP, Campbell C. Uptake in cancer screening programmes: a priority in cancer control. *Br J*
484 *Cancer* 2009;101 Suppl 2:S55-9. doi: 10.1038/sj.bjc.6605391

- 1
2
3 485 14. Cancer Research UK. About cancer screening 2020 [Available from:
4
5 486 <https://www.cancerresearchuk.org/about-cancer/screening>].
6
7 487 15. Smith RA, Andrews KS, Brooks D, et al. Cancer screening in the United States, 2019: A review of
8
9 488 current American Cancer Society guidelines and current issues in cancer screening. *CA*
10
11 489 *Cancer J Clin* 2019;69(3):184-210. doi: 10.3322/caac.21557
12
13 490 16. Joseph DA, King JB, Dowling NF, et al. Vital signs: colorectal cancer screening test use—United
14
15 491 States, 2018. *Morbidity and Mortality Weekly Report* 2020;69(10):253.
16
17 492 17. Power E, Wardle J. Change in public awareness of symptoms and perceived barriers to seeing a
18
19 493 doctor following Be Clear on Cancer campaigns in England. *British Journal of Cancer*
20
21 494 2015;112(1):S22-S26.
22
23 495 18. Smith LK, Pope C, Botha JL. Patients' help-seeking experiences and delay in cancer
24
25 496 presentation: a qualitative synthesis. *Lancet* 2005;366(9488):825-31. doi: 10.1016/S0140-
26
27 497 6736(05)67030-4
28
29 498 19. Balasooriya-Smeekens C, Walter FM, Scott S. The role of emotions in time to presentation for
30
31 499 symptoms suggestive of cancer: a systematic literature review of quantitative studies.
32
33 500 *Psychooncology* 2015;24(12):1594-604. doi: 10.1002/pon.3833
34
35 501 20. Taber JM, Leyva B, Persoskie A. Why do people avoid medical care? A qualitative study using
36
37 502 national data. *J Gen Intern Med* 2015;30(3):290-7. doi: 10.1007/s11606-014-3089-1
38
39 503 21. American Cancer Society. Lifetime risk of developing or dying from cancer. 2014
40
41 504 22. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk
42
43 505 prediction models with electronic health records data: a systematic review. *J Am Med Inform*
44
45 506 *Assoc* 2017;24(1):198-208. doi: 10.1093/jamia/ocw042
46
47 507 23. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw Open*
48
49 508 2018;1(4):e181404. doi: 10.1001/jamanetworkopen.2018.1404
50
51 509 24. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380(14):1347-
52
53 510 58. doi: 10.1056/NEJMra1814259
54
55 511 25. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective
56
57 512 Study Using Statewide Electronic Health Records and Machine Learning. *J Med Internet Res*
58
59 513 2018;20(1):e22. doi: 10.2196/jmir.9268
60

- 1
2
3 514 26. Hao S, Fu T, Wu Q, et al. Estimating One-Year Risk of Incident Chronic Kidney Disease:
4
5 515 Retrospective Development and Validation Study Using Electronic Medical Record Data From
6
7 516 the State of Maine. *JMIR Med Inform* 2017;5(3):e21. doi: 10.2196/medinform.7954
8
9 517 27. Martin Noguero T, Paulano-Godino F, Martin-Valdivia MT, et al. Strengths, Weaknesses,
10
11 518 Opportunities, and Threats Analysis of Artificial Intelligence and Machine Learning
12
13 519 Applications in Radiology. *J Am Coll Radiol* 2019;16(9 Pt B):1239-47. doi:
14
15 520 10.1016/j.jacr.2019.05.047
16
17 521 28. Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology - new tools for
18
19 522 diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019;16(11):703-15. doi:
20
21 523 10.1038/s41571-019-0252-y
22
23 524 29. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference,
24
25 525 and prediction: Springer Science & Business Media 2009.
26
27 526 30. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis
28
29 527 and prediction. *Comput Struct Biotechnol J* 2015;13:8-17. doi: 10.1016/j.csbj.2014.11.005
30
31 528 31. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling
32
33 529 cancer risk using structured clinical data. *Artif Intell Med* 2018;90:1-14. doi:
34
35 530 10.1016/j.artmed.2018.06.002
36
37 531 32. Munn Z, Peters MDJ, Stern C, et al. Systematic review or scoping review? Guidance for authors
38
39 532 when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*
40
41 533 2018;18(1):143. doi: 10.1186/s12874-018-0611-x
42
43 534 33. Peters MD. In no uncertain terms: the importance of a defined objective in scoping reviews. *JBIR*
44
45 535 *Database System Rev Implement Rep* 2016;14(2):1-4. doi: 10.11124/jbisrir-2016-2838
46
47 536 34. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018;319(13):1317-
48
49 537 18. doi: 10.1001/jama.2017.18391
50
51 538 35. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of
52
53 539 nonrandomized studies in meta-analyses. *Eur J Epidemiol* 2010;25(9):603-5. doi:
54
55 540 10.1007/s10654-010-9491-z
56
57 541 36. Waegeman W, De Baets B, Boullart L. ROC analysis in ordinal regression learning. *Pattern*
58
59 542 *Recognition Letters* 2008;29(1):1-9.
60

- 1
2
3 543 37. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of
4
5 544 machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*
6
7 545 2019;110:12-22. doi: 10.1016/j.jclinepi.2019.02.004
8
9 546 38. Hart GR, Roffman DA, Decker R, et al. A multi-parameterized artificial neural network for lung
10
11 547 cancer risk prediction. *PLoS One* 2018;13(10):e0205264. doi: 10.1371/journal.pone.0205264
12
13 548 39. Hornbrook MC, Goshen R, Choman E, et al. Early Colorectal Cancer Detected by Machine
14
15 549 Learning Model Using Gender, Age, and Complete Blood Count Data. *Dig Dis Sci*
16
17 550 2017;62(10):2719-27. doi: 10.1007/s10620-017-4722-8
18
19 551 40. Miotto R, Li L, Kidd BA, et al. Deep Patient: An Unsupervised Representation to Predict the Future
20
21 552 of Patients from the Electronic Health Records. *Sci Rep* 2016;6:26094. doi:
22
23 553 10.1038/srep26094
24
25 554 41. Richter AN, Khoshgoftaar TM. Efficient learning from big data for cancer risk modeling: A case
26
27 555 study with melanoma. *Comput Biol Med* 2019;110:29-39. doi:
28
29 556 10.1016/j.combiomed.2019.04.039
30
31 557 42. Roffman D, Hart G, Girardi M, et al. Predicting non-melanoma skin cancer via a multi-
32
33 558 parameterized artificial neural network. *Sci Rep* 2018;8(1):1701. doi: 10.1038/s41598-018-
34
35 559 19907-9
36
37 560 43. Schneider J, Layefsky E, Udaltsova N, et al. Validation of an Algorithm to Identify Patients at Risk
38
39 561 for Colorectal Cancer Based on Laboratory Test and Demographic Data in Diverse,
40
41 562 Community-Based Population. *Clin Gastroenterol Hepatol* 2020 doi:
42
43 563 10.1016/j.cgh.2020.04.054
44
45 564 44. Stark GF, Hart GR, Nartowt BJ, et al. Predicting breast cancer risk using personal health data and
46
47 565 machine learning models. *PLoS One* 2019;14(12):e0226765. doi:
48
49 566 10.1371/journal.pone.0226765
50
51 567 45. Wang HH, Wang YH, Liang CW, et al. Assessment of Deep Learning Using Nonimaging
52
53 568 Information and Sequential Medical Records to Develop a Prediction Model for
54
55 569 Nonmelanoma Skin Cancer. *JAMA Dermatol* 2019 doi: 10.1001/jamadermatol.2019.2335
56
57 570 46. Wang YH, Nguyen PA, Islam MM, et al. Development of Deep Learning Algorithm for Detection of
58
59 571 Colorectal Cancer in EHR Data. *Stud Health Technol Inform* 2019;264:438-41. doi:
60
572 10.3233/SHTI190259

- 1
2
3 573 47. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian
4
5 574 network for pancreatic cancer prediction. *J Biomed Inform* 2011;44(5):859-68. doi:
6
7 575 10.1016/j.jbi.2011.05.004
8
9 576 48. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for
10
11 577 detection of colorectal cancer in primary care by analysis of complete blood counts: a
12
13 578 binational retrospective study. *J Am Med Inform Assoc* 2016;23(5):879-90. doi:
14
15 579 10.1093/jamia/ocv195
16
17 580 49. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast
18
19 581 cancer for white females who are being examined annually. *J Natl Cancer Inst*
20
21 582 1989;81(24):1879-86. doi: 10.1093/jnci/81.24.1879
22
23 583 50. Lorena AC, Jacintho LF, Siqueira MF, et al. Comparing machine learning classifiers in potential
24
25 584 distribution modelling. *Expert Systems with Applications* 2011;38(5):5268-75.
26
27 585 51. Usher-Smith JA, Walter FM, Emery JD, et al. Risk Prediction Models for Colorectal Cancer: A
28
29 586 Systematic Review. *Cancer Prev Res (Phila)* 2016;9(1):13-26. doi: 10.1158/1940-
30
31 587 6207.CAPR-15-0274
32
33 588 52. Gray EP, Teare MD, Stevens J, et al. Risk Prediction Models for Lung Cancer: A Systematic
34
35 589 Review. *Clin Lung Cancer* 2016;17(2):95-106. doi: 10.1016/j.clcc.2015.11.007
36
37 590 53. Usher-Smith JA, Emery J, Kassianos AP, et al. Risk prediction models for melanoma: a
38
39 591 systematic review. *Cancer Epidemiol Biomarkers Prev* 2014;23(8):1450-63. doi:
40
41 592 10.1158/1055-9965.EPI-14-0295
42
43 593 54. Rastogi T, Hildesheim A, Sinha R. Opportunities for cancer epidemiology in developing countries.
44
45 594 *Nat Rev Cancer* 2004;4(11):909-17. doi: 10.1038/nrc1475
46
47 595 55. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a
48
49 596 simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
50
51 597 doi: 10.1186/1471-2288-14-137
52
53 598 56. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical
54
55 599 prediction model. *BMJ* 2020;368:m441. doi: 10.1136/bmj.m441
56
57 600 57. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995.
58
59 601 Montreal, Canada.
60

- 1
2
3 602 58. Bergquist SL, Brooks GA, Keating NL, et al. Classifying Lung Cancer Severity with Ensemble
4
5 603 Machine Learning in Health Care Claims Data. *Proc Mach Learn Res* 2017;68:25-38.
6
7 604 59. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a
8
9 605 framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38. doi:
10
11 606 10.1097/EDE.0b013e3181c30fb2
12
13 607 60. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive
14
15 608 analytics. *BMC Med* 2019;17(1):230. doi: 10.1186/s12916-019-1466-7
16
17 609 61. Downing M. Barts Health using AI to prioritise care for colon cancer patients 2020 [Available from:
18
19 610 [https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-](https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-colon-cancer-patients-8867)
20
21 611 [colon-cancer-patients-8867](https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-high-risk-colon-cancer-patients-8867) accessed 2/11/2020].
22
23 612

24
25 613 **FIGURE LEGEND**

26
27 614 **Figure 1.** PRISMA flowchart depicting the search strategy.
28
29 615

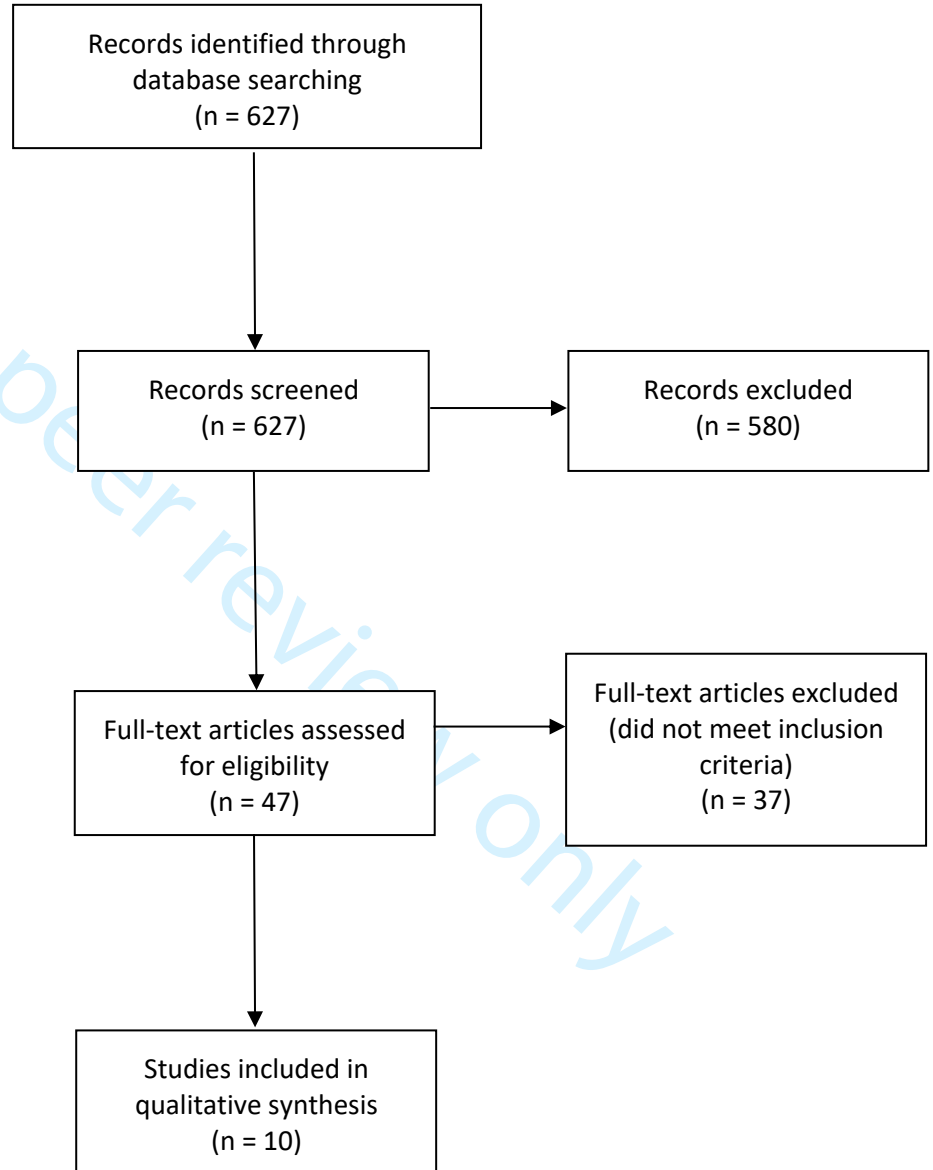
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

Identification

Screening

Eligibility

Included



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

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2 and 3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	7
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	7
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	NA
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	9
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	9
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	9
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	9
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	13, 14, 15, and 16
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	13, 14, 15, and 16
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	NA

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	9 and 10
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	11
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	13, 14, 15, and 16
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	NA
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	13, 14, 15, and 16
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	11, 12, 17, and 18
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	18, 19, 20, 21, and 22
Limitations	20	Discuss the limitations of the scoping review process.	22 and 23
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	23
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	23 and 24

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* 2018;169:467–473. doi: 10.7326/M18-0850.



St. Michael's

Inspired Care. For peer review only - <http://bmjopen.bmj.com/site/about/guidelines.xhtml>
Inspiring Science.