BMJ Open

Predicting population health with machine learning: a scoping review

Jason Denzil Morgenstern ,[1] Emmalin Buajitti ,[2,3] Meghan O'Neill ,[2] Thomas Piggott,[1] Vivek Goel,[2,3] Daniel Fridman,[4] Kathy Kornas,[2] Laura C Rosella [2,3,5]

To cite: Morgenstern JD, Buajitti E, O'Neill M, *et al*. Predicting population health with machine learning: a scoping review. *BMJ Open* 2020;**10**:e037860. doi:10.1136/bmjopen-2020-037860

► Prepublication history and additional materials for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-037860).

Received 19 February 2020
Revised 27 July 2020
Accepted 01 October 2020

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[3]Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
[4]Hospital for Sick Children, Toronto, Ontario, Canada
[5]Vector Institute, Toronto, Ontario, Canada

Correspondence to
Dr Laura C Rosella;
laura.rosella@utoronto.ca

ABSTRACT

Objective To determine how machine learning has been applied to prediction applications in population health contexts. Specifically, to describe which outcomes have been studied, the data sources most widely used and whether reporting of machine learning predictive models aligns with established reporting guidelines.

Design A scoping review.

Data sources MEDLINE, EMBASE, CINAHL, ProQuest, Scopus, Web of Science, Cochrane Library, INSPEC and ACM Digital Library were searched on 18 July 2018.

Eligibility criteria We included English articles published between 1980 and 2018 that used machine learning to predict population-health-related outcomes. We excluded studies that only used logistic regression or were restricted to a clinical context.

Data extraction and synthesis We summarised findings extracted from published reports, which included general study characteristics, aspects of model development, reporting of results and model discussion items.

Results Of 22 618 articles found by our search, 231 were included in the review. The USA (n=71, 30.74%) and China (n=40, 17.32%) produced the most studies. Cardiovascular disease (n=22, 9.52%) was the most studied outcome. The median number of observations was 5414 (IQR=16 543.5) and the median number of features was 17 (IQR=31). Health records (n=126, 54.5%) and investigator-generated data (n=86, 37.2%) were the most common data sources. Many studies did not incorporate recommended guidelines on machine learning and predictive modelling. Predictive discrimination was commonly assessed using area under the receiver operator curve (n=98, 42.42%) and calibration was rarely assessed (n=22, 9.52%).

Conclusions Machine learning applications in population health have concentrated on regions and diseases well represented in traditional data sources, infrequently using big data. Important aspects of model development were under-reported. Greater use of big data and reporting guidelines for predictive modelling could improve machine learning applications in population health.

Registration number Registered on the Open Science Framework on 17 July 2018 (available at https://osf.io/rnqe6/).

INTRODUCTION

Predictive models have a long history in clinical medicine. One well-known example is the Framingham Risk Score, which was first developed in 1967.[1] Such models have proliferated throughout clinical practice to inform management and interventions, including preventive approaches. More recently, researchers have developed prediction models beyond individual clinical applications, for population health uses.[2 3] While there is no universal definition of population health, it generally encompasses 'the health outcomes of a group of individuals, including the distribution of such outcomes within the group'.[4] Similarly to clinical medicine, population-level models can be used to identify high-risk groups, directing the implementation of preventive interventions. Additionally, population health prediction models can inform policy-makers about future disease burden and help to assess the impact of public health actions. Thus far, most predictive modelling in both medicine and population health has used parametric statistical regression models. More recently, there has been increasing interest in the use of a broader range of machine learning methods for prediction tasks.[5–7]

Strengths and limitations of this study

- ► Our review is one of the first syntheses of machine learning applications in population and public health.
- ► We used a robust search strategy, including nine peer-reviewed databases, grey literature and reference searching, to comprehensively describe the literature.
- ► We compared reported study characteristics to established predictive modelling reporting guidelines, which provide an objective measure of the quality of reporting.
- ► Since both machine learning and population health have broad definitions, there may be some relevant articles that were not included.
- ► Given our focus on prediction, we could not address many other important intersections of machine learning and population health, such as surveillance and health promotion.

Machine learning can be loosely defined as the study and development of algorithms that learn from data with little or no human assistance.[8] These approaches have been increasingly applied in the past two decades as a result of the enabling growth of big data reserves and computational power.[9] Recent machine learning applications to prediction in population health contexts include forecasting childhood lead poisoning,[10] yellow fever incidence[11] and the onset of suicidal ideation.[12]

The distinction between machine learning algorithms and parametric regression models is debated.[13] Regression models tend to impose more structure on the data, requiring greater human input for the verification of distributional assumptions and incorporation of domain knowledge in choosing the input parameters.[14] Algorithms employed in machine learning often derive more structure directly from the data, making fewer distributional assumptions about the data or variables. The literature remains divided on the relative advantages of more traditional approaches compared with newer methods[15]; however, given the wide variation in applications and the data used in these examples, broad assessments of superiority are often not appropriate. Also, there are debates regarding the differences in developing and validating machine learning approaches for health applications.[15 16]

Population health applications of prediction models are relatively new compared with clinical applications; correspondingly, the role of machine learning in these applications has been far less studied and discussed in the health literature. The goals of our review are to determine how machine learning has been applied to prediction in population health, the nature of the models and data used, and how the models have been developed. We also sought to assess how well the published literature aligns with recommended guidelines for reporting of predictive models and machine learning, by extracting features related to model development and performance that are highlighted by two such guidelines.[16 17]

METHODS

We based our scoping review on the framework proposed by Arksey and O'Malley[18] and refined by the Joanna Briggs Institute.[19] We also followed the more recent Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews.[20] Our study protocol was registered on the Open Science Framework on 17 July 2018 (available at https://osf.io/rnqe6/).

Our initial goal was to scope out all machine learning applications in population health. However, the screening process identified a much larger number of publications than anticipated. Consequently, to describe the subject area comprehensively, we restricted our scope to articles predicting future outcomes.

Search strategy

Our search strategy consisted of peer-reviewed literature databases, grey literature and reference searches. First, we searched nine interdisciplinary, indexed databases (MEDLINE, EMBASE, CINAHL, ProQuest, Scopus, Web of Science, Cochrane Library, INSPEC and ACM Digital Library) on 18 July 2018 for papers published between 1980 and 2018. Our search was informed by consultation with a health science librarian, a machine learning textbook[21] and a similar registered review.[15] Online supplemental table A includes the full MEDLINE search strategy and filters, serving as an example search query for all database searches.

Our grey literature search included Google Scholar and Google. We developed a Google Scholar search based on terms related to 'machine learning' and 'population health', which was refined based on the relevance of initial results. The first 200 results were included in screening. A similar approach was used for the general Google search, which we restricted to the first 30 results. We examined relevant websites for publications. Results were limited to articles published on or before the date of the peer-reviewed literature search. Finally, we searched the references of relevant reviews for additional articles. Most of these reviews were identified during screening.

Eligibility criteria

We included articles if they used machine learning to develop a predictive model that could be applied in a population health context. Therefore, we excluded articles where the model was trained primarily on people with a pre-existing disease. We also excluded articles that were only indirectly related to population health, for example, traffic accident models that did not predict a health outcome. Studies predicting individual outcomes were included if the approach was determined to be scalable to a population level. Finally, articles using only logistic regression were excluded. See online supplemental appendix A for the full eligibility criteria.

In order to manage the scope, articles were excluded if their full text could not be retrieved with our institutional licenses and if they were not written in English. Finally, articles published prior to 1980 were excluded as earlier machine learning investigators lacked comparable amounts of digitised data, software and computational resources.

Screening process

Initially, individual reviewers screened titles for obvious irrelevance to the review topic (JDM and EB). One example of an obviously irrelevant topic was a paper describing the machine health lifespan of a piece of industrial equipment; specific examples of articles removed at this stage are listed in online supplemental appendix B. Then, we imported remaining references into Covidence systematic review management software.[22] Two reviewers screened the abstracts of remaining articles (JDM, EB, MO'N and DF). Prior to evaluating full texts using all eligibility criteria, we then screened out articles that did not focus on a prediction application (JDM, EB and MO'N). Finally, two reviewers screened the full text of remaining articles (JDM, EB and MO'N). Conflicts were resolved by discussion between at least two reviewers.

Data extraction and synthesis

Individual authors extracted article data (JDM, EB, MO and DF). We based our extraction items on features identified in a recent biomedical guideline for reporting of machine learning predictive models[16] and on the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.[17] Major extraction categories identified from these guidelines included general study characteristics (eg, geographic location and sample size), model development (eg, algorithms used and type of validation), results (eg, discrimination and calibration measures) and model discussion (eg, practical costs of errors and implementation). See online supplemental table B for a description of each extraction item.

We computed descriptive statistics for all extraction items. For categorical extracted features (eg, whether or not unstructured text was used and the method of validation used), we calculated the total number and percent of all studies in a particular category. For continuous extracted features (eg, number of observations in the study sample), we calculated the median value and the IQR (range between quartile 1 and quartile 3 in the value distribution). We also completed a narrative synthesis of discussion elements based on the text of included manuscripts.

Patient and public involvement statement

There was no patient or public involvement in this study.

RESULTS

We initially retrieved 16 162 articles, after removing duplicates (figure 1). We excluded 6494 articles after title screening, 7860 after abstract screening, 1456 when screening out non-prediction articles and 121 after full-text screening. This resulted in 231 articles being included in the final review (see online supplemental appendix C).

General study characteristics

The number of articles published in the population health prediction area that used machine learning increased dramatically after 2007 (see online supplemental figure A). Studies were undertaken worldwide, with the largest representation from the USA (n=71, 30.74%) and China (n=40, 17.32%) (table 1). Relatively few articles came from Oceania (n=2, 0.87%), Africa (n=5, 2.16%) and the Americas outside of the USA (n=13, 5.63%).

The median number of observations in each article was 5414 (IQR=16 543.5) and the median number of features (ie, independent variables) used was 17 (IQR=31) (table 1). Seventy-two studies (31.2%) did not report the number of observations. These studies often used data from reportable disease databases, which do not necessarily have a firm sampling frame, making ascertainment of the number of observations difficult.

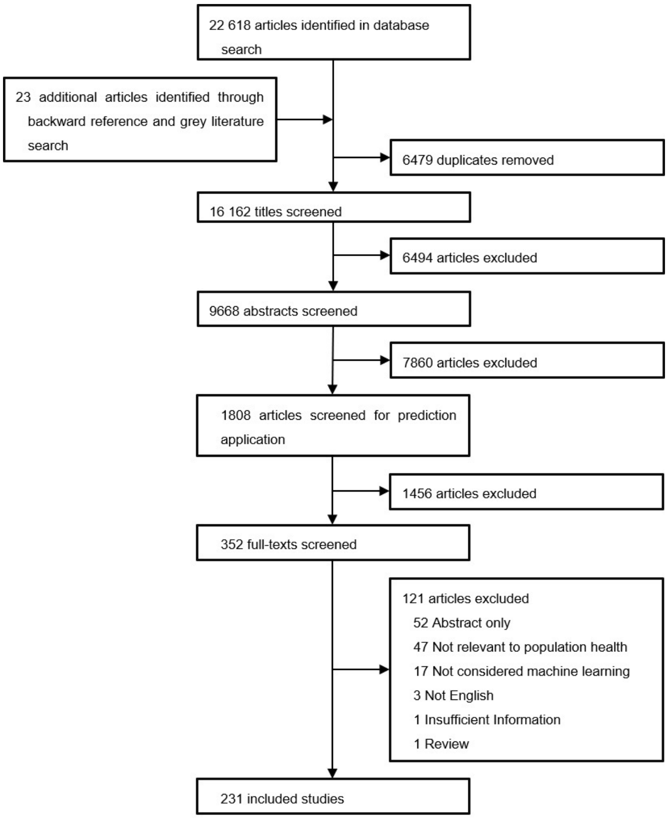
22 618 articles identified in database search
23 additional articles identified through backward reference and grey literature search
6479 duplicates removed
16 162 titles screened
6494 articles excluded
9668 abstracts screened
7860 articles excluded
1808 articles screened for prediction application
1456 articles excluded
352 full-texts screened
121 articles excluded
52 Abstract only
47 Not relevant to population health
17 Not considered machine learning
3 Not English
1 Insufficient Information
1 Review
231 included studies

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart of article screening process.

Algorithms

The most frequently used machine learning algorithms were neural networks (n=95, 41.13%), followed by support vector machines (n=59, 25.54%), single tree-based methods (n=52, 22.51%) and random forests (n=48, 20.78%) (see online supplemental table C). About half of the articles made a comparison with statistical methods (n=111, 48.1%), which were generally logistic regression or autoregressive integrated moving average models (table 1).

Outcomes

Non-communicable disease outcomes were assessed by many articles (n=95, 41.13%), with communicable diseases (n=76, 32.90%) and non-disease outcomes (n=60, 25.97%) studied somewhat less often. The outcome most frequently predicted was cardiovascular disease (n=22, 9.52%) (figure 2). Other commonly forecasted non-communicable disease outcomes were suicidality (n=13, 5.63%), cancer (n=12, 5.19%) and perinatal health (n=12, 5.19%). Influenza (n=15, 6.49%) and dengue fever (n=14, 6.06%) were the most predicted communicable disease outcomes. Aside from non-communicable and communicable diseases, mortality (n=13, 5.63%) and healthcare utilisation (n=14, 6.06%) were also frequently predicted.

Table 1 Summary statistics of included articles

Characteristic*	Number of articles†	Percent of articles‡
Region		
The USA	71	30.74%
Asia excluding China	41	17.75%
China	40	17.32%
Europe	36	15.58%
Americas excluding the USA	13	5.63%
Africa	5	2.16%
Oceania	2	0.87%
Multi-region	15	6.49%
Not reported	8	3.46%
Year published		
Before 1990	1	0.4%
1990–1999	3	1.3%
2000–2004	13	5.6%
2005–2009	18	7.8%
2010–2014	70	30.3%
2015–2018	126	54.5%
Outcome level§		
Individual risk prediction	139	60.17%
Population risk prediction	92	39.83%
Number of observations	Median=5414†	IQR=16 543.5‡
Not reported	72	31.2%
Number of features	Median=17†	IQR=31‡
Not reported	59	25.5%
Used any unstructured text		
Yes	24	10.4%
No	207	89.6%
Machine learning model was compared with other statistical methods	111	48.1%
Reported data preprocessing¶		
Yes	160	69.3%
No	71	30.7%
Reported method of feature selection		
Yes	164	71.0%
No	67	29.0%
Reported hyperparameter search		
Yes	114	49.4%
No	117	50.6%
Method of validation		
Holdout	112	48.5%
Cross-validation or bootstrap	84	36.4%
External	15	6.5%
Not reported	32	13.9%
Reported descriptive statistics**		
Yes	140	60.6%
No	91	39.4%
Discussed the practical costs of prediction errors††		
Yes	36	15.6%
No	195	84.4%
Stated rationale for using machine learning		
Yes	179	77.5%
No	52	22.5%
Discussed model usability		
Yes	91	39.4%
No	140	60.6%
Stated model limitations		
Yes	161	69.7%
No	70	30.3%
Discussed model implementation		
Yes	184	79.7%
No	47	20.3%
Dataset availability by study‡‡		
Closed	149	64.5%
Public	42	18.2%
Closed and public	38	16.5%
Unknown	1	0.4%

*Refer to online supplemental table A for a description of each characteristic and rationales for including some elements.
†In rows where the characteristic being measured is an integer count (eg, number of features), this column refers to the median value.
‡In rows where the characteristic being measured is an integer count (eg, number of features), this column refers to the IQR (quartile 3 – quartile 1).
§Individual risk prediction refers to studies that developed models to predict the health outcomes of individuals, while population risk prediction refers to studies that developed models to predict aggregated population-level health outcomes.
¶Whether any aspects of data cleaning or preprocessing were reported. Examples include how missing data were handled, whether log transformations were done and if derived variables were generated.
**Included a broad array of descriptive statistics such as sample population demographics, feature distributions and outcome distributions.
††Whether the article discussed the relative risks of false negative and false positive results based on their predictive model in contexts where it might be used.
‡‡Closed refers to datasets that were not immediately available in the public domain or were not identifiable as such.

Data

Data sources were usually structured (n=207, 89.6%) and closed, that is, not publicly available (n=189, 81.8%) (table 1). In general, high-dimensional data with many observations, such as multi-linked electronic medical records (EMRs) or internet-based data, may offer the most value for machine learning applications. These data types were represented in some of the articles captured, for which the most frequently reported data sources were

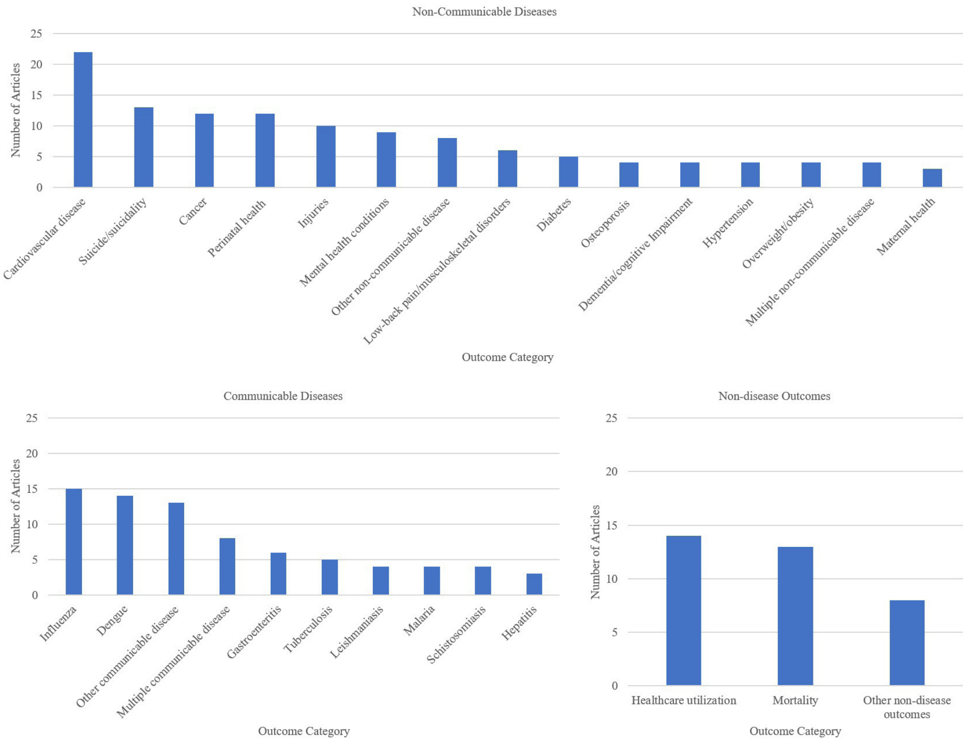

Figure 2 Number of articles by outcome.

health records (n=126, 54.5%) and investigator generated (eg, cohort studies) (n=86, 37.2%) (table 2). A large proportion of studies (n=42, 18.2%) used an environmental data source (eg, satellite imagery), mostly for prediction of infectious disease. Government databases (n=32, 13.9%) and internet-based data (n=21, 9.1%) were less frequently used. Among studies from China and the USA, 80.0% and 67.6%, respectively, used health records data, whereas 54.5% of studies overall used these data sources (see online supplemental figure B).

Features

The median number of features used in a machine learning algorithm was 17 (IQR=31; table 1). The frequency of specific feature categories used are shown in online supplemental figures C and table D. Biomedical and sociodemographic features were frequently used (see online supplemental figure C). Of these, the most commonly used were disease history (43.3%), age (48.5%) and sex/gender (41.1%). Among lifestyle features, smoking was the most frequently used (25.1%) and of environmental features, meteorology was common (17.3%). Social media posts (5.2%) and web search queries (5.2%) were not often used. In general, most studies focused on features typical of clinical prediction models, such as subject demographics, behaviours and medical histories. We observed limited use of other data, such as unstructured text or image-based features, which are difficult to parse using traditional statistical approaches and could benefit more from machine learning applications.

Model development and validation

The majority of articles reported how data preprocessing (n=160, 69.3%) and feature selection (n=164, 71%) were done (table 1). Fewer authors reported how hyperparameters were selected (n=114, 49.4%). Most studies used a holdout method of validation (n=112, 48.5%), 15 (6.5%) externally validated their models and 32 (13.9%) did not report how models were validated.

Performance metrics

Most articles reported a prediction discrimination metric (n=172, 74.46%), which quantifies a model's ability to correctly rank-order individuals (table 3).[23] Discrimination is a useful performance metric in cases where classification is the primary goal, including many machine learning relevant tasks such as image recognition. The most common discrimination metrics employed were area under the receiver operator curve (n=98, 42.42%), accuracy (n=76, 32.90%) and recall (n=68, 29.44%).

In clinical and public health settings, accurate prediction of outcome probabilities is important for the practical utility of a tool, so assessing model calibration is very important.

Few articles in our study reported a measure of calibration (n=21, 9.09%), which describes how well a model predicts the absolute probability of outcomes (table 3).[23] Calibration was mostly assessed with graphing methods (n=9, 3.90%) and Hosmer-Lemeshow statistics (n=8, 3.46%).

Some articles also reported a measure of overall model fit (n=77, 33.33%). Overall performance was usually measured with a form of mean error, such as root mean squared error (n=35, 15.15%).

Table 2 Data sources

Sources of data used*	Number	Percent
Environmental	42	18.2%
Geographical information database	12	5.2%
Meteorological/air quality datasets	32	13.9%
Satellite imagery	21	9.1%
Health records database	126	54.5%
Clinical record database†	46	19.9%
Disease registry	2	0.9%
Population health survey	15	6.5%
Reportable disease database	42	18.2%
Other health records database	30	13.0%
Government database	32	13.9%
Census	11	4.8%
Vital statistics	13	5.6%
Other government database	14	6.1%
HealthMap	3	1.3%
Private insurance data	9	3.9%
Private insurance claims	9	3.9%
Private insurance questionnaire	3	1.3%
Internet based	21	9.1%
Search engine	12	5.2%
Social media	12	5.2%
Investigator generated‡	86	37.2%
Public repositories§	19	8.2%
Health organisation reports¶	5	2.2%
Not reported	6	2.6%

*Categories are not mutually exclusive.
†Any dataset produced primarily for the purpose of delivering clinical care, such as electronic medical records and administrative healthcare databases produced by hospitals.
‡Any datasets resulting from researcher-driven studies, such as randomised controlled trials, cohort studies and case–control studies.
§Any freely available datasets such as Medical Information Mart for Intensive Care or the University of California, Irvine Machine Learning Repository.
¶Health-related reports, typically, including disease burden estimates, produced by non-governmental or governmental organisations, such as the WHO.

Table 3 Prediction performance metrics

Prediction performance metrics used	Number	Percent
Any overall performance metric	77	33.33%
RMSE	35	15.15%
MSE	26	11.26%
MAE	24	10.39%
MAPE	23	9.96%
R^2*	19	8.23%
Correlation	8	3.46%
AIC or BIC	8	3.46%
Other performance metric†	21	9.09%
Any discrimination metric	172	74.46%
Area under the curve‡	98	42.42%
Accuracy§	76	32.90%
Recall¶	68	29.44%
Precision**	39	16.88%
F statistics	10	4.33%
Likelihood ratio††	4	1.73%
Youden Index	3	1.30%
Manual or visual comparison	3	1.30%
Other discrimination metric‡‡	4	1.73%
Any calibration metric	21	9.09%
Manual or visual comparison§§	9	3.90%
Hosmer-Lemeshow	8	3.46%
Observed/xpected	5	2.16%
Other calibration metric¶¶	3	1.30%
Any reclassification metric	6	2.60%
Net Reclassification Index	5	2.16%
Integrated discrimination improvement	3	1.30%

*Includes R^2 and pseudo-R^2 metrics.
†Includes penalty error, total sum of squares, proportional reduction in error, overall prediction error, specific prediction error, Nash-Sutcliffe, root mean squared percentage Error (2), mean relative absolute error, analysis of variance F-stat, 2LogLikelihood, relative efficiency, deviance, Ljung-Box test, mean absolute deviation, SE, mean percentage error, Brier score and log score.
‡Includes c-statistic, s-index and area under the receiver operator curve.
§Includes accuracy, misclassification and error rate.
¶Includes sensitivity, specificity, true/false positive and true/false negative.
**Includes positive predictive value, negative predictive value and precision.
††Includes positive/negative likelihood ratios.
‡‡Includes G-means (2), k-statistic and Matthews correlation coefficient.
§§Includes calibration plots.
¶¶Includes mean bias (from Bland-Altman plot), calibration factoring and calibration statistic.
AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error; RMSE, rootmean squared error.

Study discussion and narrative synthesis

Most articles included some discussion of their rationale for using machine learning (n=179, 77.5%), although some articles did not mention or explain their rationale (n=52, 22.5%) (table 1). Rationale for applying machine learning approaches mainly focused on being 'state of the art' or better suited to modelling complex data than regression.

Most articles also had some discussion of the limitations of their study (n=161, 69.7%), and how the model might be implemented (n=184, 79.7%) (table 1). Frequent concerns were an inadequate sample size, too few features, questionable generalisability and a lack of interpretability. When discussing model implementation, many articles stated that predictive accuracy would be improved, but they did not frequently discuss how this could be translated to specific health-related policies or actions.

Less than half of the articles discussed model usability (n=91, 39.4%), that is, whether and how the model could practically be used in a relevant context. This is an important reporting component of the TRIPOD statement (Discuss the potential clinical use of the model and implications for future research) and is relevant for understanding real-word applications of prediction models.[17] Also, only a small number discussed the costs of prediction errors in real-world contexts (n=36, 15.6%).

See online supplemental appendix D for further narrative synthesis of discussion reporting items.

DISCUSSION

Our results show that machine learning is increasingly being applied to make predictions related to population health. However, applications of machine learning to population health prediction tasks have not capitalised fully on the opportunities presented by emerging big data resources and efficient machine learning algorithms. Furthermore, reporting of these models often does not align with established guidelines for reporting of prediction models, which limits their ability to be critically appraised, compared with existing statistical models, or implemented in clinical or public health practice.

Applications of machine learning prediction models

Nearly half of the included studies were conducted in the USA or China. Both countries produce the greatest number of scientific publications in general[24]; however, they also likely benefited from robust health data infrastructures. The USA has rapidly digitised much of its healthcare system, resulting in large EMRs linked with government data through public–private partnerships, including processes to make these data available to researchers.[25 26] Both the USA and China made greater use of health records and less use of investigator-generated data relative to other regions, which may have made machine learning projects more tractable. They also used more internet-based data, which typically includes many observations and is high dimensional, making it amenable to machine learning methods. We noted that studies from Oceania, Africa and the Americas (outside of the USA) were limited. This may be partly due to less availability of traditional sources of structured health data. However, given that machine learning methods can incorporate non-traditional data sources, there is the potential to expand use of these methods even when structured health data is unavailable.

We found that a wide range of population health outcomes have been the focus of machine learning prediction models. However, relative to morbidity and mortality, multiple outcome categories like cancer, HIV, dementia, gastroenteritis, pneumococcal disease, perinatal health, tuberculosis and malaria appear understudied.[27] Many of these conditions are most prevalent in regions with decreased access to traditional health data, perhaps stymieing research. If machine learning methods are used to leverage novel data sources for research in these regions, it could enable greater study of neglected diseases.

Most investigators did not analyse a large number of observations and features. We observed a high reliance on electronic health records and investigator-generated data, including the use of relatively small study cohorts. Small study sample sizes or narrow data collection associated with these data sources can make it difficult to achieve high sample sizes or high dimensional data, which may impact machine learning algorithm performance. Specifically, the use of smaller investigator-generated datasets may affect the performance of studied models, as machine learning algorithms generally require a high number of observations relative to features.[28] Additionally, most studies focused on features typical of clinical prediction models, such as biomedical factors and limited aspects of broader socioeconomic or environmental determinants of health. We also observed infrequent use of unstructured data and wearable data for prediction purposes. A reliance on small datasets and traditional numbers and types of features is unlikely to fully leverage any benefits of machine learning. This may be contributing to the small performance differences observed between parametric regression and machine learning models. Greater use of linked population-level databases, large EMRs, internet data and unstructured features would likely improve these approaches.

Reporting of machine learning prediction models

Based on the elements of model development that we studied, adherence to existing machine learning[16] and prediction model[17] guidelines appears limited. Most articles did not report their method of hyperparameter selection, discuss practical costs of prediction errors or consider model usability, which are needed for transparency and model assessment. Many studies did not report the number of features included, method of validation, method of feature selection or any performance metric. Given these issues, it would be difficult or impossible to compare many of these machine learning models with existing approaches. However, we acknowledge that existing guidelines were not available when many included studies were published. Future work should apply existing guidance,[16] including from TRIPOD,[17] and anticipate the forthcoming Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Machine Learning (TRIPOD-ML) statement.[29]

Lastly, we noted that included studies rarely assessed predictive performance in terms of calibration, which refers to a model's ability to accurately predict the absolute probability of outcomes.[23] In contrast, discrimination measures of predictive performance quantify a model's ability to correctly rank-order individuals. Many traditional machine learning tasks, such as image recognition, often have a high signal-to-noise ratio. In these cases, discrimination may be a suitable lone performance metric, as the algorithm can achieve near perfect performance. Conversely, health outcomes tend to be more stochastic. As a result, accurate prediction of probabilities is more important.[23] Models can have good predictive discrimination, but poor calibration, making them less useful in practice, particularly for population health applications. A further issue is that many measures of discrimination, such as accuracy and recall, artificially impose a threshold for calling events. Thresholds should ideally be ascertained by decision-makers based on their cost-utility curves.[23] Overall, applications of machine learning in population health would benefit from greater use of calibration performance metrics.

Strengths and limitations of this review

A strength of our study is that we addressed an understudied area, the intersection of machine learning and population health. Additionally, prediction is an application with untapped potential in population health, and where machine learning has the potential to make significant improvements. Our study also employed a comprehensive search strategy, including numerous multidisciplinary peer-reviewed databases, alongside a grey literature search. Furthermore, we applied insights from the field of clinical prediction modelling to population health and machine learning. Finally, given the focus on prediction, we were able to take a comprehensive approach to data extraction and synthesis.

In terms of limitations, concentrating on prediction prevented us from exploring applications of machine learning to other important aspects of population health, such as disease surveillance. These should be the focus of future research. Our review was also limited by including only English articles and articles with available full text, which may have introduced selection bias. Because of the broad scope of this review, and inconsistent reporting of model development and validation in reviewed articles, we were unable to carry out a critical appraisal of the literature and are unable to comment significantly on the overall performance of published machine learning population health prediction tools. This would be of great value for understanding the clinical and population health relevance of machine learning prediction tools. Lastly, the two main concepts underlying our review, machine learning and population health, are not universally defined. As a result, we may have excluded articles that may be relevant to these fields.

Research recommendations and conclusion

This was the first scoping review specifically focused on machine learning prediction in population health applications. Predictive modelling in population health can help to inform preventive interventions, anticipate future disease burden and assess the impact of health policies and programmes. Advances in machine learning offer opportunities to improve these models, particularly when incorporating big data. Countries with substantial EMR use and government database linkage such as Finland, Singapore and Denmark[30] likely have untapped potential for machine learning research. This is still a nascent field, but based on our findings, more research in Oceania, Africa and South America would also be particularly beneficial. Diseases with a high global burden of disease that were under-represented in our findings include malaria, tuberculosis and dementia, which may be opportune for further study.[31] Additionally, future machine learning projects could incorporate larger datasets and more non-traditional features. Greater use of resources such as HealthMap, social media, web search patterns, remote sensing and WHO reports would enable more work in regions without formal data sources and enrich research in others. Another largely untapped prospect is using machine learning and high-dimensional data to incorporate richer representations of the social determinants of health. Opportunities should continue to grow as governments increasingly digitise their health service records and link databases to both health and non-health data. Overall, as applications of machine learning in population health develop, adherence to existing guidance[16 17 29] will improve our ability to assess and advance machine learning applications. We hope that our results will help to inform future research in this area, including the development of guidelines for machine learning applications in population health. Finally, it will be important to evaluate the impact of prediction models on decisions made in population health and the practice of public health.

Twitter Laura C Rosella @LauraCRosella

Acknowledgements We are grateful to Catherine Bornbaum for her assistance with the initial design of the study.

Contributors JDM contributed to the design of the study and led the literature search, article screening, data extraction, analysis and writing of the manuscript. EB contributed to the design of the study, the literature search, article screening, data extraction and analysis. MO'N contributed to article screening, data extraction, analysis and writing of the manuscript. TP and VG contributed to the design of the study and supervised work. DF contributed to article screening and data extraction. KK contributed to the design of the study. LCR led the design of the study. All authors interpreted study results and contributed to drafting of the manuscript.

Funding This work was funded by the Canadian Institutes of Health Research (FRN: 72054363). LCR is funded by a Canada Research Chair in Population Health Analytics. The funders had no role in study design, data collection, data analysis, data interpretation or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data will be made available in a public, open access repository. The full data extraction table used for this review will be made publicly available after publication with no end date on Mendeley Data (DOI: 10.17632/7rrz9xrp2j.1).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Jason Denzil Morgenstern http://orcid.org/0000-0002-6636-462X
Emmalin Buajitti http://orcid.org/0000-0002-3194-7331
Meghan O'Neill http://orcid.org/0000-0001-9551-0967
Laura C Rosella http://orcid.org/0000-0003-4867-869X

REFERENCES

1 Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* 1967;20:511–24.
2 Nunes MB, McPherson M, Kommers P, *et al*. Proceedings of the International association for development of the information Society (IADIS) International Conference on e-learning (Lisbon, Portugal, July 20-22, 2017), 2017. Available: http://libaccess.mcmaster.ca/login?url=https://search.proquest.com/docview/2013525439?accountid=12347
3 Manuel DG, Tuna M, Bennett C, *et al*. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the cardiovascular disease population risk tool (CVDPoRT). *CMAJ* 2018;190:E871–82.
4 Kindig D, Stoddart G. What is population health? *Am J Public Health* 2003;93:380–3 http://www.ncbi.nlm.nih.gov/pubmed/12604476
5 Panch T, Pearson-Stuttard J, Greaves F, *et al*. Artificial intelligence: opportunities and risks for public health. *Lancet Digit Health* 2019;1:e13–14.
6 Aldridge RW. Research and training recommendations for public health data science. *Lancet Public Health* 2019;4:e373.
7 Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health* 2018;39:95–112.
8 Samuel AL. Some studies in machine learning using the game of Checkers. *IBM J Res Dev* 1959;3:210–29.
9 Russell SJ, Norvig P. *Artificial intelligence: a modern approach*. 2nd edn. Upper Saddle River, NJ: Prentice Hall, 2003.
10 Potash E, Brew J, Loewi A, *et al*. Predictive modeling for public health: preventing childhood lead poisoning. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min* 2015:2039–47.
11 Shearer FM, Longbottom J, Browne AJ, *et al*. Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. *Lancet Glob Health* 2018;6:e270–8.
12 De Choudhury M, Kiciman E, Dredze M. *Discovering shifts to suicidal ideation from mental health content in social media. In: Conference on Human Factors in Computing Systems - Proceedings. Association for Computing Machinery*, 2016: 2098–110.
13 Moons KG, de Groot JAH, Bouwmeester W, *et al*. Checklist for data extraction and critical appraisal for systematic reviews of prediction modelling studies: the charms checklist. *Eur J Epidemiol* 2015;30:904.
14 Breiman L. Statistical modeling: the two cultures. *Stat Sci* 2001;16:199–215.
15 Christodoulou E, Ma J, Collins GS, *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
16 Luo W, Phung D, Tran T, *et al*. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016;18:e323.
17 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015;13:1.
18 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8:19–32.
19 Joanna Briggs Institute. *Joanna Briggs institute reviewers' manual 2015 - methodology for JBI scoping reviews*. Adelaide, 2015.
20 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*.
21 Hastie T, Tibshirani R, Witten D, *et al*. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013.
22 Veritas Health Innovation. Covidence systematic review software. Available: www.covidence.org
23 Steyerberg EW. *Clinical prediction models*. New York: Springer, 2009.
24 Nature Index. Country outputs. Available: https://www.natureindex.com/country-outputs/generate/All/global/All/score [Accessed 15 Nov 2019].
25 Hecht J. The future of electronic health records. *Nature* 2019;573:S114–6.
26 Gliklich RE, Dreyer NA, Leavy MB. *Public-private partnerships*, 2014.
27 Naghavi M, Wang H, Lozano R, *et al*. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet* 2015;385:117–71.
28 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
29 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577–9.
30 OECD. *Health data governance: privacy, monitoring and research*. Paris, 2015.
31 Kyu HH, Abate D, Abate KH, *et al*. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1859–922.