Social bias in artificial intelligence algorithms designed to improve cardiovascular risk assessment relative to the Framingham Risk Score: a protocol for a systematic review

Ivneet Garcha 1, Susan Paula Phillips 2

ABSTRACT

Introduction Cardiovascular disease (CVD) prevention relies on timely identification of and intervention for individuals at risk. Risk assessment models such as the Framingham Risk Score (FRS) have been shown to over-estimate or underestimate risk in certain groups, such as socioeconomically disadvantaged populations. Artificial intelligence (AI) and machine learning (ML) could be used to address such equity gaps to improve risk assessment; however, critical appraisal is warranted before ML-informed clinical decision-making is implemented.

Methods and analysis This study will employ an equity-lens to identify sources of bias (ie, race/ethnicity, gender and social stratum) in ML algorithms designed to improve CVD risk assessment relative to the FRS. A comprehensive literature search will be completed using MEDLINE, Embase and IEEE to answer the research question: do AI algorithms that are designed for the estimation of CVD risk and that compare performance with the FRS address the sources of bias inherent in the FRS? No study date filters will be imposed on the search, but English language filters will be applied. Studies describing a specific algorithm or ML approach that provided a risk assessment output for coronary artery disease, heart failure, cardiac arrhythmias (ie, atrial fibrillation), stroke or a global CVD risk score, and that compared performance with the FRS will be eligible for inclusion. Papers describing algorithms for the diagnosis rather than the prevention of CVD will be excluded. A structured narrative review analysis of included studies will be completed.

Ethics and dissemination Ethics approval was not required. Ethics exemption was formally received from the General Research Ethics Board at Queen’s University. The completed systematic review will be submitted to a peer-reviewed journal and parts of the work will be presented at relevant conferences.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- Previous reviews of cardiovascular risk algorithms focus on statistical precision, performance and bias, while this systematic review employs a robust qualitative analytic approach to study inherent social biases.
- Medical and engineering studies will be identified by searching MEDLINE, EMBASE and IEEE, thus ensuring that all machine learning (ML) algorithms relevant to this review will be captured.
- All article screening and review, data extraction, and risk of bias assessment based on an adaptation of Prediction Risk Of Bias ASsessment Tool, a validated tool, will be done independently by two reviewers.
- This review is not designed to provide either an analysis of overall performance of ML algorithms, nor a meta-analysis, as methodologies and outcome measures are heterogeneous.
- Algorithm outputs will be reported, assessed for bias and compared with Framingham Risk Score outputs, but we will not engage in a quantitative analysis of precision and performance of the included algorithms.

INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death globally.1 CVD risk factors, such as hypertension, hyperlipidaemia, age, smoking and diabetes, are largely modifiable and provide an avenue for primary care physicians to exercise preventive efforts. However, prevention of CVD outcomes requires accurate identification of and timely intervention for individuals at risk.

There are a number of prediction models that are used to screen individuals for risk of developing heart disease or CVD outcomes. The Framingham Risk Score (FRS) is one of the most commonly used and widely validated of these tools. Since the inception of the FRS, there has been a significant improvement in CVD outcomes. The reasons for this improvement are multi-faceted but the role of the FRS as one part of these prevention efforts remains crucial.2 However, the FRS is not without its shortcomings. It has been shown to both over-estimate the risk of CVD...
for all patients, and to under-estimate risk in socioeconomically disadvantaged populations, racialised groups and women. These misclassifications of risk highlight a number of limitations of the FRS. First, it employs a simple regression fitting approach that implicitly assumes that CVD risk factors are related to outcomes in a linear fashion. As a result, the complex interplay and interaction among different CVD risk factors in shaping outcomes is oversimplified or missed entirely. Second, the FRS has limited generalisability because it is primarily derived from and validated in white populations of European descent and may not accurately predict risk in diverse populations. This limitation is particularly problematic from an equity perspective. Risk stratification serves as an important tool for primary care physicians to identify patients at higher risk and to connect them with appropriate resources. If patients from equity-seeking groups, who already experience marginalisation within the health system, are consistently misclassified as lower risk, they may have less access to health resources despite having more need for them. Thus, the FRS may not only produce inaccurate predictive assessments, but could lead to the inappropriate allocation of resources based on need. The result will be a magnification of inequity with traditionally marginalised groups receiving insufficient intervention while their counterparts are subjected to unnecessary interventions.

Artificial intelligence (AI) and machine learning (ML) offer the promise of addressing limitations of traditional CVD risk estimation tools such as the FRS. Regression models adjust for rather than analysing the interconnected impact between risk factors and outcomes. In assuming linear relationships between single risk factors and outcomes, they generally cannot match the ability of ML to elucidate non-linear relationships and the interplay among various risk factors. Further, regression analyses can identify the independent associations of, for example, sex, race and income with heart disease, but cannot determine whether there are intersections of these variables. AI is a ‘machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. ML is a branch of AI. It is a computational technique focused on using data and algorithms to mimic human learning. The learnt patterns from ML algorithms are used to guide tasks and make predictions. Researchers are increasingly turning to ML to create new risk assessment tools that will allow for more precise and more inclusive risk assessment across diverse populations, that is, personalise assessments based on a patient’s comprehensive and specific health profile. AI has the potential for precision in risk estimation, increased generalisability, and ultimately greater equity in health resource allocation.

The use of AI/ML also precipitates an array of clinical, social and ethical/legal issues that must be considered. The rapidity with which AI informed clinical decision-making is being touted and implemented highlights the urgent need for oversight and critical appraisal of AI technologies and ML algorithms. AI is not value free and can never be more accurate than the information that informs it. This information is generated from human decision-making and judgement and can embed a range of biases arising from and not limited to racial and gender disparities, commercial interests, funding sources, varying definitions of risk, flawed data sets used to train AI, and the legal landscape in which the AI technology is developed, implemented and used.

Identifying such bias will help inform the development of more equitable AI algorithms for CVD risk assessment. More specifically, in the context of the FRS, it is essential to ensure that ML algorithms are actually improving the precision of risk estimates for all populations. Perhaps even more important is that ML algorithms do not reproduce the biases and limitations inherent in the FRS. Rethinking risk estimation without accounting for the equity-compromising limitations of FRS will, potentially, exacerbate health inequities in the context of accessing treatment to prevent CVD. Thus, it is important to identify any biases or limitations now, so that ML innovations can have maximal and equitable impact once we reach the stage of general implementation in a clinical context.

Objectives
The aim of this systematic review is to identify ML algorithms designed to increase accuracy of CVD risk assessment relative to the FRS and critically appraise these algorithms using an equity-lens. Thus, this systematic review will address the following question: do ML algorithms designed to improve CVD risk prediction relative to the FRS, address and correct for known social biases inherent in the FRS tool?

METHODS AND ANALYSIS
A comprehensive literature search using MEDLINE, Embase and IEEE will be performed to identify all papers that propose an ML algorithm for CVD risk assessment and compare the performance of the proposed algorithm with the FRS. The search strategy, selection criteria, and inclusion and exclusion criteria for this review are reported in the sections that follow.

Search strategy
The search strategies for the three databases searched in this review (MEDLINE, EMBASE and IEEE) were developed in collaboration with a Health Sciences Librarian at Queen’s University with expertise in systematic review searches. Several preliminary searches were completed prior to finalising the search terms to ensure the search was oversampling and capturing a high proportion of relevant and eligible studies. The finalised search strategy is included in online supplemental appendix 1. The same search strategy will be used for MEDLINE and EMBASE. The use of expansive terms such as ‘algorithm’
and ‘machine learning’ were purposeful because they captured the broad array of ML techniques that could be used to develop algorithms relevant to our study. The third database to be searched is IEEE, which publishes in the field of ML and AI. For this database, a broad search with only the term ‘Framingham’ will be used to capture as many relevant articles as possible. Narrower searches were trialled, but this all-encompassing search yielded a greater number of relevant results.

Both qualitative and quantitative studies will be reviewed for inclusion. No study design or date filters will be imposed on the search. English language filters will be applied. However, those studies in languages other than English for which English abstracts or translations are available will be reviewed. As relevant studies are identified through the search strategy, two reviewers will independently check for additional relevant cited and citing articles. Additionally, the reviewers will verify that the searches are capturing five previously identified papers deemed relevant to the research question.

**Information sources**

Literature search strategies were developed using Medical Subject Headings (MeSH) and text words related to cardiovascular risk assessment, AI and ML, and the FRS. We will search MEDLINE (OVID interface, 1948 onwards), EMBASE (OVID interface, 1980 onwards) and IEEE (1988, onwards).

To ensure literature saturation, we will scan the reference lists of included studies or relevant reviews identified through the search. Any studies deemed relevant and meeting the inclusion criteria will be added to the review.

**Study records**

Literature search results will be uploaded to Covidence, a non-profit, internet-based software program that facilitates collaboration among reviewers during the study selection process. The program auto-populates abstracts, which will be used for the initial stages of the study selection process. One of the reviewers will then manually upload full texts of each paper into Covidence. The research team developed a screening process based on the inclusion/exclusion criteria outlined below. This process was piloted and refined prior to implementation.

**Eligibility criteria**

Studies are to be selected according to the criteria outlined below.

**Study design: inclusion/exclusion**

We will include studies that: (1) described a specific algorithm or ML approach; (2) the algorithm provided a risk assessment output for one of the following CVD outcomes: coronary artery disease, heart failure, cardiac arrhythmias (ie, atrial fibrillation), stroke or a global CVD risk score and (3) provided a comparative analysis with the FRS.

Papers describing algorithms for the diagnosis rather than the prevention of CVD will be excluded. Those describing algorithms that were modelling risk of developing a cardiovascular risk factor instead of one of the CVD outcomes described in the next section will also be excluded. No year of publication restriction will be applied, that is, papers from all years of publication up to the date of final searches will be included. Conference presentations and abstracts with no associated peer-reviewed paper will be excluded. Commentaries, letters and editorials not proposing a specific ML algorithm, and unpublished work or papers that have not undergone peer review will also be excluded.

**Participants**

We will include studies examining the general adult human population or healthy adult humans (18 years or older) with no history of CVD. We will also include studies on populations with specific CVD risk factors, such as obesity, but otherwise exclude studies restricted to populations with specific diseases or conditions (eg, rheumatoid arthritis, HIV, metabolic disorders) or those with a history of CVD.

**Outcomes**

The endpoints of interest in this systematic review are the risk of developing:

- CVD.
- Coronary artery disease.
- Heart failure.
- Cardiac arrhythmias.
- Stroke.
- Major adverse cardiovascular events.

These outcomes will not be the focus of the analysis but will be used to identify algorithms with a specific focus on CVD risk estimation and prevention. These endpoints will guide selection of algorithms that will be critically appraised for bias from an equity perspective.

When the defined endpoints above are not reported, we will extend analysis to include the relevant surrogate outcome of all-cause mortality, so long as the focus of the algorithm is CVD risk estimation.

Outcomes will be collected as reported. Due to possible variation in cardiovascular risk definitions, we will record definitions of cardiovascular risk as reported in individual studies.

**Timing**

Restriction due to length of follow-up are not applicable.

No year of publication restrictions will be applied to the search.

**Setting**

There will be no restrictions by type of setting.

**Language**

We will include articles reported in the English language (those studies in languages other than English for which
English abstracts or translations are available will be reviewed).

Selection process
Two reviewers will independently screen all results by title and abstract using the inclusion/exclusion criteria. Covidence will be used to record where there is consensus and/or disagreement between the reviewers. Where there is disagreement, consensus will be reached through discussion. Where there is uncertainty about eligibility, additional information will be sought from the study authors. Reasons for exclusion will be recorded using the Covidence platform. Neither of the reviewers will be blinded with respect to journal titles, study authors or institutions. All eligible studies will move to full text review, where both review authors will read the full text of each paper to determine relevance based on the inclusion/exclusion criteria. Relevant citations identified from the papers will also be added to the review to undergo the full screening process. Again, Covidence will be used to track agreement and disagreement; the latter will be resolved through discussion.

Data collection process and data items
A standardised form will be used to extract data on the following: purpose of the ML algorithm or study, comparison of the algorithm’s performance with Framingham, specifics of algorithm development (including the type of ML approach, funding, competing interests, ethics approval and involvement of community stakeholders), sociodemographic characteristics included in training data and potential sources of bias (eg, race, gender, social strata), privacy/ethical concerns, definition of CVD risk, algorithm output (including global CVD risk, risk of stroke, heart failure, arrhythmias or coronary artery disease, etc), intended end user of the algorithm and clinical utility attributed to algorithm by authors. The collection of this information is purposeful and aligns with this review’s objective of assessing whether ML algorithms are addressing social biases inherent to the FRS. The data extraction document captures the complete process from algorithm development to implementation. Social biases of the FRS and potentially of ML algorithms can arise from the populations used in development and validation (eg, homogeneity by race, sex, etc), the design of data analysis (eg, no risk stratification by socioeconomic strata), from failure to consider diversity of populations in implementation (potential disparities in access).

Assessment of all these dimensions, including whether the algorithms use representative populations in the training and validation sets, whether subgroup analyses to assess for model performance are completed in under-represented groups, whether models are used to identify risk factors unique to different subgroups, etc, will be crucial in assessing whether ML is addressing the existing social exclusivity and bias critiques of the FRS.

The form (online supplemental appendix 2) will be piloted by both reviewing authors and revised prior to use in the review in order to ensure the reviewers are using the form in a consistent manner in order to minimise the risk of bias. Two reviewers will carry out independent single extraction of data using the standardised form for each study that is selected for full text review.

Outcomes and prioritisation
The key outcomes of this systematic review will be: (1) a summary of sources of bias identified in each study; (2) a thematic summary of sources of biases across the included studies and (3) a descriptive analysis of whether ML algorithms are addressing the misclassification of risk inherent in the FRS.

Risk of bias in individual studies
To assess risk of bias within individual studies a modified Prediction Risk Of Bias ASsessment Tool (PROBAST) will be used. This 21-item tool is used to assess risk of bias of diagnostic and prognostic prediction model studies. It is organised into four distinct domains: participants, predictors, outcome and analysis. Our modified PROBAST template is incorporated into the data extraction document (online supplemental appendix 3). The rationale for modifying PROBAST is to conduct a risk of bias assessment tailored to the overall aims of this study.

Risk of bias assessments will be completed independently by the two review authors. Any disagreements will be resolved by discussion to reach consensus.

Data synthesis
A systematic thematic narrative synthesis approach will be used to analyse each paper’s extracted data. The thematic analysis will explore the findings from each paper and any relationships between included studies. Information will be presented in text and in summarising tables to highlight the characteristics and findings of the included studies. Findings will be organised by key question and by comparisons of relevant outcomes. Each study’s risk of bias will be reported. Because this review will not be engaging in a quantitative analysis nor looking for quantitative outcomes, the concern of studies presenting with overlapping populations is not applicable to our review.

Patient and public involvement
None.

ETHICS AND DISSEMINATION
Ethics approval was not required. Ethics exemption was formally received from the Research Ethics Board at Queen’s University. The completed systematic review will be submitted to a peer reviewed journal and parts of the work will be presented at relevant conferences.

Contributors IG and SPP developed the idea for the study and oversaw planning. IG and SPP developed the selection criteria, search strategy and data extraction criteria. IG and SPP drafted and edited the manuscript. IG and SP read, provided feedback and approved the final manuscript.

Funding This systematic review is funded by a Research Initiation Grant from the Centre for Studies in Primary Care in the Department of Family Medicine at Queen’s Centre for Studies in Primary Care in the Department of Family Medicine at Queen’s University. The completed systematic review will be submitted to a peer reviewed journal and parts of the work will be presented at relevant conferences.

Contributors IG and SPP developed the idea for the study and oversaw planning. IG and SPP developed the selection criteria, search strategy and data extraction criteria. IG and SPP drafted and edited the manuscript. IG and SP read, provided feedback and approved the final manuscript.

Funding This systematic review is funded by a Research Initiation Grant from the Centre for Studies in Primary Care in the Department of Family Medicine at Queen’s University. The completed systematic review will be submitted to a peer reviewed journal and parts of the work will be presented at relevant conferences.
REFERENCES


13 Noble SJ. Algorithms of oppression: how search engines reinforce racism. NYU Press.


Appendix 1 – Search Strategy

MEDLINE AND EMBASE SEARCH TERMS:
(cardiovascular disease OR (cardio* or cardia* or heart* or coronary* or angina* or ventric* or myocard* or pericard* or ischem* or arrhythmi* or atrial fibrillat* or tachycardi* or endocardi* or hypertension OR (hypertensi* or peripheral arter* disease*) OR ((high or increased or elevated) adj2 blood pressure) OR hyperlipidemia OR (hyperlipid* or hyperlip?emia* or hypercholesterol* or hypercholester?emia* or hyperlipoprotein?emia* or hypertriglycerid?emia*) OR (arteriosclerosis or arteriolosclerosis or atherosclerosis) OR diabetes OR stroke* OR (cerebrovascular accident* or CVA* or cerebrovasc* or cerebral vascular or apoplexy))

AND
(artificial neural network or machine learning or expert system or artificial intelligence or deep learning or neural networks or cognitive computing or computer aided or algorithm)

AND (prevent* or decision* or risk assessment)

AND Framingham

FILTERS:
Timing: No year of publication restrictions will be applied to the search. Restriction due to length of follow up are not applicable.

Setting: No restrictions by type of setting.

Language: English language only

IEEE

The only search term used on the IEEE database was ‘Framingham’ to capture as many relevant articles as possible. A broad search yielded a greater number of more relevant results. English language filters were applied, but no year of publication, setting, or other restrictions were applied.
## Appendix 2 – Data extraction document with modified PROBAST

**Citation of paper: [insert here]**

<table>
<thead>
<tr>
<th>Question</th>
<th>Relevant data from paper</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td></td>
</tr>
<tr>
<td>What is the stated purpose of the algorithm/study?</td>
<td></td>
</tr>
<tr>
<td>Is the algorithm designed to improve risk estimation compared to FRS?</td>
<td></td>
</tr>
<tr>
<td>Is the algorithm designed to address any shortcomings of the FRS?</td>
<td></td>
</tr>
<tr>
<td>Is the algorithm achieving its stated purpose?</td>
<td></td>
</tr>
<tr>
<td><strong>Algorithm Development</strong></td>
<td></td>
</tr>
<tr>
<td>Where is the algorithm developed?</td>
<td></td>
</tr>
<tr>
<td><strong>Ethics and Conflicts of Interest</strong></td>
<td></td>
</tr>
<tr>
<td>Does the paper include information about how the algorithm/authors was funded?</td>
<td></td>
</tr>
<tr>
<td>Does the author note any competing interests?</td>
<td></td>
</tr>
<tr>
<td>Has the algorithm/paper been peer reviewed and approved by ethics?</td>
<td></td>
</tr>
<tr>
<td>Do the authors address relevant privacy/ethics issues?</td>
<td></td>
</tr>
<tr>
<td>Have community stakeholders or stated end-users been involved in the development of the algorithm?</td>
<td></td>
</tr>
<tr>
<td><strong>Transparency</strong></td>
<td></td>
</tr>
<tr>
<td>Do the authors provide the algorithm code or describe the data processing process for the algorithm?</td>
<td></td>
</tr>
<tr>
<td>Where is the dataset coming from? Any potential sources of bias inherent in the data set?</td>
<td></td>
</tr>
<tr>
<td>What are the risk factor variables that are fed into the algorithm?</td>
<td></td>
</tr>
<tr>
<td><strong>look for race, gender, age, social strata, etc.</strong></td>
<td></td>
</tr>
<tr>
<td>Inputs</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
</tr>
<tr>
<td>Does the algorithm include the same risk factors as FRS?</td>
<td></td>
</tr>
<tr>
<td><strong>note any differences</strong></td>
<td></td>
</tr>
<tr>
<td>Does disease definition in the algorithm align with standard and accepted disease definition?</td>
<td></td>
</tr>
<tr>
<td>Do the risk factor labels create potential for bias, i.e., are they an appropriate proxy for the development of CVD risk?</td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td></td>
</tr>
<tr>
<td>Have the authors done basic checks (representation, calibration)?</td>
<td></td>
</tr>
<tr>
<td>Were the characteristics the same in the test and validation data sets?</td>
<td></td>
</tr>
<tr>
<td>Outputs</td>
<td></td>
</tr>
<tr>
<td>What are the target health conditions (i.e., risk is assigned for which outcomes)?</td>
<td></td>
</tr>
<tr>
<td>Does the algorithm produce an output that is identical or comparable to the output produced by the FRS?</td>
<td></td>
</tr>
<tr>
<td>Does the algorithm improve risk estimation compared to the Framingham Risk Score?</td>
<td></td>
</tr>
<tr>
<td>By how much?</td>
<td></td>
</tr>
<tr>
<td>Do the authors carry out analysis of risk by social strata?</td>
<td></td>
</tr>
<tr>
<td>If so, does the output produced align with current evidence on CVD risk/disease burden in different sub-populations?</td>
<td></td>
</tr>
<tr>
<td>Does the output of the algorithm address a gap in clinical care or have clinical utility? If so, what purpose does the output serve?</td>
<td></td>
</tr>
<tr>
<td>Who is the intended end user (physicians, patients, data scientists, etc)?</td>
<td></td>
</tr>
</tbody>
</table>