

MIT Open Access Articles

Bias in machine learning applications to address non-communicable diseases at a population-level: a scoping review

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Birdi, S., Rabet, R., Durant, S. et al. Bias in machine learning applications to address non-communicable diseases at a population-level: a scoping review. BMC Public Health 24, 3599 (2024).

As Published: <https://doi.org/10.1186/s12889-024-21081-9>

Publisher: BioMed Central

Persistent URL: <https://hdl.handle.net/1721.1/157937>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution



SYSTEMATIC REVIEW

Open Access



Bias in machine learning applications to address non-communicable diseases at a population-level: a scoping review

Sharon Birdi¹, Roxana Rabet¹, Steve Durant¹, Atushi Patel¹, Tina Vosoughi¹, Mahek Shergill^{1,2}, Christy Costanian¹, Carolyn P. Ziegler²⁵, Shehzad Ali^{6,26,27}, David Buckeridge⁷, Marzyeh Ghassemi⁸, Jennifer Gibson⁹, Ava John-Baptiste¹⁰, Jillian Macklin^{1,11}, Melissa McCradden^{5,12,13}, Kwame McKenzie^{14,15}, Sharmistha Mishra^{16,17,18,19,20}, Parisa Naraei²¹, Akwasi Owusu-Bempah²², Laura Rosella^{5,23,28,29}, James Shaw²⁴, Ross Upshur^{4,5,9} and Andrew D. Pinto^{1,3,4,5*}

Abstract

Background Machine learning (ML) is increasingly used in population and public health to support epidemiological studies, surveillance, and evaluation. Our objective was to conduct a scoping review to identify studies that use ML in population health, with a focus on its use in non-communicable diseases (NCDs). We also examine potential algorithmic biases in model design, training, and implementation, as well as efforts to mitigate these biases.

Methods We searched the peer-reviewed, indexed literature using Medline, Embase, Cochrane Central Register of Controlled Trials and Cochrane Database of Systematic Reviews, CINAHL, Scopus, ACM Digital Library, Inspec, Web of Science's Science Citation Index, Social Sciences Citation Index, and the Emerging Sources Citation Index, up to March 2022.

Results The search identified 27 310 studies and 65 were included. Study aims were separated into algorithm comparison ($n = 13$, 20%) or disease modelling for population-health-related outputs ($n = 52$, 80%). We extracted data on NCD type, data sources, technical approach, possible algorithmic bias, and jurisdiction. Type 2 diabetes was the most studied NCD. The most common use of ML was for risk modeling. Mitigating bias was not extensively addressed, with most methods focused on mitigating sex-related bias.

Conclusion This review examines current applications of ML in NCDs, highlighting potential biases and strategies for mitigation. Future research should focus on communicable diseases and the transferability of ML models in low and middle-income settings. Our findings can guide the development of guidelines for the equitable use of ML to improve population health outcomes.

Keywords Population health, Non-communicable disease, Machine learning, Artificial intelligence

*Correspondence:

Andrew D. Pinto

andrew.pinto@utoronto.ca

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Background

Non-communicable diseases (NCDs), which include cardiovascular diseases, diabetes, cancers, and chronic respiratory diseases, are the leading cause of both burden of disease and death, globally, with a disproportionately higher rate of mortality in low- and middle-income countries (LMICs) [1, 2]. Population-level approaches of strengthening screening and detection are critical for identifying populations at high-risk of NCDs and informing early interventions [1]. One potential avenue for streamlining these interventions and lessening the burden of NCDs on the global population, is through artificial intelligence (AI) [3].

The increasing global interest in AI, particularly machine learning (ML), stems from the availability of large datasets and ever-growing computational power [4]. With its ability to learn and adapt from experience without explicit programming, ML has become crucial in various fields, such as healthcare [5]. However, alongside its remarkable potential, there are significant concerns associated with the widespread adoption of ML, notably the potential for algorithmic bias. *Algorithmic bias* in the context of AI and health systems is defined as: “the instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems” [6]. These biases, rooted in historical and systemic inequities, persistently affect marginalized groups, which reinforce prejudices. Marginalized groups, referring to individuals or communities who experience social, economic, or political disadvantages and discrimination, often bear the brunt of these amplified inequities in access to healthcare services and outcomes [7]. Reinforcing prejudices in this context means that predictive models, when trained on biased data or making decisions that align with historical disparities, inadvertently magnify these inequities, leading to deeper disparities in healthcare access, diagnosis, and treatment along socioeconomic, racial, gender, and ethnic lines [8–11].

Such algorithmic biases can manifest differently across different types of ML. For example, in supervised learning, which uses labelled datasets to classify data or predict outcomes, biases can enter the model through incomplete training data or data that are not representative and lead to inaccurate predictions for diverse populations. In unsupervised learning, biases can take the form of social biases, which encompass a range of prejudicial attitudes rooted in societal factors such as race, gender, and socioeconomic status. These biases may inadvertently emerge during algorithmic analysis of unlabelled data, potentially leading to unfair or discriminatory

outcomes, highlighting the importance of addressing them to foster equitable ML practices [12–14].

Recently, ML has been acknowledged for improving clinical care, yet less attention has been paid to its applications in population and public health and the potential for biases to arise during model design and development. Our objective was to conduct a scoping review to (1) identify studies that employ ML to address NCDs within the context of population and public health, and (2) to assess any algorithmic bias reporting that may have been exhibited during the design, training, and implementation of ML models, and how model developers mitigated these biases. Examining ML’s role in NCD surveillance informs more effective NCD management and resource allocation, while also addressing algorithmic bias detection to mitigate structural and systemic causes of marginalization in NCD research [12–14].

Methods

This scoping review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) statement [15]. The protocol for this review was submitted to Open Science Framework (available from osf.io/vkf24/) [16].

Databases

Due to the multidisciplinary nature of our area of interest, we considered many information sources covering both ML and NCDs. We searched the peer-reviewed, indexed literature using the following databases: *Medline (Ovid)*, *Embase (Ovid)*, *Cochrane Central Register of Controlled Trials and Cochrane Database of Systematic Reviews (Ovid)*, *CINAHL (EBSCOhost)*, *Scopus*, *ACM Digital Library*, *Inspec (Elsevier)*, and *Web of Science’s Science Citation Index*, *Social Sciences Citation Index*, and *Emerging Sources Citation Index*. All languages were included in the search. Commentaries, letters, editorials, conference proceedings were excluded. The databases were searched from 2000 to March 4–7, 2022 (inclusive). The range of publication dates was chosen to identify ML models that use the latest computing approaches and data.

Search strategy

A health information specialist (CZ) with Library Services, Unity Health Toronto, carried out comprehensive searches using a combination of subject headings and keywords, adapted for each database, for the broad concepts of AI and ML (e.g., artificial neural networks, decision trees, support vector machines) combined using the Boolean operator AND with the following five NCDs: cancers of the lung, trachea, and bronchus, ischemic heart disease, type 2 diabetes, chronic obstructive

pulmonary disease, Alzheimer's disease, and other dementias. Lung, tracheal, and bronchial cancers were chosen specifically as they represent a significant public health burden with high preventability, making them a priority area for exploring the applications of ML. We identified the aforementioned non-communicable diseases (NCDs) as part of the five primary clusters representing the greatest burden of morbidity and mortality caused by NCDs: cancer, cardiovascular disease, diabetes, chronic respiratory diseases, and neurological disorders as specified by the NCD Alliance [17]. Prior to de-duplication, the search yielded 48 701 results. After de-duplication in EndNote, 27 310 references remained. All the search strategies as run are available in Additional file 1 and have also been posted publicly on the Open Science Framework [18].

Eligibility criteria

All studies were required to meet the following eligibility criteria concerning the research focus, at both title/abstract and full-text screening : (1) relevant to population-level health and/or a public health approach; (2) pertain to at least one of the following conditions: cancer of the lung, trachea, and bronchus, ischemic heart disease, type 2 diabetes, chronic obstructive pulmonary disease, Alzheimer's disease, and other dementias; (3) describe the use of at least one ML model to address a real-world population or public health challenge. There were no language restrictions for the studies reviewed. All study designs were included.

Studies were excluded if: (1) they were not relevant to population-level health and/or a public health approach (i.e., the study focused on individual-level, clinical

applications of ML); (2) focus was not any of the conditions mentioned in the inclusion criteria or studies that focused on complications and conditions associated with the condition itself; (3) no-real world data was used; (4) commentaries, letters, editorials, conference proceedings, and dissertations (Table 1).

Screening process

DistillerSR was used to manage citations. We trained research assistants to review the citations and test the criteria on 50 randomly selected citations. The training was repeated with randomly selected blocks of 50 citations until inter-rater reliability was met ($\kappa > 0.9$). Reviewers screened the studies via a two-phase process: the title/abstract phase, referred to as first-level screening, and the full-text phase, referred to as second-level screening. The reviewers utilized the eligibility criteria to evaluate and determine the inclusion/exclusion of studies, which were then recorded in DistillerSR.

During first-level screening, two independent reviewers screened titles and abstracts of all imported studies to select studies for final review. If eligibility criteria were fully met, the studies were included. Studies that did not meet at least one inclusion criterion, as agreed upon by the reviewers, were excluded. Conflicts regarding the eligibility of certain studies were resolved through discussion and consensus among the reviewers. If consensus could not be reached, the research associate (CC) decided on inclusion/exclusion. Second-level screening involved reviewing the full-text of all studies that passed the title and abstract screening. This process was performed by a sole reviewer (SB), who excluded any studies

Table 1 Inclusion and exclusion criteria for the study articles

Inclusion Criteria	Exclusion Criteria
(1) Population-wide implications and/or a public health approach, which includes those pertaining to subsets of the general population at a certain point in life-course (e.g., seniors, children).	(1) Did not have a population-wide implication and/or public health approach, this included studies that focused on a population that was defined by one or multiple diseases, studies that focused on domains outside of public health systems or conventional population systems, studies that focused on high-risk groups (e.g., smokers) or in a specialised medical setting (e.g., hospitalized patients), or studies that focused on any subset of the population defined by socio-demographic characteristics other than age, such as ethnicity and sex.
(2) Pertained to at least one of the following conditions: cancer of the lung, trachea, and bronchus, ischemic heart disease, type 2 diabetes, chronic obstructive pulmonary disease, Alzheimer's disease, and other dementias.	(2) Focus was not any of the conditions mentioned in the inclusion criteria or studies that focused on complications and conditions associated with the condition itself (e.g., diabetic retinopathy).
(3) Described the use of at least one ML model (e.g., artificial neural networks, decision trees, support vector machines) to address a real-world population or public health challenge. There were no language restrictions for the studies reviewed.	(3) No-real world data was used, including general discussions of ML, studies that incorporated data from animal models or in-silico experiments, and proof-of-concept studies.
	(4) Commentaries, letters, editorials, conference proceedings, and dissertations.

that did not meet the same inclusion criteria as the first phase.

Data collection processing and synthesis

Four independent reviewers extracted data (AP, RR, SB, TV). An Excel data extraction form was developed based on the JBI Manual for Evidence Synthesis [19]. Two reviewers (SB, RR) pilot-tested this form on ten randomly selected studies. The four study team members independently extracted data from all included studies; the data extraction was then vetted by one study team member (SB).

The following data were extracted: author(s), title, journal, year, ML application type(s), intended purpose of ML, study design, intervention (if applicable), results, jurisdiction, data sources, unit(s) of analysis, sample size, demographics, identification of any potential algorithmic bias in the ML model (biases related to gender, sex, ethnicity, socioeconomic status), LMIC transferability, bias mitigation strategies, NCDs targeted, target population and setting, intended users, and impact reported by the author. We also noted if information was unavailable from an article or if any additional sources of algorithmic bias (e.g., age-related bias) were discussed. Narrative syntheses were performed on the extraction categories. The studies were summarized into a table outlining ML applications, ML application aims, jurisdictions, data sources, NCDs studied, considerations of biases and their mitigation (Table 2). The narrative synthesis and synthesis of study characteristics (Table 3) are presented in the Results section.

Results

Study selection

Our initial search yielded 27 310 citations. Following title/abstract screening, 275 abstracts remained. Following full-text screening by SB, 65 articles met eligibility criteria and were included in the final review (Fig. 1).

Publication and study characteristics

Table 2 presents a summary of the data extracted from each included study. Most of the studies ($n=59$, 90.77%) were published between 2017 and 2023; five studies ($n=5$, 7.69%) were published between 2011 and 2016; and one study ($n=1$, 1.54%) was published between 2005 and 2010 (Fig. 2).

Table 3 presents a synthesis of the characteristics of the included studies and the frequency with which each of the following subcategories was reported.

Application aims

Studies could be classified as either comparing ML models/approaches ($n=13$, 20.00%) [33, 34, 37, 38,

42–44, 52, 54, 59, 65, 77, 82] or using disease modelling for population-health related outputs ($n=52$, 80.00%) [20–32, 35, 36, 39–41, 45–51, 53, 55–58, 60–64, 66–76, 78–81, 83, 84]. The modelling of NCDs included measuring incidence in the population ($n=18$, 34.62%) [24, 25, 28, 36, 39, 40, 47, 53, 60–63, 68–70, 79, 81, 83], measuring risk in the population ($n=32$, 61.54%) [20–23, 26, 27, 30–32, 35, 41, 45, 46, 48–51, 55, 57, 58, 64, 66, 67, 71–76, 78, 80, 84] and evaluating the effectiveness of an intervention on outcomes as defined by study authors ($n=2$, 3.85%) [29, 56].

Data sources

Data sources used by the studies included longitudinal survey data ($n=27$, 41.54%) [20, 27–31, 33, 34, 38–41, 43–45, 49, 51, 54, 55, 57, 63, 64, 71, 74, 75, 78, 82], biomedical databases ($n=9$, 13.85%) [21, 24, 26, 36, 50, 53, 56, 59, 80], electronic medical records ($n=16$, 24.62%) [22, 25, 32, 35, 37, 52, 58, 62, 65, 66, 70, 73, 77, 81, 83, 84], social media textual elements ($n=1$, 1.54%) [23], administrative claims ($n=8$, 12.31%) [47, 48, 60, 61, 67, 69, 72, 79, 82], laboratory data ($n=1$, 1.54%) [42], cellular data ($n=1$, 1.54%) [76], search-engine queries ($n=1$, 1.54%) [68], and wearable sensors ($n=1$, 1.54%) [46].

Non-communicable diseases targeted

Almost 50% ($n=32$, 49.23%) [27, 32–35, 37, 39, 43, 46, 47, 50, 52, 55, 56, 58, 59, 61, 62, 64, 66, 67, 69, 71, 72, 74, 76–78, 80, 82–84] of included studies focused on type 2 diabetes. Almost a quarter examined Alzheimer's and other dementias ($n=14$, 21.54%) [25, 28–31, 36, 40, 41, 44, 48, 53, 60, 63, 75]. Around 30% of included studies focused on cardiovascular and respiratory diseases ($n=19$, 29.23%) [20–24, 26, 38, 42, 45, 49, 51, 54, 57, 65, 68, 70, 73, 79, 81], spanning specifically ischemic heart disease ($n=10$, 52.63%) [21, 23, 26, 38, 45, 54, 57, 65, 70, 73], chronic obstructive pulmonary disease ($n=4$, 21.05%) [24, 51, 68, 81], and cancer of the lung, trachea, and bronchus ($n=5$, 26.32%) [20, 22, 42, 49, 79].

Technical approaches

The approach most employed within the applications studied was supervised learning, aimed at resolving problems of, or completing tasks involving classification and/or regression. Included in this paradigm are approaches involving constructs such as decision trees, ensembles (in turn including bagging, boosting, and random forest constructs), algorithms such as k-nearest neighbor, and systems such as artificial neural and naïve Bayesian networks. In terms of specific technologies, some algorithms as well as constructs employed within the studies considered were support vector machine ($n=21$, 32.31%) [21, 23, 28, 33, 35, 36, 42, 44, 48, 52, 59, 62, 63, 65, 72–78],

Table 2 Summary of studies that use machine learning to address population and public health challenges surrounding non-communicable diseases (from 2000-present)

Author(s)	Machine Learning Application(s)	Year	Application Aim	Jurisdiction	Data Source	Non-Communicable Disease	Potential Bias; if yes, type?	Mitigation Strategy
Adams et al. [20]	Not Specified	2021	Modelling risk in population	Canada	Longitudinal survey	Cancer of Lung, Trachea, & Bronchus	No	No
Alaa et al. [21]	SVM, RF, DT	2019	Modelling risk in population	United Kingdom	Biomedical Database	CVD	Yes, ethnicity	No
Alexander et al. [22]	NLP	2019	Modelling risk in population	Australia	EMR	Cancer of Lung, Trachea, & Bronchus	No	No
Andy et al. [23]	SVM	2021	Modelling risk in population	United States	Social media textual elements	CVD	No	No
Baechle et al. [24]	NLP	2017	Modelling disease incidence in population	United States (USA)	Biomedical Database	COPD	No	No
Balaji et al. [25]	RF	2022	Modelling disease incidence in population	Germany	EMR	Alzheimer's & Other Dementias	No	No
Barbieri et al. [26]	Not Specified	2022	Modelling risk in population	New Zealand	Biomedical Database	CVD	Yes, sex	Yes, developed sex-specific deep learning models
Birk et al. [27]	Generalized Linear Model (GLM), Generalized linear mixed model (GLMM), RF, Elastic Net	2021	Modelling risk in population	India	Longitudinal survey	T2D	Yes, not specified	Yes, used GLMM to avoid introducing bias in training set that occurs by assuming responses within families are not correlated
Burnham et al. [28]	SVM	2014	Modelling disease incidence in population	Australia	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Byeon [29]	RF, DT	2021	Evaluating effectiveness of intervention	Korea	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Dallora et al. [30]	DT	2020	Modelling risk in population	Sweden	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Danoso et al. [31]	RF, XGBoost	2021	Modelling risk in population	United Kingdom	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Esmaily et al. [32]	ANN, SVM	2018	Comparison of models/approaches	Iran	Longitudinal survey	T2D	No	No
Esmaily et al. [33]	RF, DT	2015	Modelling risk in population	Iran	EMR	T2D	No	No
Fazakis et al. [34]	BN, RF, DT, LR	2021	Comparison of models/approaches	United Kingdom	Longitudinal survey	T2D	No	No
Ferdousi et al. [35]	RF, DT, BN, MLP, LR, kNN, SVM-Polykernel & SVM-RBFKernel, AdaBoos, Bagging,	2021	Modelling risk in population	Bangladesh	EMR	T2D	No	No
Ford et al. [36]	BN, RF, SVM, ANN	2019	Modelling disease incidence in population	United Kingdom	Biomedical Database	Alzheimer's & Other Dementias	No	No

Table 2 (continued)

Author(s)	Machine Learning Application(s)	Year	Application Aim	Jurisdiction	Data Source	Non-Communicable Disease	Potential Bias; if yes, type?	Mitigation Strategy
Gholipour et al. [37]	ANN, Multiple Regression	2018	Comparison of models/approaches	Iran	EMR	T2D	No	No
Goldman et al. [38]	ANN	2021	Comparison of models/approaches	United States	Longitudinal survey	CVD	No	No
Haneef et al. [39]	LDA, LR, Flexible Discriminant Analysis, DT, Boosted C5, XGBoost	2021	Modelling disease incidence in population	France	Longitudinal survey	T2D	Yes, age-related bias	Yes, random resampling to balance the training and data set (lack of older population in dataset)
Hu et al. [40]	RF, XGBoost, BN, LR	2021	Modelling disease incidence in population	China	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Jia et al. [41]	Markov Modelling	2020	Modelling risk in population	United States	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Kamis et al. [42]	RF, DT, SVM	2021	Comparison of models/approaches	United States	Laboratory Data	Cancer of Lung, Trachea, & Bronchus	No	No
Kim et al. [43]	DNN	2021	Comparison of models/approaches	South Korea	Longitudinal survey	T2D	No	No
Kim et al. [44]	DNN, RF, Adaboost, MLP, BN, SVM	2021	Comparison of models/approaches	South Korea	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Kim et al. [45]	ANN	2017	Modelling risk in population	South Korea	Longitudinal survey	CVD	No	No
Lam et al. [46]	Unsupervised Learning	2021	Modelling risk in population	United Kingdom	Wearable Sensor	T2D	No	No
Liao et al. [47]	RF, GB, Bagging	2019	Modelling disease incidence in population	United States	Administrative Claims	T2D	No	No
Lim et al. [48]	RF, MLP, SVM	2021	Modelling risk in population	South Korea	Administrative Claims	Alzheimer's & Other Dementias	No	No
Lim et al. [49]	ANN, Deep Belief Network (DBN)	2018	Modelling risk in population	South Korea	Longitudinal survey	Cancer of Lung, Trachea, & Bronchus	Yes, ethnicity	No
Liu et al. [50]	Ensemble Methods (Voting and Stacking)	2019	Modelling risk in population	China	Biomedical Database	T2D	No	No
Liu et al. [51]	ANN, DT, LR	2019	Modelling risk in population	China	EMR	T2D	No	No
Mani et al. [52]	BN, LR, kNN, RF, SVM	2012	Comparison of models/approaches	United States	EMR	T2D	No	No
Mar et al. [53]	RF	2022	Modelling disease incidence in population	Spain	Biomedical Database	Alzheimer's & Other Dementias	Yes, age-related bias	No
Masih et al. [54]	MLP	2021	Comparison of models/approaches	United States	Longitudinal survey	CVD	No	No

Table 2 (continued)

Author(s)	Machine Learning Application(s)	Year	Application Aim	Jurisdiction	Data Source	Non-Communicable Disease	Potential Bias, if yes, type?	Mitigation Strategy
Moon et al. [55]	Not Specified	2021	Modelling risk in population	South Korea	Longitudinal survey	T2D	No	No
Nayak et al. [56]	Local Linear Wavelet Neural Network (LLWN), Structured Singular Value (SSV), Simplex Method based Social Spider Optimization (SMSSO)	2022	Evaluating effectiveness of intervention	India	Biomedical Database	T2D	No	No
Neumann et al. [57]	Not Specified	2022	Modelling risk in population	Australia	Longitudinal survey	CVD	No	No
Ooka et al. [58]	RF	2021	Modelling risk in population	Japan	EMR	T2D	Yes, not specified	No
Owusu et al. [59]	BN, LR, MLP, SVM, DT	2017	Comparison of models/approaches	United Kingdom	Biomedical Database	T2D	Yes, sex	Combined human expertise with machine power to represent best strategy to test hypothesis on potential disease predictors
Park et al. [60]	Not Specified	2020	Modelling disease incidence in population	South Korea	Administrative Claims	Alzheimer's & Other Dementias	No	No
Park et al. [61]	ANN, Multiple Regression (MRM), Sequential Neural Network (SNN)	2001	Modelling disease incidence in population	United States	Administrative Claims	T2D	No	No
Patil et al. [62]	SVM	2022	Modelling disease incidence in population	India	EMR	T2D	No	No
Pekkala et al. [63]	SVM	2016	Modelling disease incidence in population	Finland	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Piko et al. [64]	LR	2020	Modelling risk in population	Hungary	Longitudinal survey	T2D	Yes, sex	Data of each sex was analysed separately; saw no prominent differences
Priyanga et al. [65]	SVM, kNN, ANN, BN	2020	Comparison of models/approaches	India	EMR	CVD	No	No
Ravaut et al. [66]	DT	2021	Modelling risk in population	Canada	EMR	T2D	Yes, unspecified	No
Razavian et al. [67]	Not Specified	2015	Modelling risk in population	United States	Administrative Claims	T2D	No	No
Rehman et al. [68]	RF	2021	Modelling disease incidence in population	United States	Search-Engine Queries	COPD	No	No

Table 2 (continued)

Author(s)	Machine Learning Application(s)	Year	Application Aim	Jurisdiction	Data Source	Non-Communicable Disease	Potential Bias, if yes, type?	Mitigation Strategy
Shangguan et al. [51]	BN	2021	Modelling risk in population	China	Longitudinal survey	COPD	Yes, sex and socioeconomic status	Adjusted for sex and income as risk factors of COPD
Su et al. [69]	ANN	2021	Modelling disease incidence in population	Taiwan	Administrative Claims	T2D	No	No
Su et al. [70]	RF	2020	Modelling disease incidence in population	China	EMR	CVD	No	No
Syed et al. [71]	Synthetic Minority Over-sampling Technique (SMOTE)	2020	Modelling risk in population	Saudi Arabia	Longitudinal survey	T2D	No	No
Uddin et al. [72]	SVM, RF, kNN, ANN, LR	2022	Modelling risk in population	Australia	Administrative Claims	T2D	Yes, not specified	No
Wang et al. [73]	SVM	2021	Modelling risk in population	China	EMR	CVD	No	No
Wang et al. [74]	SVM, RF	2021	Modelling risk in population	China	Longitudinal survey	T2D	No	No
Wu et al. [75]	Regularized Logistic Regression (r-LR), SVM, RF, Super Learner (SL)	2022	Modelling risk in population	China	Longitudinal survey	Alzheimer's & Other Dementias	No	No
Xie et al. [76]	BN, DT, LR, ANN, RF, SVM	2019	Modelling risk in population	United States	Cellular Data	T2D	No	No
Xiong et al. [77]	MLP, AdaBoost, RF, SVM, GB	2019	Comparison of models/approaches	China	EMR	T2D	Yes, ethnicity and socioeconomic status (SES)	No
Yang et al. [78]	LDA, SVM, RF	2020	Modelling risk in population	China	Longitudinal survey	T2D	Yes, not specified	Yes, not specified
Yeh et al. [79]	DNN	2021	Modelling disease incidence in population	Taiwan	Administrative Claims	Cancer of Lung, Trachea, & Bronchus	Yes, ethnicity	No
Yun et al. [80]	Deep Learning (DL)	2022	Modelling risk in population	United Kingdom	Biomedical Database	T2D	No	No
Zafari et al. [81]	Multilayer Neural Networks (MLNN), Extreme Gradient Boosting (XGB)	2022	Modelling disease incidence in population	Canada	EMR	COPD	No	No
Zhang et al. [82]	RF, ANN, DT	2020	Comparison of models/approaches	Australia	Longitudinal survey	T2D	Yes, sex and ethnicity	No
Zheng et al. [83]	NLP, RF	2016	Modelling disease incidence in population	United States	EMR	T2D	No	No

Abbreviations: ANN artificial neural network, BN naive bayesian network, COPD chronic obstructive pulmonary disease, CVD cardiovascular disease, DT Decision Tree, DNN deep neural network, EMR electronic medical record, GB gradient boosting, kNN k-Nearest Neighbour, LDA linear discriminant analysis, LR logistic regression, MLP multilayer perceptron, NLP natural language processing, RF random forest, SVM support vector machine, T2D type 2 diabetes

Table 3 Characteristics of included studies

	Characteristic	Frequency	%
Years of Publication	2005–2010	1	1.54
	2011–2016	5	7.69
	2017–2023	59	90.77
Jurisdictions	Bangladesh	1	1.54
	Canada*	3	4.62
	China	10	15.38
	Finland*	1	1.54
	France*	1	1.54
	Germany*	1	1.54
	Hungary*	1	1.54
	India	4	6.15
	Iran	3	4.62
	Japan*	1	1.54
	New Zealand*	1	1.54
	Saudi Arabia*	1	1.54
	South Korea*	8	12.31
	Spain*	1	1.54
	Sweden*	1	1.54
	Taiwan*	2	3.08
United Kingdom*	7	10.77	
United States*	13	20.00	
Application Aims	Modelling Risk in Population	32	61.54
	Modelling Disease Incidence in Population	18	34.62
	Evaluating Effectiveness of Intervention	2	3.85
	Comparison of Models/Approaches	13	20.00
Data Sources	Longitudinal Survey Data	27	41.54
	Biomedical databases	9	13.85
	Electronic medical records	16	24.62
	Social media textual elements	1	1.54
	Administrative claims	8	12.31
	Laboratory data	1	1.54
	Cellular data	1	1.54
	Search engine queries	1	1.54
Non-communicable diseases	Type 2 diabetes	32	49.23
	Alzheimer's & other dementias	14	21.54
	Cardiovascular disease	10	15.38
	Chronic obstructive pulmonary disease	4	6.15
	Cancer or the lung, trachea, and bronchus	5	7.69
Major technical approaches	Support vector Machine	21	32.31
	Multilayer perceptron	6	9.23
	Random forest	27	41.54
Bias considerations	No consideration	49	75.38
	Ethnicity-related bias	5	7.69
	Sex-related bias	5	7.69
	Age-related bias	2	3.85
	Socioeconomic status	2	3.85
Implementation of bias mitigation strategies	Not specified bias	5	7.69
	Yes	7	10.76
	No	58	89.23

Numerous ML approaches were used simultaneously in certain studies

More than one bias was considered in certain studies

*High-income economy (i.e., for 2023, with a gross national income (GNI) per capita of \$13 205 or more)

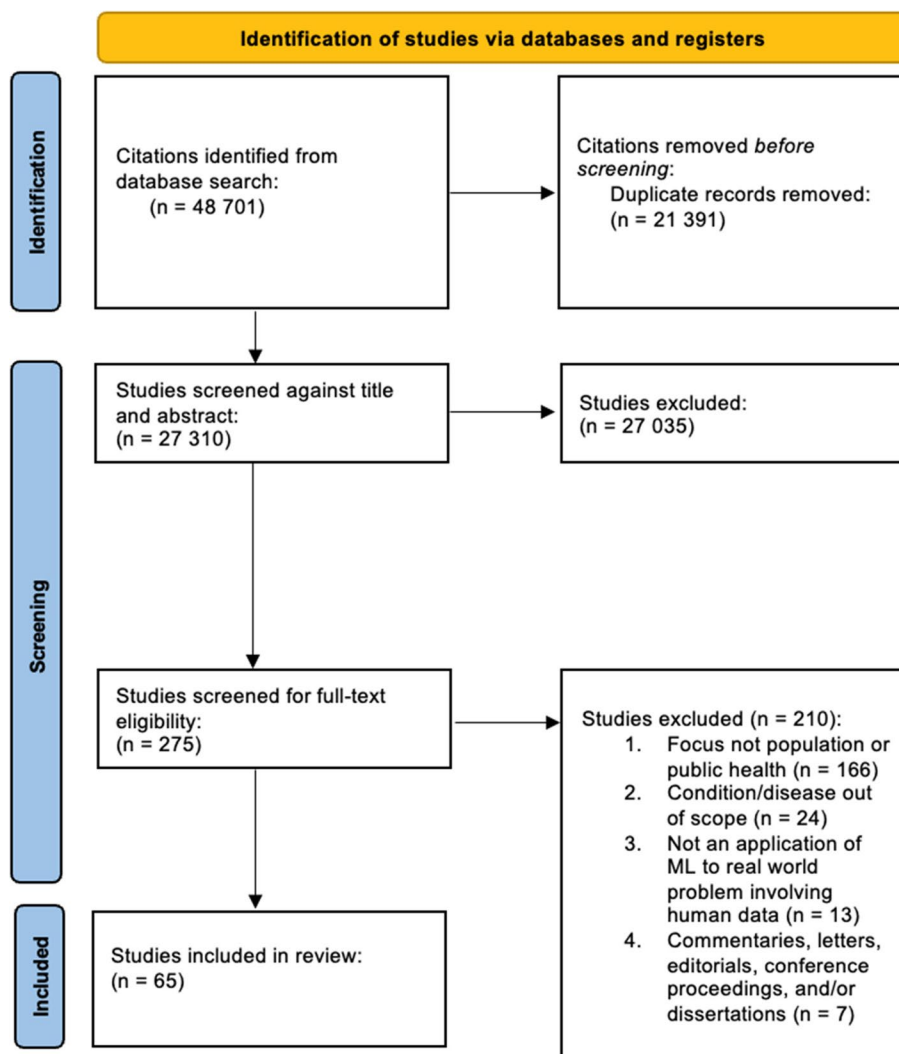


Fig. 1 PRISMA-ScR flow diagram

multilayer perceptron ($n=6$, 9.23%) [35, 44, 48, 54, 59, 77], random forest ($n=27$, 41.54%) [21, 25, 27, 29, 31, 32, 34–36, 40, 42, 44, 47, 48, 52, 53, 58, 68, 70, 72, 74–78, 82, 83]. Numerous ML approaches were used simultaneously in certain studies.

Consideration of bias and its mitigation

Although all the reviewed articles recommended applying their ML models in their specific NCD contexts, less than one-third ($n=16$, 24.62%) [21, 26, 27, 39, 49, 51, 53, 58, 59, 64, 66, 72, 77–79, 82] addressed the possibility of algorithmic bias that may arise from the implementation of their ML models. Of those studies mentioning bias potential, only 7 (7 out of 16, 43.75%) [26, 27, 39, 51, 59, 64, 78] outlined practical steps taken to mitigate bias. For the most part, those methods concerned mitigating sex-related bias and affected model design.

Countries and other regional divisions represented

Areas from which samples were drawn included Australia ($n=5$, 7.69%) [22, 28, 57, 72, 82], Bangladesh ($n=1$, 1.54%) [35], Canada ($n=3$, 4.62%) [20, 66, 81], China ($n=10$, 15.38%) [40, 50, 51, 70, 73–75, 77, 78, 84], Finland ($n=1$, 1.54%) [63], France ($n=1$, 1.54%) [39], Germany ($n=1$, 1.54%) [25], Hungary ($n=1$, 1.54%) [64], India ($n=4$, 6.15%) [27, 56, 62, 65], Iran ($n=3$, 4.62%) [32, 33, 37], Japan ($n=1$, 1.54%) [58], New Zealand ($n=1$, 1.54%) [26], Saudi Arabia ($n=1$, 1.54%) [71], South Korea ($n=8$, 12.31) [29, 43–45, 48, 49, 55, 60], Spain ($n=1$, 1.54%) [53], Sweden ($n=1$, 1.54%) [30], Taiwan ($n=2$, 3.08%) [69, 79], the United Kingdom ($n=7$, 10.77%) [21, 31, 34, 36, 46, 59, 80], and the United States ($n=13$, 20.00%) [23, 24, 38, 41, 42, 47, 52, 54, 61, 67, 68, 76, 83].

Most studies used datasets drawn from areas defined by the World Bank as high-income economies (i.e. for

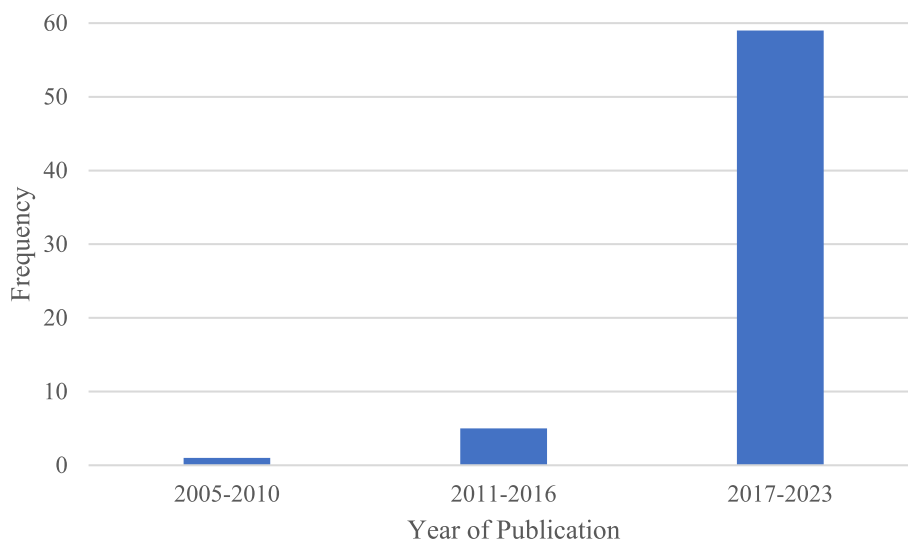


Fig. 2 Distribution of included studies by year of publication

2023, those with a gross national income (GNI) per capita of \$13 205 or more) ($n=15/19$ total countries; 78.95%) [85]. The remainder originated from countries classified as lower-middle-income (i.e. those with a GNI per capita of \$1 086 to \$4 255 [85], here specifically Bangladesh, India, and Iran) or upper-middle income (i.e. those with a GNI per capita of \$4 256 to \$13 204, here China) [85].

Discussion

Summary

In summary, we identified 65 peer-reviewed studies published since 2005 that applied ML methods to evaluate NCDs using a population health lens. Only 65 out of 27 310 references were eligible for our study, illustrating the lack of studies that comment on ML applications in population and public health, specifically concerning NCDs. The initial large reference yield may have been due to the novelty of ML and, after the onset of COVID-19, the increasing interest into population and public health. Although the literature adequately addressed types of data sources, to truly engage with issues of health equity, more work must be done to address algorithmic biases in ML which leaves a gap for researchers to explore.

Study selection and methodological considerations

We employed a rigorous selection process to determine which research studies would be included in our analysis. This process involved applying specific criteria, which ultimately led to the exclusion of certain studies. The reasons behind the exclusion of these studies are comprehensively outlined in Table 1. Some studies were later eliminated from consideration for various

reasons, despite initially meeting our inclusion criteria. Firstly, some studies' samples were obtained in ways that were not representative of the broader population. For instance, while one study by Muro et al. (2021) geared at identifying predictors of COPD diagnosis using data from many of the same individuals' annual medical check-up information across 21 years, these individuals were all employees of Hitachi, Ltd. [86], which could have systematically influenced some aspect of the data collection. Secondly, other studies' objectives, upon full-text examination, differed from what we identified during our initial screening. For example, one study titled "Predicting Lung Cancer in the United States: A Multiple Model Examination of Public Health Factors" appeared initially to model disease incidence and specify risk factors but ultimately focused on which emitted compounds are most harmful, and how population health can be improved by initiatives geared at transitioning the USA from non-renewable to renewable energy sources [42]. Because this study did not ultimately overview participants' data (i.e., no sample size was mentioned), it did not meet our inclusion criteria.

Future directions in NCD research and public health interventions

The diseases evaluated in this review are representative of the global burden of mortality from NCDs, emphasizing diabetes, ischemic heart disease, cancers, and chronic respiratory diseases [87]. Study-focus distribution was also indicative of respective disease-category burdens. For instance, 49.23% of studies centered on T2D. Mortality from diabetes is increasing at a higher rate than other NCDs [88]. While overall NCD age-standardized mortality rates decreased by 22% globally between 2000 and

2019 for those between the ages of 30 and 70, diabetes age-standardized mortality for the same group increased by 3% worldwide [88]. At the same time, there has been a notable 13% increase in mortality rates attributable to T2D in LMICs [88]. The burden of disease is of particular relevance to LMICs, where there is already a high burden of infectious diseases. However, this was not reflected in the ML applications examined across this review. In contrast to the substantial 48.48% of studies that predominantly focused on Type 2 Diabetes (T2D), a significantly smaller fraction of studies (around 16%) tackled cardiovascular diseases (CVDs). Although T2D mortality is increasing where CVDs' is decreasing, T2D is still directly responsible for fewer deaths: approximately two million relative to 17.9 million from CVDs [87]. As such, CVD-centered ML applications in health may be useful to prioritize, considering the vast spectrum of conditions which could be categorized as CVDs [89].

Advancements in ML approaches

From a technical standpoint, supervised learning was the most popular algorithm found in our search. Unsupervised learning was also employed in some studies, such as in Lam et al. (2021) [46]. These approaches can uncover patterns in data and identify subpopulations, making them particularly useful for exploratory analysis. Specifically, Lam et al. demonstrated the potential of continuous or periodic self-monitoring for early detection and screening of disease progression among subpopulations at risk of T2D, particularly those in a prediabetic state [46]. Principal component analysis (PCA), a popular dimensionality reduction technique, was used by Kim et al. (2021) [44] to predict not only future dementia patients but also other types of diseases using data that include limited input variables, making it useful in places with limited access to resources. The findings suggest that PCA can serve as a cost-effective tool for predicting future cases of dementia and other diseases, even with limited input variables [44]. Natural language processing (NLP) and text mining techniques were used by Alexander et al. (2020), Zheng et al. (2016), and Baechele et al. (2017) to extract information from electronic health records to identify disease patterns and risk factors [22, 24, 83]. The results of these studies demonstrate the potential of NLP and text mining techniques in extracting population health data from large-scale electronic health records, which could contribute to developing more targeted public health interventions.

Risk modelling was the most popular application of ML. Ravaut et al. (2021) and Barbieri et al. (2022) established a machine-learning model at a population level that accurately predicts the onset of T2D and CVD using

administrative health data up to 5 years in advance [26, 66]. The studies suggested that using ML and administrative health data can create effective population health planning tools to differentiate high-risk from low-risk populations for diabetes. This can assist in directing investments and interventions toward preventing NCDs and could also aid in mitigating individual-level complications.

Transferability of ML applications to resource-limited settings

There is a notable disparity in the frequency of ML applications between high-income and low- and middle-income countries. Populations classified as "low-income" by the World Bank were not included in the studies considered within this review. The application of ML models in jurisdictions that lack robust health records may be limited as these approaches rely on large-scale data sets to learn patterns and make predictions [90]. However, one study led by researchers from the United States explored the use of several ML techniques as a lower-cost alternative to prediabetes screening in resource-limited settings [27]. The authors used survey data from an FFQ completed by individuals from a rural region of Hyderabad, India, to calculate each participant's Global Diet Quality Score (GDQS) and predict their risk for T2D development. The global applicability of the GDQS combined with ML techniques served as a low-cost, easy-to-use method for identifying populations at high risk of developing diabetes, bypassing the need to screen all individuals using laboratory-based tests [27]. An example of a promising data source is social media textual elements, such as Facebook posts, to help predict the risk of an NCD. In a study by Andy et al., the discriminatory ability of social media posts to predict the 10-year risk of CVD was compared to that of pooled cohort risk equations [23]. The study results present a novel outlook for utilizing new and emerging digital data sources to identify potential risk factors by analyzing information recorded over several time points [23]. Accessing the rapidly generated data on social media platforms (e.g., posts) from consenting individuals offers an opportunity to collect and analyze unscripted information that can differ from the standard survey assessments.

Geographical representation and generalizability of data

Within countries, there was inadequate representation of different regions. For instance, one study collected a diverse set of demographic variables (i.e., participant diet and level of cultural participation) [74] but was limited in its generalizability to a broader Chinese population because it focused on data collected solely from the Shanxi region. Similarly, another study featuring a

sample from China focused on Nanjing. While focusing on regional samples provides insight into specific sub-populations' health in China, it also highlights a shortfall in data on other subpopulations. This could be because population health data and public health initiatives appear to be provincially governed.

This trend appears to be applicable to several nations. Other studies focused on regions such as East Azerbaijan [37] and Mashhad, Iran [33], and eight cities in Tamil Nadu, India [65]. One study aimed at predicting participation in a cognitive health promotion program among older adults in Seoul, South Korea, who had not been diagnosed with MND. This study focused on correlating intent to participate with various demographic factors such as level of education, smoking status, and cohabitation status [29]. The authors recognized the complex biopsychosocial nature of cognitive health and employed a methodology that equalized city-level representation, such as stratified clustered sampling of all 25 districts in Seoul [29]. Since most of South Korea's population resides in urban regions [91], generalizability to the entire population may not be as significant.

Addressing algorithmic bias in ML

With respect to algorithmic bias, there was an overall lack of discussion on identifying, defining, and mitigating bias in population health settings. Chen et al. (2021) explored how the potential for ML to exacerbate existing health disparities, especially during model development, is a concern that requires more attention [92]. The article stresses the importance of health data in ML models and notes how collected data can be biased, with a larger portion of the dataset leaning towards a specific biological sex or gender-identity, for example [92]. In this case, the model cannot be initialized due to imbalanced baseline representation [92]. A study by Barbieri et al. attempts to mitigate sex-related bias to detect CVD by developing sex-specific ML models, emphasising the improved calibration and discrimination enabling 5-year risk prediction [26]. Yet, it also emphasizes the need to further explore these models in countries with larger administrative health datasets [26]. However, even with larger datasets, algorithmic biases are still present [92]. People made vulnerable by social and economic policies, including transgender and gender-nonconforming individuals, undocumented immigrants, and racialized populations are often underrepresented, misrepresented, or missing from collected health data [92]. Demographic data collected in countries such as Canada and France, where race and ethnicity are not recorded in their nationalized health databases, makes race-based disparities extremely difficult to explore [92]. Ultimately, representative data collection is important in ensuring that datasets reflect

the public population [92]. Moreover, the lack of reporting on bias in studies on population health must be addressed if ML is regarded as being able to revolutionize global healthcare systems [92, 93].

Strengths and limitations

This review is novel in examining how ML has been applied to population and public health by a range of applications such as prediction, surveillance, and evaluating the effectiveness of interventions. Notably, we identified potential algorithmic biases and mitigation strategies. This review has several limitations. Firstly, a grey literature search was not conducted, thus possibly introducing selection bias. Next, we did not perform duplicate screening during the full-text screening phase to adhere to project timelines and resources. Similarly, while data extraction was vetted by the lead author, it was not conducted in duplicate. This approach may introduce the potential for bias, particularly in areas requiring subjective judgment, such as interpreting the inclusion criteria and determining which biases were discussed in the studies and how they were mitigated. Although our reviewer (SB) had extensive experience in the topic area, the lack of duplicate screening may affect the reliability of our findings. Additionally, although we did not place any restrictions on language, non-English articles were translated via Google Translate which is susceptible to some level of error. Finally, the terms *population health* and *machine learning* are not universally defined. Although we tried to encompass subtypes of machine learning in our search strategy, we may have excluded articles that could have relevance to the field. Along the same lines, increased recognition of the complexities of NCD-NCD interplays and, more broadly, the finer aspects of keyword delineation will characterize future work. This includes the preferred terminology used by structurally disadvantaged communities to describe their experiences and the terminological conventions used to discuss ML applications in health in languages other than English. Finally, this review acknowledges the limitation of focusing solely on lung, tracheal, and bronchial cancers. While this allowed for in-depth analysis within this specific scope, future research incorporating a wider range of cancer types is necessary to gain a more comprehensive understanding of ML applications in oncology.

Conclusion

This review provides an overview of current ML applications as well as the potential for bias and bias mitigation strategies. This was the first scoping review focused on ML applications for studying NCDs. LMIC transferability of such ML models was not discussed much, leaving a gap for researchers to investigate data

transparency methods, such as making codes and protocols open source. As the field of ML continues to evolve, there will be ample opportunity to capitalize upon the use of technology to improve population health (e.g., identifying high-risk subgroups); we hope our results will help to guide future research, such as the development of guidelines for the equitable use of machine learning.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12889-024-21081-9>.

Additional file 1.

Acknowledgements

The authors thank Pruthuvie Chandras for reviewing and assessing the eligibility of articles.

Authors' contributions

ADP, SD, CZ – conceptualization and methodology. SB, RR, AP, TV – data collection. SB – formal analysis. SB, RR, AP, TV – writing—original draft. All authors (SB, RR, SD, AP, TV, MS, CC, CZ, SA, DB, MG, JG, AJP, JM, MM, KM, SM, PN, AOB, LR, JS, RU, ADP) – writing—review and editing. All authors contributed to revising the manuscript for important intellectual content, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Funding

This project was supported in part by the Canadian Institutes for Health Research (#460906). ADP, SA, DB, LR, SM (CRC grant number 950-23264), and JS are supported by Canada Research Chairs. ADP is supported as a Clinician-Scientist by the Department of Family and Community Medicine, Faculty of Medicine at the University of Toronto and at St. Michael's Hospital, the Li Ka Shing Knowledge Institute, St. Michael's Hospital, and a CIHR Applied Public Health Chair in Upstream Prevention. The opinions, results and conclusions reported in this article are those of the authors and are independent from any funding sources.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Upstream Lab, MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, Unity Health Toronto, 30 Bond Street, Toronto, ON M5B 1W8, Canada. ²Michael G. DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada. ³Department of Family and Community Medicine, St. Michael's Hospital, Toronto, ON, Canada. ⁴Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, ON, Canada. ⁵Division of Clinical Public Health, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. ⁶Department of Epidemiology and Biostatistics, Western Centre for Public Health & Family Medicine, Western University, London, ON, Canada. ⁷Department of Epidemiology, Biostatistics

and Occupational Health, School of Population and Global Health, McGill University, Montreal, QC, Canada. ⁸Department of Electrical Engineering and Computer Science (EECS) and Institute for Medical Engineering & Science (IMES), MIT, Cambridge, MA, USA. ⁹Joint Centre for Bioethics, University of Toronto, Toronto, ON, Canada. ¹⁰Departments of Epidemiology & Biostatistics, Anesthesia & Perioperative Medicine, Schulich Interfaculty Program in Public Health, Western University, London, ON, Canada. ¹¹Undergraduate Medical Education, Faculty of Medicine, University of Toronto, Toronto, ON, Canada. ¹²Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada. ¹³Genetics & Genome Biology, SickKids Research Institute, Toronto, ON, Canada. ¹⁴Wellesley Institute, Toronto, ON, Canada. ¹⁵CAMH, Toronto, ON, Canada. ¹⁶Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, ON, Canada. ¹⁷MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, Unity Health Toronto, Toronto, ON, Canada. ¹⁸Institute of Medical Science, Faculty of Medicine, University of Toronto, Toronto, Canada. ¹⁹Institute of Health Policy, Management and Evaluation, Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. ²⁰ICES, Toronto, ON, Canada. ²¹Department of Computer Science, Toronto Metropolitan University, Toronto, ON, Canada. ²²Department of Sociology, Faculty of Arts & Sciences, University of Toronto, Toronto, ON, Canada. ²³Institute for Better Health, Trillium Health Partners, Toronto, ON, Canada. ²⁴Department of Physical Therapy, Faculty of Medicine, University of Toronto, Toronto, ON, Canada. ²⁵Library Services, Unity Health Toronto, St. Michael's Hospital, Toronto, ON, Canada. ²⁶Division of Epidemiology, Dalla Lana School of Public Health, Toronto, ON, Canada. ²⁷Department of Laboratory Medicine and Pathobiology, Temerty Faculty of Medicine, Toronto, ON, Canada. ²⁸Department of Health Sciences, University of York, York, UK. ²⁹WHO Collaborating Centre for Knowledge Translation and Health Technology Assessment in Health Equity, Ottawa Centre for Health Equity, Ottawa, ON, Canada.

Received: 21 March 2024 Accepted: 12 December 2024

Published online: 28 December 2024

References

- World Health Organization. Noncommunicable diseases. World Health Organization; 2022. <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>.
- Boutayeb A, Boutayeb S. The burden of non communicable diseases in developing countries. *Int J Equity Health*. 2005;4(1): 2.
- Artificial intelligence and data technology provide smarter health care – 4 solutions that have made a difference for noncommunicable diseases. Available from: <https://www.who.int/europe/news/item/14-12-2021-artificial-intelligence-and-data-technology-provide-smarter-health-care-4-solutions-that-have-made-a-difference-for-noncommunicable-diseases>. Cited 2023 Oct 4.
- Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol*. 2018;15(3):504–8.
- Berente N, Bin Gu, Recker J, Santhanam R. Managing artificial intelligence. *Manage Inform Syst Q*. 2021;45(3):1433–50.
- Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9(2). Available from: <https://doi.org/10.7189/jogh.09.020318>. Cited 2023 Apr 25.
- Sevelius JM, Luis G-M, -, Zamudio-Haas S, Mccree B, Ngo A, Jackson A, et al. NOTES FROM THE FIELD Research with Marginalized Communities: Challenges to Continuity During the COVID-19 Pandemic. *AIDS Behav*. 2020;24:2009–12. <https://doi.org/10.1007/s10461-020-02920-3>. Cited 2023 Sep 19.
- Buckeridge DL. Precision, equity, and public health and epidemiology informatics - a scoping review. *Yearb Med Inf*. 2020;29(1):226–30.
- Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):167–79.
- Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health*. 2019;1(4):e157-9.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56. Available from: <https://pubmed.ncbi.nlm.nih.gov/30617339/>. Cited 2023 Aug 29.

12. Rajkumar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–58.
13. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/30128552/>. Cited 2023 Aug 29.
14. Benjamin R. Assessing risk, automating racism. *Science*. 1979;366(6464):421–2. Available from: <https://www.science.org/doi/10.1126/science.aaz3873>. Cited 2023 Aug 29.
15. 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(7):467–73. Available from: <https://annals.org>. Cited 2023 Apr 17.
16. OSF. Available from: <https://osf.io/>. Cited 2023 Sep 19.
17. Bridging the Gap on NCDs: From global promises to local progress - Discussion paper | NCD Alliance. Available from: <https://ncdalliance.org/resources/bridging-the-gap-on-ncds-from-global-promises-to-local-progress-policy-brief>. Cited 2023 Sep 26.
18. Ziegler, Carolyn P. Literature search strategies for: machine learning applications to address non-communicable diseases at a population-level: a scoping review. Open Science Framework (OSF); 2023. osf.io/j7p9h.
19. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. *JBI Manual for Evidence Synthesis*. JBI; 2024. Available from: <https://synthesismanual.jbi.global>. <https://doi.org/10.46658/JBIMES-24-01>
20. Adams SJ, Mondal P, Penz E, Tyan CC, Lim H, Babyn P. Development and cost analysis of a lung nodule management strategy combining artificial intelligence and lung-RADS for baseline lung cancer screening. *J Am Coll Radiol*. 2021;18(5):741–51.
21. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14(5):e0213653.
22. Alexander M, Solomon B, Ball DL, Sheerin M, Dankwa-Mullan I, Preininger AM, et al. Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. *JAMIA Open*. 2020;3(2):209–15.
23. Andy AU, Guntuku SC, Adusmalli S, Asch DA, Groeneveld PW, Ungar LH, et al. Predicting cardiovascular risk using social media data: performance evaluation of machine-learning models. *JMIR Cardio*. 2021;5(1):e24473.
24. Baechle C, Agarwal A, Zhu X. Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients. *J Big Data*. 2017;4(1):1–18.
25. Balaji C, Suresh DS. Multi-class Recognition of Alzheimer's and Parkinson's diseases using Bag of Deep reduced Features (BoDrF) with Improved Chaotic Multi Verse Harris Hawks Optimization (CMVHHO) and Random Forest (RF) based classification for early diagnosis. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2022; 11(3):774–85.
26. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, et al. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. *Int J Epidemiol*. 2022;51(3):931–44.
27. Birk N, Matsuzaki M, Fung TT, Li Y, Batis C, Stampfer MJ, et al. Exploration of machine learning and statistical techniques in development of a low-cost screening method featuring the global diet quality score for detecting prediabetes in rural India. *J Nutr*. 2021;151:1105–1185.
28. Burnham SC, Faux NG, Wilson W, Laws SM, Ames D, Bedo J, et al. A blood-based predictor for neocortical Aβ burden in Alzheimer's disease: results from the AIBL study. *Mol Psychiatry*. 2014;19(4):519–26.
29. Byeon H. Can the Random Forests Model Improve the Power to Predict the Intention of the Elderly in a Community to Participate in a Cognitive Health Promotion Program? *Iran J Public Health*. 2021;50:315 <http://ijph.tums.ac.ir>.
30. Dallora AL, Minku L, Mendes E, Rennemark M, Anderberg P, Berglund JS. Multifactorial 10-year prior diagnosis prediction model of dementia. *Int J Environ Res Public Health*. 2020;17(18):1–18.
31. Danso SO, Zeng Z, Muniz-Terrera G, Ritchie CW. Developing an explainable machine learning-based personalised dementia risk prediction model: a transfer learning approach with ensemble learning algorithms. *Front Big Data*. 2021;4:613047.
32. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *J Res Health Sci*. 2018;18(2):e00412.
33. Esmaily H, Tayefi M, Ghayour-Mobarhan M, Amirabadizadeh A. Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes. *Iran Biomed J*. 2015;22(5):303–11.
34. Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, Moustakas K. Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access*. 2021;9:103737–57.
35. Ferdousi R, Hossain MA, El Saddik A. Early-stage risk prediction of non-communicable disease using machine learning in health CPS. *IEEE Access*. 2021;9:96823–37.
36. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC Med Inf Decis Mak*. 2019;19(1):248.
37. Gholipour K, Asghari-Jafarabadi M, Iezadi S, Jannati A, Keshavarz S. Modelling the prevalence of diabetes mellitus risk factors based on artificial neural network and multiple regression. *East Mediterr Health J*. 2018;24(8):770–7.
38. Goldman O, Raphaeli O, Goldman E, Leshno M. Improvement in the prediction of coronary heart disease risk by using artificial neural networks. *Qual Manag Health Care*. 2021;30(4):244–50.
39. Haneef R, Kab S, Hrzic R, Fuentes S, Fosse-Edorh S, Cosson E, et al. Use of artificial intelligence for public health surveillance: a case study to develop a machine learning-algorithm to estimate the incidence of diabetes mellitus in France. *Arch Public Health*. 2021;79(1):1.
40. Hu M, Shu X, Yu G, Wu X, Välimäki M, Feng H. A risk prediction model based on machine learning for cognitive impairment among Chinese community-dwelling elderly people with normal cognition: development and validation study. *J Med Internet Res*. 2021;23(2):e20298.
41. Jia Y, Chang CCH, Hughes TF, Jacobsen E, Wang S, Berman SB, et al. Predictors of dementia in the oldest old: a novel machine learning approach. 2020. Available from: <http://links.lww.com/WAD/A287>.
42. Kamis A, Cao R, He Y, Tian Y, Wu C. Predicting lung cancer in the united states: a multiple model examination of public health factors. *Int J Environ Res Public Health*. 2021;18(11):6127.
43. Kim H, Lim DH, Kim Y. Classification and prediction on the effects of nutritional intake on overweight/obesity, dyslipidemia, hypertension and type 2 diabetes mellitus using deep learning model: 4–7th Korea national health and nutrition examination survey. *Int J Environ Res Public Health*. 2021;18(11):5597.
44. Kim J, Lim J. A deep neural network-based method for prediction of dementia using big data. *Int J Environ Res Public Health*. 2021;18(10):5386.
45. Kim JK, Kang S. Neural network-based coronary heart disease risk prediction using feature correlation analysis. *J Healthc Eng*. 2017;2017:2780501.
46. Lam B, Catt M, Cassidy S, Bacardit J, Darke P, Butterfield S, et al. Using wearable activity trackers to predict type 2 diabetes: machine learning-based cross-sectional study of the UK Biobank accelerometer cohort. *JMIR Diabetes*. 2021;6(1):e23364.
47. Liao X, Kerr D, Morales J, Duncan I. Application of machine learning to identify clustering of cardiometabolic risk factors in U.S. adults. *Diabetes Technol Ther*. 2019;21(5):245–53.
48. Lim SJ, Lee Z, Kwon LN, Chun HW. Medical health records-based mild cognitive impairment (Mci) prediction for effective dementia care. *Int J Environ Res Public Health*. 2021;18(17):9223.
49. Lim K, Lee BM, Kang U, Lee Y, Lim K, Lee BM, et al. An optimized DBN-based coronary heart disease risk prediction. *Int J Comput Commun Control*. 2018;13:492.
50. Liu Y, Ye S, Xiao X, Sun C, Wang G, Wang G, et al. Machine learning for tuning, selection, and ensemble of multiple risk scores for predicting type 2 diabetes. *Risk Manag Healthc Policy*. 2019;12:189–98.
51. Shangguan C, Yu L, Liu G, Song Y, Chen J. Risk assessment of chronic obstructive pulmonary disease using a Bayesian network based on a provincial survey. *Pol Arch Intern Med*. 2021;131(4):345–55.
52. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc*. 2012;2012:606.
53. Mar J, Gorostiza A, Arrospide A, Larrañaga I, Alberdi A, Cernuda C, et al. Estimation of the epidemiology of dementia and associated

- neuropsychiatric symptoms by applying machine learning to real-world data. *Rev Psiquiatr Salud Ment.* 2022;15(3):167–75.
54. Masih N, Naz H, Ahuja S. Multilayer perceptron based deep neural network for early detection of coronary heart disease. *Health Technol (Berl).* 2021;11(1):127–38.
 55. Moon S, Jang JY, Kim Y, Oh CM. Development and validation of a new diabetes index for the risk classification of present and new-onset diabetes: multicohort study. *Sci Rep.* 2021;11(1):15748.
 56. Nayak M, Das S, Bhanja U, Senapati MR. Predictive Analysis for Cancer and Diabetes Using Simplex Method Based Social Spider Optimization Algorithm. *IETE J Res.* 2022; 69(10):7342–56. <https://doi.org/10.1080/03772063.2022.2027276>.
 57. Neumann JT, Thao LTP, Callander E, Chowdhury E, Williamson JD, Nelson MR, et al. Cardiovascular risk prediction in healthy older people. *Geroscience.* 2022;44(1):403–13.
 58. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. *BMJ Nutr Prev Health.* 2021;4(1):140–8.
 59. Owusu Adjah ES, Montvida O, Agbeve J, Paul SK. Data mining approach to identify disease cohorts from primary care electronic medical records: a case of diabetes mellitus. *Open Bioinforma J.* 2017;10(1):16–27.
 60. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ Digit Med.* 2020;3(1):46.
 61. Park J, Edington DW. A sequential neural network model for diabetes prediction. *Artif Intell Med.* 2001;23(3):277–93.
 62. Patil R, Tamane S, Rawandale SA, Patil K. A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. *Int J Electr Comput Eng.* 2022;12(1):524–33.
 63. Pekkala T, Hall A, Lötjönen J, Mattila J, Soininen H, Ngandu T, et al. Development of a late-life dementia prediction index with supervised machine learning in the population-based CAIDE study. *J Alzheimer's Disease.* 2017;55(3):1055–67.
 64. Piko P, Werissa NA, Fiatal S, Sandor J, Adany R. Impact of genetic factors on the age of onset for type 2 diabetes mellitus in addition to the conventional risk factors. *J Pers Med.* 2021;11(1):1–17.
 65. Priyanga P, Pattankar VV, Sridevi S. A hybrid recurrent neural network-logic-based whale optimization framework for heart disease prediction with electronic health records. *Comput Intell.* 2021;37(1):315–43.
 66. Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and validation of a machine learning model using administrative health data to predict onset of Type 2 diabetes. *JAMA Netw Open.* 2021;4:e2111315.
 67. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data.* 2015;3(4):277–87.
 68. Rehman NA, Counts S. Neighborhood level chronic respiratory disease prevalence estimation using search query data. *PLoS One.* 2021;16(6 June):e0252383.
 69. Su TJ, Lee FC, Pan TS, Wang SM. A hybrid fuzzy neural network analysis to the risk factors of type 2 diabetes. *Int. J. Mod. Phys. B, Condensed matter physics, statistical physics, applied physics.* 2021;35(14n16):2140035–.
 70. Su X, Xu Y, Tan Z, Wang X, Yang P, Su Y, et al. Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model. *J Clin Lab Anal.* 2020;34(9):e23421.
 71. Syed AH, Khan T. Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in Saudi arabia: a retrospective cross-sectional study. *IEEE Access.* 2020;8:199539–61.
 72. Uddin S, Imam T, Hossain ME, Gide E, Sianaki OA, Moni MA, et al. Intelligent type 2 diabetes risk prediction from administrative claim data. *Inf Health Soc Care.* 2022;47(3):243–57.
 73. Wang C, Zhao Y, Jin B, Gan X, Liang B, Xiang Y, et al. Development and validation of a predictive model for coronary artery disease using machine learning. *Front Cardiovasc Med.* 2021;8:8.
 74. Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, et al. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Med Inf Decis Mak.* 2021;21(1):105.
 75. Wu Y, Jia M, Xiang C, Lin S, Jiang Z, Fang Y. Predicting the long-term cognitive trajectories using machine learning approaches: a Chinese nationwide longitudinal database. *Psychiatry Res.* 2022;310:114434.
 76. Xie Z, Nikolayeva O, Luo J, Li D. Building risk prediction models for type 2 diabetes using machine learning techniques. *Prev Chronic Dis.* 2019;16(9):E130.
 77. Xiong XL, Zhang RX, Bi Y, Zhou WH, Yu Y, Zhu DL. Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese Adults. *Curr Med Sci.* 2019;39(4):582–8.
 78. Yang T, Zhang L, Yi L, Feng H, Li S, Chen H, et al. Ensemble learning models based on noninvasive features for type 2 diabetes screening: Model development and validation. *JMIR Med Inf.* 2020;8(6):e15431.
 79. Yeh MCH, Wang YH, Yang HC, Bai KJ, Wang HH, Li YCJ. Artificial intelligence-based prediction of lung cancer risk using nonimaging electronic medical records: deep learning approach. *J Med Internet Res.* 2021;23:e26256 JMIR Publications Inc.
 80. Yun JS, Kim J, Jung SH, Cha SA, Ko SH, Ahn YB, et al. A deep learning model for screening type 2 diabetes from retinal photographs. *Nutr Metabol Cardiovasc Dis.* 2022;32(5):1218–26.
 81. Zafari H, Langlois S, Zulkernine F, Kosowan L, Singer A. AI in predicting COPD in the Canadian population. *Biosystems.* 2022;211:104585.
 82. Zhang L, Shang X, Sreedharan S, Yan X, Liu J, Keel S, et al. Predicting the development of type 2 diabetes in a large australian cohort using machine-learning techniques: longitudinal survey study. *JMIR Med Inf.* 2020;8(7):e16850.
 83. Zheng L, Wang Y, Hao S, Shin AY, Jin B, Ngo AD, et al. Web-based real-time case finding for the population health management of patients with diabetes mellitus: a prospective validation of the natural language processing-based algorithm with statewide electronic medical records. *JMIR Med Inf.* 2016;4(4):e6328.
 84. Liu S, Gao Y, Shen Y, Zhang M, Li J, Sun P. Application of three statistical models for predicting the risk of diabetes. *BMC Endocr Disord.* 2019;19(1):126.
 85. WDI - The World by Income and Region. Available from: <https://dataopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html>. Cited 2023 Apr 24.
 86. Muro S, Ishida M, Horie Y, Takeuchi W, Nakagawa S, Ban H, et al. Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study. *JMIR Med Inform* 2021;9(7):e24796. <https://medinform.jmir.org/2021/7/e24796>. 2021 Jul 6;9(7):e24796. Cited 2023 Apr 18.
 87. World Health Organization. WHO Independent high-level commission on noncommunicable diseases: final report: it's time to walk the talk. 2019. Available from: <https://apps.who.int/iris/handle/10665/330023>. Cited 2023 Apr 24.
 88. Country snapshot of diabetes prevention and control in the Americas - PAHO/WHO | Pan American Health Organization. Available from: <https://www.paho.org/en/node/84821>. Cited 2023 Apr 24.
 89. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol.* 2017;70(1):1–25. Available from: <https://pubmed.ncbi.nlm.nih.gov/28527533/>. Cited 2023 Dec 10.
 90. Morgenstern JD, Buajitti E, O'neill M, Piggott T, Goel V, Fridman D, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10:e037860 <http://bmjopen.bmj.com/>.
 91. The Statistic of Seoul | 서울정책아카이브 Seoul Solution. Available from: <https://www.seoulsolution.kr/en/content/statistic-seoul>. Cited 2023 Apr 25.
 92. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci.* 2021;4(1):123–44.
 93. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol.* 2019;16(7):391–403.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.