Bias Mitigation in Primary Healthcare Artificial Intelligence Models: Scoping Review

Maxime Sasseville, Steven Ouellet, Caroline Rhéaume, Malek Sahlia, Vincent Couture, Philippe Després, Jean-Sébastien Paquette, David Darmon, Frédéric Bergeron, Marie-Pierre Gagnon

Submitted to: Journal of Medical Internet Research
on: May 06, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **29**
Multimedia Appendixes **30**
Multimedia Appendix 1 ..... **30**
Multimedia Appendix 2 ..... **30**

Bias Mitigation in Primary Healthcare Artificial Intelligence Models: Scoping Review

Maxime Sasseville[1, 2] RN, PhD; Steven Ouellet[1] PhD; Caroline Rhéaume[2, 3, 4] MD, PhD; Malek Sahlia[5] MSc, MEng; Vincent Couture[1] PhD; Philippe Després[6] PhD; Jean-Sébastien Paquette[2, 3] MD; David Darmon[7] PhD; Frédéric Bergeron[8] MSI; Marie-Pierre Gagnon[1, 2] PhD

[1]Faculté des sciences infirmières, Université Laval Quebec CA
[2]VITAM Research Center on Sustainable Health Quebec CA
[3]Faculté de médecine, Université Laval Quebec CA
[4]Research Center of Quebec Heart and lungs Institute Quebec CA
[5]ENSI – École Nationale des Sciences de l'Informatique, Université de La Manouba La Manouba TN
[6]Département de physique, de génie physique et d'optique, Université Laval Quebec CA
[7]Département d'enseignement et de recherche en médecine générale, UFR, RÉTINES, HEALTHY, Université Côte d'Azur Nice FR
[8]Bibliothèque – Direction des services-conseils, Université Laval Quebec CA

Corresponding Author:
Maxime Sasseville RN, PhD
Faculté des sciences infirmières, Université Laval
1050, avenue de la Médecine
Quebec
CA

Abstract

Background: Artificial intelligence (AI) predictive models in primary healthcare can potentially lead to benefits for population health. Algorithms can identify more rapidly and accurately who should receive care and health services, but they could also perpetuate or exacerbate existing biases toward diverse groups. We noticed a gap in actual knowledge about which strategies are deployed to assess and mitigate bias toward diverse groups, based on their personal or protected attributes, in primary healthcare algorithms.

Objective: To describe attempts, strategies, and methods used to mitigate bias in primary healthcare artificial intelligence models. To identify which diverse groups or protected attributes have been considered. To evaluate the results on bias attenuation and AI models performance of these attempts, strategies, and methods.

Methods: We conducted a scoping review informed by the Joanna Briggs Institute (JBI) review recommendations. An experienced librarian developed a search strategy in four databases (Medline (OVID), CINAHL (EBSCO), PsycInfo (OVID), and Web of Science) to identify sources published between 2017-01-01 and 2022-11-15. We imported data in Covidence and pairs of reviewers independently screened titles and abstracts, applied the selection criteria, and performed full-text screening. Any discrepancies regarding the inclusion of studies were resolved through consensus. Based on reporting standards for AI in health care, we performed data extraction - study objectives, models' main features, diverse groups concerned, mitigation strategies deployed, and results. Using the Mixed-Methods Appraisal Tool (MMAT), we appraised the quality of studies.

Results: After removing 585 duplicates, we screened 1018 titles and abstracts. From remaining 189 after exclusion, we excluded 172 full texts and included 17 studies. The most investigated personal or protected attributes were Race (or Ethnicity) in (12/17), and Sex (mostly identified as Gender in studies), using binary "male vs female" in (10/17) of included studies. We grouped studies according to bias mitigation attempts into the following categories: 1) existing AI models or datasets, 2) sourcing data such as Electronic Health Records, 3) developing tools with "human-in-the-loop" and 4) identifying ethical principles for informed decision-making. Mathematical and algorithmic preprocessing methods, such as changing data labeling and reweighing, along with a natural language processing method using data extraction from unstructured notes, showed the greatest potential. Other methods to enhance model fairness include group recalibration and the application of the equalized odds metric, which either exacerbated predictions errors between groups or resulted in overall models miscalibrations.

Conclusions: Results suggests that biases toward diverse groups can be more easily mitigated when data are open-sourced, multiple stakeholders are involved, and during the algorithm' preprocessing stage. Further empirical studies, considering more

diverse groups, such as nonbinary gender identities or Indigenous peoples in Canada, are needed to confirm and to expand this knowledge. Clinical Trial: OSF Registries qbph8; https://osf.io/qbph8

(JMIR Preprints 06/05/2024:60269)
DOI: https://doi.org/10.2196/preprints.60269

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Bias Mitigation in Primary Healthcare Artificial Intelligence Models: Scoping Review

Authors: Maxime Sasseville; Steven Ouellet; Caroline Rhéaume; Malek Sahlia; Vincent Couture; Philippe Després; Jean-Sébastien Paquette; David Darmon; Frédéric Bergeron; Marie-Pierre Gagnon

Abstract

Background: Artificial intelligence (AI) predictive models in primary healthcare can potentially lead to benefits for population health. Algorithms can identify more rapidly and accurately who should receive care and health services, but they could also perpetuate or exacerbate existing biases toward diverse groups. We noticed a gap in actual knowledge about which strategies are deployed to assess and mitigate bias toward diverse groups, based on their personal or protected attributes, in primary healthcare algorithms.

Objectives: To describe attempts, strategies, and methods used to mitigate bias in primary healthcare artificial intelligence models. To identify which diverse groups or protected attributes have been considered. To evaluate the results on bias attenuation and AI models performance of these attempts, strategies, and methods.

Methods: We conducted a scoping review informed by the Joanna Briggs Institute (JBI) review recommendations. An experienced librarian developed a search strategy in four databases (Medline (OVID), CINAHL (EBSCO), PsycInfo (OVID), and Web of Science) to identify sources published between 2017-01-01 and 2022-11-15. We imported data in Covidence and pairs of reviewers independently screened titles and abstracts, applied the selection criteria, and performed full-text screening. Any discrepancies regarding the inclusion of studies were resolved through consensus. Based on reporting standards for AI in health care, we performed data extraction - study objectives, models' main features, diverse groups concerned, mitigation strategies deployed, and results. Using the Mixed-Methods Appraisal Tool (MMAT), we appraised the quality of studies.

Results: After removing 585 duplicates, we screened 1018 titles and abstracts. From remaining 189 after exclusion, we excluded 172 full texts and included 17 studies. The most investigated personal or protected attributes were Race (or Ethnicity) in (12/17), and Sex (mostly identified as Gender in studies), using binary "male vs female" in (10/17) of included studies. We grouped studies according to bias mitigation attempts into the following categories: 1) existing AI models or datasets, 2) sourcing data such as Electronic Health Records, 3) developing tools with "human-in-the-loop" and 4) identifying ethical principles for informed decision-making. Mathematical and algorithmic preprocessing methods, such as changing data labeling and reweighing, along with a natural language processing method using data extraction from unstructured notes, showed the greatest potential. Other methods to enhance model fairness include group recalibration and the application of the equalized odds metric, which either exacerbated predictions errors between groups or resulted in overall models miscalibrations.

Conclusions: Results suggests that biases toward diverse groups can be more easily mitigated when data are open-sourced, multiple stakeholders are involved, and during the algorithm' preprocessing stage. Further empirical studies, considering more diverse groups, such as nonbinary gender identities or Indigenous peoples in Canada, are needed to confirm and to expand this knowledge.

Trial Registration: OSF Registries qbph8; https://osf.io/qbph8

Keywords: Artificial Intelligence; Algorithms; Expert System; Decision Support; Bias; Community Health Services; Primary Health Care; Health Disparities; Social Equity; Scoping Review

Introduction

Computer science developments have led to artificial intelligence (AI) models learning from large datasets and capable of independent analysis **[1-4]**. Great progress has been made for these tasks with the development of machine learning (ML). This branch of AI is devoted to understanding, generating, and reasoning based on data without implicit human instructions **[2,3]** Such ML algorithms are using data sets, called "training data sets", to capture the patterns required for clustering tasks or prediction modeling **[3,4]**. These models are now used in multiple contexts and industries to predict the likelihood of an event or to support human decision making **[4]**. In healthcare, AI models applied in radiology can potentially detect accurately and predict the progression of cancerous tumors **[5]**. Algorithms could also be useful in community-based primary health care (CBPHC) for identifying individuals, such as heart failure or diabetes outpatients, who require specific health care services **[6]**. As defined by the Canadian Institutes of Health Research, CBPHC encompasses a comprehensive array of services aimed at community well-being, incorporating primary prevention (including public health), health promotion, disease prevention, diagnosis, treatment, and management of chronic and episodic illnesses, rehabilitation support, and end-of-life care **[7]**.

Despite the potential AI benefits, such as compensating for workers shortage and maximizing access to CBPHC **[6]**, algorithm biases toward diverse groups can hinder their application in healthcare settings. These biases can be perpetuated when protected attributes **[1]**, as identified by the PROGRESS-Plus framework **[8]**, are sub or misrepresented in algorithms' training data **[1,9]**. Strategies aimed at identifying and mitigating bias, defined as a persistent inclination either in favor or toward something **[9]**, in predictive models are in development and are beginning to be empirically applied **[10,11]**. In computer science, algorithmics fairness attempts can be performed using 1) pre-processing, 2) in-processing or even, 3) post-processing strategies such as in the case of "out-of-the-box" commercial AI models **[4]**. Academic disciplines other than computer science, such as medicine, management, and ethics, are closely involved and intertwined in these issues of identifying potential bias toward diverse groups in AI models **[1,3]**. However, there remains knowledge gap in identifying which strategies and methods have been empirically applied to mitigate bias toward diverse groups in CBPHC algorithms **[10,12]**.

To address this gap, we conducted a scoping review aiming to identify and describe: 1) the attempts to made to mitigate bias in primary health care AI models, 2) which diverse groups or protected attributes have been considered, and 3) the results regarding on bias attenuation and the overall performance of the models.

Methods

Search Strategy

We conducted a scoping review informed by the Joanna Briggs Institute (JBI) **[13]**, and we used the Population [or Participant], Concept, and Context (PCC) Framework **[14]** as shown in **Table 1**.

Table 1. PCC (Population [or Participant], Concept, and Context) framework used for the search strategy[a]

PCC elements [**14**]	Definition (per JBI Reviewer's Manual Ch 11)	PCC elements applied in this review
Population	"Important characteristics of participants, including age and other qualifying criteria" (11.2.4)	Any diverse groups **[8]**, based on their personal or protected attributes **[1]**.

Concept	"The core concept examined by the scoping review should be clearly articulated to guide the scope and breadth of the inquiry. This may include details that pertain to elements that would be detailed in a standard systematic review, such as the 'interventions' and/or 'phenomena of interest' and/or 'outcomes'" (11.2.4)	Strategies, attempts or methods to assess and mitigate bias in artificial intelligence.
Context	"May include…cultural factors such as geographic location and/or specific racial or gender-based interests. In some cases, context may also encompass details about the specific setting."	Community-based primary health care **[7]**.

[a]PCC framework [**14**]. *Topic: Bias Mitigation in Primary Healthcare Artificial Intelligence Models. Primary review questions: 1) What are the attempts to mitigate bias in primary health care AI models, 2) Which diverse groups or protected attributes have been considered, and 3) What are the results on bias attenuation and model performance?*

An experienced librarian carried out a search strategy in November 2022, focusing on the main concepts of our research questions and adapted it for 4 relevant databases (Medline (OVID), CINAHL (EBSCO), PsycInfo (OVID), and Web of Science). The search covered the last 5 years (2017-2022) aligned with the concept's emergence time frame. Details of the search strategy can be found in [**Multimedia Appendix 1**]. Additionally, we searched the reference lists of included studies to identify additional references.

Data Collection

We imported all records (n = 1603) into the web-based collaborative tool Covidence (*Veritas Health Innovation*) **[15]**, and 585 duplicates were identified and removed. At the first "Title and abstract" screening phase, seven reviewers independently evaluated titles and abstracts by applying inclusion and exclusion criteria. All 1018 titles and abstracts underwent dual screening (**Table 2**). Any discrepancies regarding the inclusion of studies were resolved through consensus.

Table 2. Inclusion and exclusion criteria.

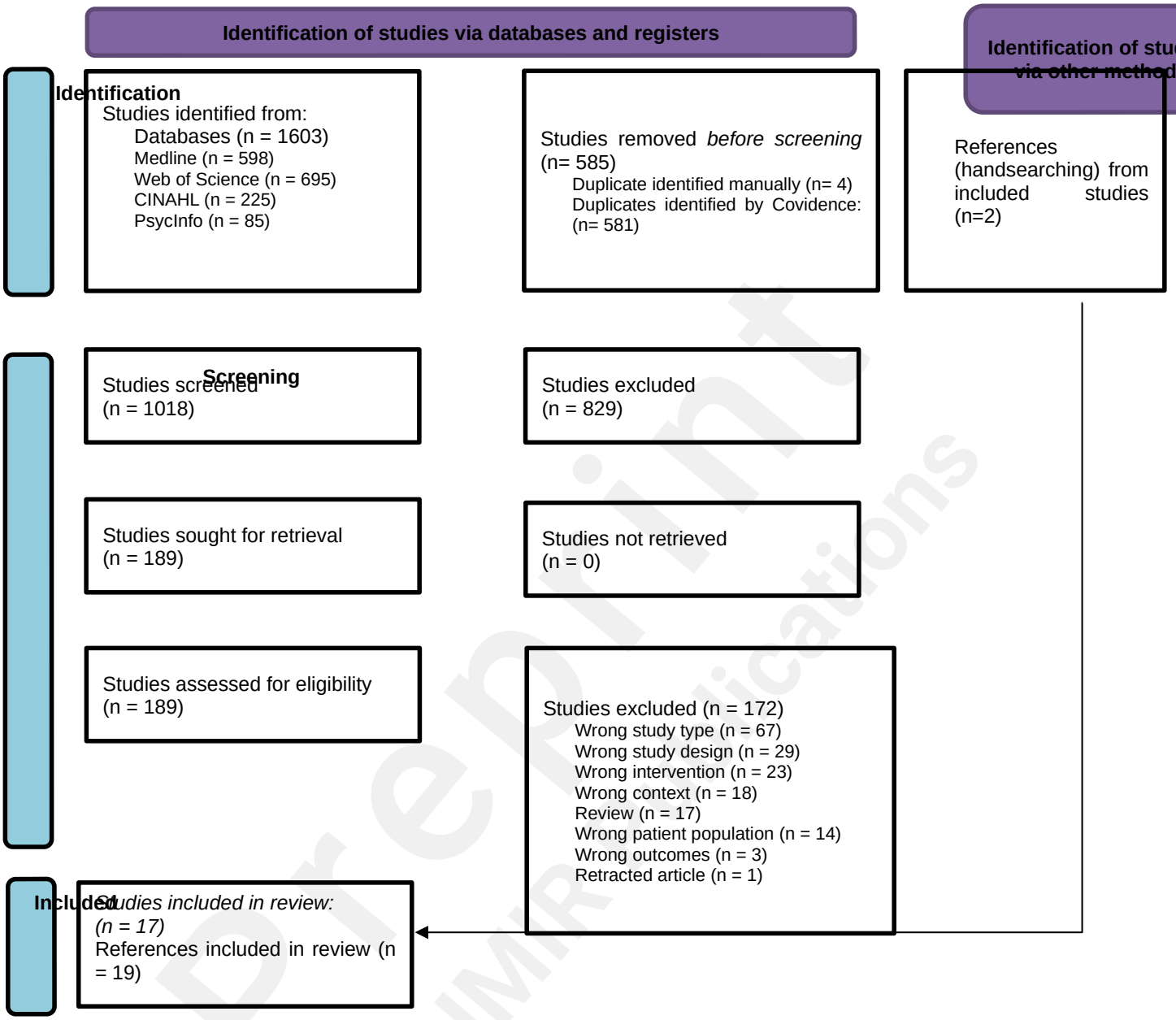
PCC (Population, Concept, and Context) elements **[14]**	Inclusion criteria	Exclusion criteria
Population	• Any populations targeted by CBPHC[a] interventions.	• Any populations targeted by hospital or specialized care interventions.
Concept	• All methods or strategies deployed to assess and mitigate bias toward diverse groups or protected attributes in AI models. • All mitigation methods or strategies deployed to promote and increase equity, diversity, and inclusion in CBPHC algorithms.	• Methods or strategies deployed to assess and mitigate bias on the AI model itself (e.g., biased prediction of a treatment effects), rather than bias related to individuals' characteristics/protected attributes. • Strategies, methods, or interventions not CBPHC related. • CBPHC interventions that do not include any algorithm/AI[b] system.
Context	• Include all CBPHC algorithms (AI) applications that can perpetuate/introduce potential biases toward diverse groups in terms of their characteristics/protected attributes.	• Algorithms used by primary health care providers for support in administrative tasks and for operational aspects, rather than for clinical decisions.
Study design and time frame	• All empirical studies, published in English or French, between 2017 and 2022.	• Reviews, opinion, commentary, editorial content, conference paper, communication, protocol, magazine articles, etc.

[a]CBPHC: Community-based primary health care.

[b]AI: Artificial intelligence.

Of the remaining 189 articles assessed for eligibility at the "Full text review" stage, we searched and obtained all missing full texts of selected references and imported them into Covidence. Five reviewers independently applied the same selection criteria, and all exclusion motives were recorded in Covidence. All full texts underwent dual screening, and any discrepancies were resolved through team consensus. All exclusion motives of the 172 references excluded were recorded in Covidence. As for the previous stage, any discrepancies concerning the included studies were resolved by consensus. We also hand-searched [**16**] and identified two relevant articles **[17,18**] related to two included studies [**19,20**], which were added to Covidence for extraction. The PRISMA flow diagram can be consulted in **Figure 1**.

Figure 1. PRISMA flow diagram



Data Extraction

One senior reviewer performed the extraction of the 17 included studies and at least one senior researcher validated all of them. Based on reporting standards for AI in health care **[21]**, we extracted general information (title of paper, year of publication, lead author, country); study objective; discipline/study design; AI model features/study population and setting; AI model architecture/evaluation; bias assessment method/strategy features and deployment; diverse groups concerned; bias mitigation results; and impact on AI model performance/accuracy results.

Quality Assessment

One senior reviewer appraised the quality of the 17 included studies by applying the Mixed-Methods

Appraisal Tool (MMAT) **[22,23]** and at least one senior researcher validated each of them. Most studies had a quantitative descriptive study design (14/17), while two employed a mixed methods design, and one used a qualitative design. All studies showed high quality, receiving scores of three or four stars (on a possibility of five). All MMAT scores can be found in [**Multimedia Appendix 2**].

Data Synthesis

Informed by the JBI recommendations **[24]**, we synthesized data using structured narrative summaries around our review concepts (e.g., model data source, model input, model output, diverse groups/protected attributes) concerned, mitigation strategies deployed, and results on bias mitigation and models overall performance. We reported our results based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) **[25]**.

Results

Overview

Of the 17 included studies, published between 2019 and 2022, we identified seven studies in the discipline of data science or informatics, seven in medical informatics, one in medical ethics/informatics, one in medical ethics using a Delphi method, and one in management care ethics using a user-centered design. Most studies have been conducted in the United States of America (15/17), one in the United Kingdom, and one in Italy. The main characteristics of the included studies can be consulted in **Table 3**.

Table 3. Characteristics of included studies

Lead author and year/ Country/ Discipline/ Study design	Objectives	AI model features/ Data source/Model input	Model output/ Computed result of the model	Diverse groups (protected attributes) concerned	Mitigation strategies deployed	Results on protected attributes bias mitigation and on models' performance
Alday et al, 2022 **[19]**; Reyna et al, 2021**[17]** (added by handsearching) USA Informatics. Case study (a reduced-bias machine learning design model proposal)	1) To compare performance across demographics of electrocardiogram (ECG) automatic classification algorithms **[19].** 2) To propose a machine-learning method aimed at mitigating bias and enhance health equity **[19]**.	Data from various databases/recordings comprised a range of reduced-lead ECGs from patients with cardiovascular diseases **[17]**. The assessment method focused on algorithms designed for two-lead ECGs **[19].**	Open-source algorithms capable of automatically detecting cardiac abnormalities in ECG recordings using only the provided data and routine demographic information **[17].**	The data from the training, validation, and test datasets were organized based on sex (male vs. female), age (grouped by decade), and race (Asian, Black, White, and Other) when it was possible **[19].**	A constrained optimization scheme (a mathematical framework) was proposed to incorporate measures of impartiality and health equity into the objectives of machine learning design **[19].**	The AI model was retrained with an additional constraint aimed at minimizing performance disparities across sex, race, and age simultaneously. This adjustment led to a small decrease in overall performance but significantly reduced bias **[19].**
Bhanot et al, 2021 **[26]** USA Informatics. Application of novel fairness metrics.	1) To ensure alignment between the distributions of real and synthetic data. 2) To generate synthetically data without significant deviations that could lead to discrimination against specific subgroups.	Data from three published synthetic research datasets: the American Time Use Survey (ATUS) dataset was the only primary care relevant among the three.	To generate synthetic sleep data for various age groups and *genders (*sex was identified as gender by the authors*) on the American Time Use Survey (ATUS) dataset.	Different protected attributes such as age, *gender (*sex was identified as gender by the authors*), and race.	Two fairness metrics were devised for synthetic data evaluation, examining all subgroups delineated by protected attributes to assess bias in the American Time Use Survey (ATUS) dataset.	Analysis using covariate-level disparity metrics showed potential discrepancies in the representativeness of synthetic data across both univariate and multivariate subgroup levels.
Fletcher et al, 2021 **[27]** USA	To applicate three principles (appropriateness, bias, and fairness)	Data from a simple set of diagnostic tools (e.g., a complete battery of pulmonary	To help predict the individual risk of several pulmonary	Age, Sex, and SES status measure or proxy: The medical care in rural India	The systematic bias was examined by testing the accuracy of the model using	The Allergic Rhinitis (AR) disease model showed increased stability with greater

Medical Ethics and Informatics. (Case study in Pune, India).	using a case of machine learning pulmonary disease diagnosis and screening.	function tests (PFT), which included spirometry, body plethysmography, etc).	diseases, and for the purpose of general practitioner doctors' decision support.	suffers from a low doctor-patient ration of 1:1700 which results in relatively high levels of underdiagnosis and misdiagnosis.	equal size homogenous training sets.	diversity in SES representation meaning that inclusivity in training data improved performance.
Foryciarz et al, 2022 **[28]** USA Informatics.	1) To measure the impact of two fairness methods, 2) To demonstrate principles applicable for conducting contextually pertinent fairness assessments of models used in clinical settings,	Data from an updated dataset from pooled cohorts focusing on 10-year Atherosclerotic Cardiovascular Disease (ASCVD) risk predictions was used.	The 10-year Atherosclerotic Cardiovascular Disease (ASCVD) risk predictions aim to inform a clinician-patient shared decision-making on initiating statin therapy.	Current ASCVD models do not inherently account for variations in race, ethnicity, and *gender-specific groups. *(sex was identified as gender by the authors).	Two algorithmic fairness strategies—group recalibration and equalized odds—was examined as means to refine risk estimations, ensuring they align with the underlying assumptions of the decision rules within the guidelines.	Compared with an unconstrained model, group-recalibration enhances calibration at specific thresholds for each group but exacerbates differences in false positive and false negative rates between groups. An equalized odds constraint, aimed at balancing error rates across groups, achieves this but misalign the overall model.
Ghai and Mueller, 2022 **[29]** USA Informatics. Case study.	To develop and propose a visual interactive tool (*D-BIAS*) that embodies human-in-the-loop AI approach for auditing and mitigating social biases from tabular datasets.	The Adult Income dataset was used as a tool for bias identification and mitigation. A random sample of 3000 points was used for faster computation.	The prediction task of this AI model is to classify if a person's income will be greater or lesser than $50k/year based on their personal attributes.	Age, work class, education, marital status, race, *gender (*sex was identified as gender by the authors*), and income.	The tool was evaluated by experimenting three datasets and a formal user study.	The tool significantly reduces bias compared to the baseline method across various fairness metrics with minimal data distortion and slight utility loss. The human-in-the-loop approach notably surpasses other methods in trust, interpretability, and accountability.
Hane and Wasserman, 2022	To implement practical tools aimed at	From a data set of 1,511,260 members in a commercial or	To predict next year's inpatient stays. Risk scores	Age, *gender (*sex was identified as gender by the*	Pragmatic tools were proposed to a fairer use of risk scores.	These tools aid stakeholders in seeing the suitable risk score

[30] USA Informatics.	facilitating the fairer utilization of risk scores in outreach programs.	Medicare Advantage plan from the "*Optum Labs Data Warehouse (OLDW)*".	aims to select which patients will receive aid and support.	*authors*), and race/ethnicity.	The method output charts allow users to select the optimal risk thresholds to trigger outreach.	thresholds for different patient groups to ensure Equality of Opportunity (EOp). They are applicable regardless of the machine learning or statistical model used in score generation.
Juhn et al, 2022 **[31]** USA Medical informatics. Case study.	To evaluate how disparities in data quality within electronic health records (EHRs) impact the varying performance of AI models across different socioeconomic status (SES) levels.	Data from a prior study included the training of two machine learning models. Variables, such as sociodemographic factors, risk factors, and asthma outcomes, were extracted from electronic health records (EHRs) over a preceding three-year period.	To estimate 1-year asthma exacerbation (AE) risk among children with asthma.	Authors focused to quantify bias in model performance by socioeconomic status (SES), and considered other readily available demographic characteristics (e.g., age, sex, and race/ethnicity), and pediatric chronic conditions.	The balanced error rate (BER) across various SES levels was compared and assessed using the HOUsing-based SocioEconomic Status measure (HOUSES) index, along with the incompleteness of EHR information on asthma care in relation to SES as a potential source of bias.	Children with asthma from lower SES backgrounds exhibited higher BER compared to those from higher SES backgrounds, with a notable ratio difference Also, they had a greater proportion of missing information relevant to asthma care, such as missing asthma severity and undiagnosed asthma despite meeting asthma criteria.
Khurshid et al, 2022 **[32]** USA Medical Informatics.	To develop a multi-institutional EHR cohort named "Community Care Cohort Project (C3PO)" with a focus on cardiovascular disease with two main objectives: (1) To mitigate ascertainment bias and (2) reduce data missingness.	Data from 520,868 individuals aged 18–90 who received regular primary care, with at least two visits within 1–3 consecutive years.	To predict more accurately cardiovascular disease risk with C3PO than for Convenience Samples Models such as Pooled cohort equations (PCE) or Cohorts for Heart and Aging Research in Genomic Epidemiology for Atrial Fibrillation (CHARGE-AF).	Potential racial discrimination examining four categories: (White Women, Black Women, White Men, Black Men) in C3PO versus Convenience Samples.	Development and implementation of a deep natural language processing (NLP) model by extracting four vital sign features from unstructured notes, and evaluation of effectiveness by comparing sample sizes before and after missing data recovery.	NLP helped find missing vital signs in EHR by 31%. C3PO risk models worked better compared to the Convenience Samples. By looking at patients who regularly visit their primary care doctor and using NLP to find missing information, prediction can be apply it to more people and therefore in fairness.

Martinez-Martin et al, 2021 **[33]** USA Medical Ethics. Delphi study.	To establish consensus statements on fundamental ethical principles guiding the use of digital phenotyping in mental health applications within the United States.	Digital phenotyping is studying behavior using data from digital devices. Data includes information collected in real-life settings. By monitoring things such as pulse rate or finger taps, or voice characteristics,	To assess behavior, physical health, and cognitive performance continuously. This helps in understanding someone's mental state or predicting their future actions.	Data streams may not adequately include people of different racial, socioeconomic, or disability status.	Delphi study was used to address ethical issues raised by mental health applications of digital phenotyping, such as privacy and data protection, consent, transparency, potential for bias in outcomes, and accountability.	This study revealed a consensus on ethical concerns concerning the development of mental health apps using digital phenotyping, including privacy, transparency, consent, accountability, and fairness.
Nong et al, 2022 **[34]** USA Management Care Ethics. Qualitative analysis of user-centered design (n=46) and expert interviews (n=10).	To (1) identify user requirements for informed decision-making and utilization of predictive models and (2) anticipate and reflect equity concerns in the information provided about models.	This is an exploratory study based in a single academic institution (interviews were conducted at a large medical institution with predictive analytics infrastructure).	Semi structured interviews were conducted to understand participants' needs, concerns, and preferences related to a hypothetical hepatitis C model prototype that predicted serious illness to allocate treatment resources.	All minority groups susceptible to disparities in treatment access, such as racial and ethnic minorities.	A user-centered design study at an academic medical center with clinicians and stakeholders to identify elements required for decision-making related to predictive models. Equity focused interviews with experts were also conducted.	Four elements were identified: details of (1) the model's developers and users, (2) the methodology employed, (3) the model's peer review processes and updates, and (4) the model's validation with different populations. Equity-related concerns raised in interviews with experts focused on the purpose or application of the model and its link to systemic inequalities.
Obermayer et al, 2019 **[35]** USA Data Science	To investigate the potential racial bias of a widely used algorithm that affects patients, particularly black patients, who when at risk scores similar to white patients, appear to have	Data from all primary care patients enrolled in risk-based contracts from 2013 to 2015 in a large academic hospital. The main sample consisted of 6079 patients who self-identified as Black and 43,539 patients who self-identified as	The stated aim of the algorithm studied was to predict individuals with higher health needs to offer interventions and resources to meet those needs.	Race: Black patients versus White patients. The algorithm was taking a large set of raw insurance claims data over a year. However, the algorithm specifically excluded race.	The existing model infrastructure (excluding race, as before), was used but the label was changed. Rather than future cost, an index variable that combined health prediction with cost prediction was created.	The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that less money is spending caring for Black patients than for White patients. Remedying this disparity would increase the percentage of Black patients receiving

	more severe health conditions.	White without another race or ethnicity.				additional help from 17.7 to 46.5%.
Panigutti et al, 2021 **[36]** Italy Medical Informatics. Use case.	To present a tool, named "FairLens", that can audit black-box model acting as a clinical decision support system (DSS).	Data from the MIMIC-IV (Medical Information Mart for Intensive Care) database. Although not primary care dataset focused, MIMIC may still be pertinent for primary care clinicians.	A fictional clinical decision support system (DSS). Most health systems used these programs considering effective at improving outcomes while reducing costs.	This tool presents patient data based on demographic attributes such as age, ethnicity, *sex (*identified as gender), and health insurance (SES proxy).	The reliability of this tool in discovering biases was tested through a fictional commercial biased black-box model named "*Doctor AI*".	"FairLens" could reveal biases injected into the fictitious DSS when other standard multi-label performance measures failed to detect them. Experts could explore a particular misclassification by identifying elements in the clinical history of patients in the groups concerned.
Park et al, 2022 **[20]** Singh and Long, 2018 **[18]** (added by handsearching). USA Informatics.	To 1) analyze the susceptibility of commonly used machine learning approaches for sex bias in mobile mental health assessment; 2) explore the use of an algorithmic disparate impact remover (DIR) approach to reduce bias levels while maintaining high accuracy **[20]**.	Using the data set (n=55) obtained in a previous study **[18]**, preprocessing and model training were carried out **[20]**. From a sample of 55 participants, 21 (38%) declared themselves to be women or female (minority class), and 34 (62%) described themselves as man or male **[18]**.	To detect and predict mental health problems (automated ML algorithms used the smartphone's characteristics, automatically classifying a person's level of mental health and general well-being with an accuracy of around 80% **[18]**.	Accuracy levels and differences in accuracy across **genders* (*sex was identified as gender by the authors*), were computed using five different machine learning models **[20]**.	Random forest model, which yielded the highest accuracy, was selected for a more detailed audit, and computed multiple metrics that are commonly used for fairness in the machine learning literature. Then, the disparate impact remover (DIR) approach was applied to reduce bias in the algorithm **[20]**.	The highest accuracy observed for mental health assessment was 78.57%. However, auditing based on *gender [sex as gender] revealed that performance was statistically higher for males than females. This disparity was considerably reduced after applying the DIR approach by adjusting the data used for modelling **[20]**.
Park et al, 2021 **[37]** USA Medical Informatics	To assess methods for mitigating bias in machine learning models within a practical clinical context.	Based on data from a cohort built using the "IBM MarketScan Medicaid database" (2014-2018), containing de-	Prediction of 2 binary outcomes: postpartum depression (PPD) and postpartum mental health	Binary race (black and white individuals): cohort study including 314,903 white pregnant women	A 3-fold approach: 1) Reweighing, a preprocessing method; 2) Prejudice Remover for logistic regression, an in-	A reweighing method was associated with a greater reduction in algorithmic bias for postpartum depression and mental health service

Cohort study.		identified claims records from around seven million Medicaid enrollees in several US states.	service utilization.	and 217,899 black pregnant women with Medicaid coverage.	processing method; and 3) Models training without the race variable for comparison, an approach known as fairness through unawareness.	utilization prediction between white and black pregnant women than simply excluding race from the prediction models.
Seker et al, 2022 **[38]** USA Medical Informatics.	To examine bias within electronic health record (EHR) data by: (1) measuring the level of discrimination, (2) addressing bias through re-balancing class labels, and (3) evaluating and comparing the modeling outcomes before and after processing.	An integrated dataset was created by adding zip code-level information to 19,367 electronic health records (EHRs) of patients diagnosed with chronic diseases (including asthma, diabetes, various heart conditions, stroke, etc.) sourced from the University of Arkansas for Medical Sciences Clinical Data Warehouse.	Patients are prioritized based on their risk level (multimorbidity) to ensure they receive care tailored to their specific needs. A simplified risk prediction model was developed to examine the effect of preprocessing data to mitigate bias on predictions.	Urban and rural residency status. (Place of residence) was identified as a potential bias against rural patients in algorithms using diagnostic data for risk assessment.	The level of discrimination favoring metropolitan patients in multimorbidity classification were measured in a model using the biased data to determine optimal classification performance. Then, a new training dataset and model, without bias, were developed and assessed against unchanged validation data.	The classification performance of the new model on unchanged data showed no significant deviation from the performance of the initial optimal model trained on biased data, indicating that bias can be effectively mitigated through preprocessing.
Straw and Wu, 2022 **[39]** UK Medical Informatics.	Following a literature review of the Indian Liver Patient Dataset (ILPD) publications: To replicate the machine learning (ML) models from these previous studies and subsequently examine them for bias.	The ILPD was originally collected from India and consists of 583 patient records, of which 416 have liver disease. The ILPD was imported from the "*UCI machine learning Repository, Irvine, Ca.*".	To diagnose liver diseases: Predictive machine learning models may benefit patient care if liver diseases can be diagnosed at an earlier stage.	Unisex biochemical thresholds used in research (based only on males' features) may disadvantage female patients in practice.	Four experiments trained on sex-unbalanced/balanced data, with and without feature selection. Random forests (RFs), support vector machines (SVMs), Gaussian Naïve Bayes and logistic regression (LR) classifiers were built, running experiments 100	Published models were reproduced achieving accuracies of >70% and demonstrated a previously unobserved performance disparity. Across all classifiers females suffer from a higher false negative rate (FNR). RF and LR classifiers are reported as the most effective models. However, in their experiments they

					times, and reporting average results with Standard deviation.	demonstrated the greatest false negative rate (FNR) disparity.
Yan et al, 2022 **[40]** USA Medical Informatics.	To illustrate the impact of observability on differential bias for a clinical prediction model (CPM).	Data of all discharges were collected in years 2018-2019 from an hospital EHRs. Development of a simple 30-day readmission CPM using age, length of stay, admission source, and five comorbidities.	Development of a simple 30-day readmission CPM using age, length of stay, admission source, and five comorbidities.	Place of residence: Each patient was categorized being local (within the hospital's county) or non-local.	The CPM performance was compared with and without this variable (of being local or not). Differences in observing an outcome that may induce differential bias were investigated.	There is not a meaningful difference between the local and non-local groups. Living locally only impacting the observability of the outcome (i.e., 30-day readmission) and cannot determine if there is a differential bias or not.

Diverse Groups Considered

The most studied protected attribute considered was Race (or Ethnicity) in 12 (out of 17) and Sex (using binary male vs female) in 10 (out of 17) studies. All studies are not differencing (biological) Sex and (socially construed) Gender, and five of them identified Sex as Gender. Race or Ethnicity was most identified as (White or Black), (Black or Non-Black) or, in one study as (Asian, Black, White, and Other).
The other protected attributes considered by studies are identified as follows: Age (7/17), Socioeconomic status or proxy - such as income, work class, education, healthcare insurance (5/17), Place of residence (2/17); Marital status (1/17), and Disability status (1/17).

Categorization of Bias Mitigation Strategies Deployed

We identified a heterogeneity in studies, which employed various strategies and methods to assess and mitigate bias in algorithms affecting diverse groups. We categorized these efforts into four groups: 1) addressing bias in existing AI models or datasets, 2) sourcing data from sources such as Electronic Health Records (EHRs), 3) developing tools that incorporate a 'human-in-the-loop' approach, and 4) identifying ethical principles to guide informed decision-making.

Attempts in Existing AI models or Datasets

We identified seven studies attempting to mitigate biases in existing AI models or Datasets **[19,27,28,35,20,37,39]**.
A debiasing attempt was performed on an insurance coverage algorithm aiming to identify who could benefit from health resources accordingly to their health needs **[35].** Risk scores of individuals were calculated based on future costs rather than uncontrolled or unmanaged illnesses indicators, putting Black patients at disadvantage. Changing data labeling to future illness rather than future costs, the percentage of Black patients that can benefit of health resources increased significantly **[35]**.
Another cohort study **[37]** using a Medicaid enrollees' dataset also showed that reweighing was associated with greater reduction in bias for postpartum depression risk scores between White and Black than training without the race variable for comparison. Initially, it was found that the White race had higher rates of postpartum depression and mental health service utilization. However, after comparing rates of postpartum depression between races based on population surveys, it emerged that the higher rates in White women could be due to disparities in the timely assessment, screening, and detection of symptoms in Black women **[37]**.

Three other studies showed that 1) retraining models data in incorporating health equity measures resulted in a slight decrease in models' performance for detecting abnormal ECG, but significantly reduced gender, race and age biases **[19]**; 2) adding greater diversity on the training data of a predictive pulmonary diseases model improved its performance **[27]** and; 3) although achieving high accuracy for assessing mental health, another model performance was statistically higher and more accurate for Men than for Women **[18]**. The use of an algorithmic disparate remover, by adjusting the modeling data, significantly reduced this disparity while maintaining high accuracy **[20]**.

Another attempt to assess bias was initiated by replicating models predicting liver disease **[39]**. Importing an existing data set reproduced predictive models with high accuracy but revealed a previously unobserved bias toward women having higher false negative rate (FNR).

We identified only one in-processing debias attempt **[28]**. Two algorithmic fairness strategies – group recalibration and equalized odds – were used to recalibrate a predictive model of cardiovascular diseases not initially adjusted for attributes such as sex or race, resulting in an exacerbation of false positive and negative rates differences between groups or in an overall model miscalibration.

Attempts in Sourcing Data

We identified five studies attempting to mitigate biases in sourcing data **[26,31,32,38,40]**.
Based on published synthetic datasets, such as the American Time Use Survey dataset analysis; using fairness metrics showed potential discrepancies in the representativeness between real and synthetic data across age, sex, and race **[26]**.
Four other studies investigated electronic health records (EHRs) datasets **[31,32,38,40]**. A natural language processing (NLP) model using vital sign features extraction from unstructured notes was developed comparing risk scores with two convenience samples. This method reduced the missingness of vital signs by 31%, and therefore possible discrimination toward diverse groups, such as Black Men or Black Women **[32]**. Based on data from a previous study, two machine learning models were trained comparing balanced error rate (BER) against different socioeconomic status levels and incompleteness of electronic health records data **[31]**. Asthmatic children with lower socioeconomic status had larger BER than those with higher socioeconomic status, and had higher missing information of asthma care, severity, or undiagnosed asthma despite they are meeting asthma criteria **[31]**.
Potential place of residence bias based on electronic health records was examined by two studies **[38,40]**. Re-balancing class labels, by adding information (zip-code level) to 19,367 electronic health records in the preprocessing step showed no significant deviation relative to performance indicating that bias can be mitigated through preprocessing **[38]**. Meanwhile, a simple 30-day readmission prediction was developed in which each patient was categorized as local (near) or not (far) **[40]**. The performance with and without this variable were assessed without significant differences. Considering living locally only impacting the observability of the outcome (e.g., a patient can be readmitted in another hospital), differential bias assessment cannot be uniquely on observed data **[40]**.

Attempts in Developing Tools with “Human-in-the-loop”

We identified three studies attempting to mitigate biases in bringing « human-in-the-loop **[29,30,36]**.

These studies led to the development of "human-in-the-loop" tools: 1) an interactive visual tool for auditing and mitigating bias from tabular datasets, which was tested by running experiments on three datasets and with user participation, significantly reduced bias compared to another commercial debiasing toolkit [29]; 2) pragmatic tools have been developed for better use of risk scores with a Medicare members’ dataset, while decision users can use these tools to identify appropriate risk scores for each subgroup in order to achieve equality of opportunity [30], and; 3) a tool called “FairLens” capable of identifying and explaining biases has been tested using a fictitious black box model serving as a decision support system (DSS) [36]. Empirically validated by injecting biases into this fictitious DSS, this tool outperformed other standard measures and enabled experts to see problematic groups or affected patients allowing for possible misclassification [36].

Attempts at Identifying Ethical Principles for Informed Decision-making

We identified two empirical studies attempting to mitigate biases by identifying ethical principles for informed decision-making **[33,34]**.
To assess the possible EHRs data missingness from phenotyping technology, a Delphi study was conducted to address ethical challenges reaching a consensus about the importance of privacy, transparency, consent, accountability, and fairness **[33]**. A user-centered design study was also conducted to identify user requirements, mainly destinated to health managers and clinicians, for informed decision-making and confidence using a hepatitis C severity illness predictive model prototype **[34]**.

Discussion

Principal Findings

The reviewed studies illustrate a multifaceted approach to mitigating bias in primary care AI models. Strategies include retraining, reweighing, relabeling, adding more diversity, and trying to replicate existing modeling data **[19,20,27,35,37,39]**, or an algorithmic recalibration applied to an existing prediction model **[28]**. Other strategies include development and application of fairness metrics to ensure equitable distributions in formerly published databases **[26]**, or identification of missingness in EHRs datasets, by re-balancing class labels or adding information **[31,32,38]**. Another group of strategies include the introduction of visual interactive tools for human-in-the-loop bias auditing **[29,30,36]**. All these attempts cover a broad spectrum of interventions from data preprocessing, algorithmic modification, to post-hoc analysis, demonstrating the complexity and variety of approaches needed to address bias in AI models in primary healthcare.
The studies collectively address a wide range of protected attributes **[1,8]**, including Race or Ethnicity **[19,26,28-37]**, Sex **[19,20,26-31,36,39]**, Age **[19,26,27,29,30,31,36]**, Socioeconomic status (SES) **[27,29,31,33,36]**, and other demographic variables such as Place of residence **[38,40]**. It underlines the recognition of the multifaceted nature of bias, which can intersect across various dimensions of identity and social determinants of health **[41,42]**. However, we have

identified disparities in the number of protected attributes studied. Race (White versus Black) and Sex (Male versus Female) are mostly investigated whereas other attributes, such as disability and gender are under-investigated or not at all.

Bias mitigation efforts revealed a nuanced landscape where efforts to reduce bias across protected attributes can result in complex trade-offs with model performance. For example, a decrease in overall model performance but significant reductions in bias following the implementation of constrained optimization was observed **[19]**. Similarly, improvements in calibration for specific groups at the cost of increased disparities in false positive and false negative rates between groups was reported **[28]**. Despite these trade-offs, the efforts are largely successful in reducing bias, as evidenced by a study **[26]** achieving fairer distributions in synthetic data, and in another **[29]** where human-in-the-loop interventions significantly reduced bias while maintaining utility.

These empirical findings reinforce theoretical insights emphasizing the importance of health equity between protected and unprotected attributes **[1,8]**. To mitigate bias in AI health models, distributive justice options for machine learning were proposed: 1) Equal patient outcomes; 2) Equal performance, and 3) Equal allocation of resources **[1]**. As these different types of fairness options are often incompatible, it seems difficult to optimize all these parameters, as an identified study demonstrate **[28]**. Trade-offs are essential, and participatory process with key stakeholders, including ethicists, clinicians, and marginalized populations is strongly encouraged **[1]**.

Comparison to Prior Work

Initiatives on the fair use of AI in healthcare, or on assessing the risk of bias in AI predictive models, have been published in recent years. Notable initiatives include CONSORT-AI and SPIRIT-AI **[43]**, which recommend guidelines for the ethical presentation of the results of trials conducted with AI in the healthcare field. To assess the risk of bias in diagnostic and prognostic prediction models studies, the "Prediction model Risk Of Bias ASsessment Tool" (PROBAST) **[44]** can be used as a list of signaling questions grouped into four categories: participants, predictors, outcomes and analysis. This tool, PROBAST, was used in a systematic scoping review to assess the quality of primary studies reporting applications of AI in community-based primary health care **[44]**.

However, the objective of our scoping review differs and is not to identify biases in the AI prediction models themselves, but rather biases toward groups sub or misrepresented in these AI models. An identified review has used and adapted PROBAST to assess protected attributes-related bias, but the AI predictive models studied were hospital-based and not primary care relevant **[46]**. We also identified a scoping review protocol that looked at bias toward diverse groups in AI systems in primary care, but unless we are mistaken, the results of this protocol have never been published **[47]**. Another identified review objective was to assess age-related bias in AI but without focusing on primary healthcare **[48]**. Finally, we identified another systematic review investigating health inequities in primary care but adopting a system-wide perspective such as patient consultation or effects on health systems **[49]**.

To our knowledge, no other published review has the objectives of identifying 1) the bias mitigation strategies or methods in primary healthcare 2) the diverse groups sub or misrepresented concerned and 3) the results on bias mitigation and AI models performance.

Strengths and Limitations

Strengths of this review are results that can be converted into recommendations guiding multiple stakeholders, such as AI developers, researchers, and decision makers. However, we acknowledge some limitations. First, we limited our search strategy to the last 5 years prior to

November 2022 and to 4 databases. Some relevant studies might not be included. Second, studies extraction and quality assessment were done once but all of them were validated by at least one senior researcher. Third, due to the heterogeneity of studies we were not able to combine results through a quantitative synthesis and stayed at a narrative level of reporting. Finally, our study mainly identified studies from North American setting, reducing transferability to another continent.

Future Directions and Dissemination Plan

This scoping review is the first step of the "Protecting and Engaging Vulnerable Populations in the Development of Predictive Models in Primary Health Care for Inclusive, Diverse and Equitable AI" (PREMIA) project. Based on results presented in this review, a group of clinicians, managers (from the Public Health Department), primary care researchers and data scientists will evaluate an existing AI predictive model using the bias mitigation strategies identified from this review. We will recruit patient and citizen partners, self-declared with a protected attribute or characteristic, to invite them to participate in this working group. Diverse groups, such as ageing, disabled, diverse races or ethnicities, will actively be consulted in the second step of PREMIA.
Indigenous peoples in Canada represent a group historically sub represented in health research, leading to inequities **[3]**. As no other study considered bias related to indigenous status, we will work with Indigenous representatives to propose methods to mitigate this bias in primary health care algorithms.
While striving to create ethically robust AI models, there is often a tension in selected studies as efforts to reduce bias can sometimes lead to a decrease in the model's overall performance. This presents a critical challenge of balancing the imperative for fairness with the need to maintain high accuracy and efficiency in algorithmic outputs.

Ethics Approval

We obtained approval by the Ethics board of the "Comité d'éthique de la recherche sectoriel en Santé des Populations et Première Ligne du CIUSSS de la Capitale-Nationale" for the Protecting and Engaging Vulnerable Populations in the Development of Predictive Models in Primary Health Care for Inclusive, Diverse and Equitable AI (PREMIA) project (#2023-2726).

Conclusion

This review identifies strategies and methods for mitigating bias in primary healthcare algorithms, considers diverse groups based on their personal or protected attributes, and examines the results on bias attenuation and model performance. Results suggest that biases toward diverse groups can be more effectively mitigated when data are open-sourced, multiple stakeholders are involved, and during the preprocessing stage of algorithm development. More empirical studies are needed, with a focus on including participants embracing more diversity, such as nonbinary gender identities or Indigenous peoples in Canada.

Acknowledgments

The Protecting and Engaging Vulnerable Populations in the Development of Predictive Models in Primary Health Care for Inclusive, Diverse and Equitable AI (PREMIA) project is funded by the International Observatory on the Societal Impacts of AI and Digital Technology (OBVIA).
The authors would like to thank Karine Gentelet for her contribution to design the study and

obtained the funding.

Data Availability

The database search strategies are available in [**Multimedia Appendix 1**]. The PRISMA Extension for Scoping Reviews (PRISMA-ScR) checklist was included in the results manuscript submission as a supplementary file. Databases created in the data collection and data extraction processes can be provided upon reasonable request.

Authors' Contribution

Sasseville, Gagnon, Rhéaume, Couture, Després, Paquette, and Darmon designed the study and obtained the funding. Sasseville, Gagnon, Ouellet and Bergeron designed the search strategy. Sasseville, Ouellet, Sahlia, Rhéaume, Gagnon, Couture, and Bergeron participated in the sources screening. Sasseville, Ouellet, and Gagnon did the data extraction. Ouellet, Sasseville and Gagnon completed a first draft of the manuscript and all authors participated in the revising and editing of the manuscript versions. All authors reviewed and approved this manuscript.

Conflict of Interest

The authors have no conflict of interest to declare.

Multimedia Appendix 1

The Databases search strategies

Multimedia Appendix 2

Quality assessment/The MMAT scores

References

1. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med*. 2018;169(12):866-872. doi:10.7326/M18-1990
2. Qu Y, Wei C, Du P, et al. Integration of cognitive tasks into artificial general intelligence test for large models. *iScience*. 2024;27(4):109550. Published 2024 Mar 22. doi:10.1016/j.isci.2024.109550
3. Gurevich E, El Hassan B, El Morr C. Equity within AI systems: What can health leaders expect? *Healthc Manage Forum*. 2023;36(2):119-124. doi:10.1177/08404704221125368
4. Alabdulmohsin I, Lucic M. A near-optimal algorithm for debiasing trained machine learning models. August 23 2022. https://doi.org/10.48550/arXiv. 2106.12887. Published online.
5. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr Radiol. Oct 2022;52(11):2087-2093.
6. Kang J, Hanif M, Mirza E, Khan MA, Malik M. Machine learning in primary care: potential to improve public health. *J Med Eng Technol*. 2021;45(1):75-80. doi:10.1080/03091902.2020.1853839
7. Community-based primary health care. Canadian Institutes of Health Research. URL: https://cihr-irsc.gc.ca/e/43626.html [accessed 2024-04-01]
8. PROGRESS-Plus. Cochrane Methods. URL: https://methods.cochrane.org/equity/projects/evidence-equity/progress-plus [accessed 2024-04-01]
9. Delgado J, de Manuel A, Parra I, et al. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J Bioeth Inq*. 2022;19(3):407-419. doi:10.1007/s11673-022-10200-z
10. Wang JX, Somani S, Chen JH, Murray S, Sarkar U. Health equity in artificial intelligence and primary care research: protocol for a scoping review. JMIR Res Protoc. Sep 17, 2021;10(9):e27799.
11. Wang H, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc. Jul 12, 2022;29(8):1323-1333.
12. Sasseville M, Ouellet S, Rhéaume C, et al. Risk of Bias Mitigation for Vulnerable and Diverse Groups in Community-Based Primary Health Care Artificial Intelligence Models: Protocol for a Rapid Review. *JMIR Res Protoc*. 2023;12:e46684. Published 2023 Jun 26. doi:10.2196/46684
13. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth*. 2020;18(10):2119-2126. doi:10.11124/JBIES-20-00167
14. Apply PCC. University of South Australia. URL: https://guides.library.unisa.edu.au/ScopingReviews/ApplyPCC [accessed 2024-04-01]
15. Veritas Health Innovation. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-01]
16. Handsearching. Cochrane. URL: https://training.cochrane.org/resource/tsc-induction-mentoring-training-guide/5-handsearching [accessed 2024-04-01]

17. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021 v1.0.3. physionet.org. Accessed January 11, 2024. https://physionet.org/content/challenge-2021/1.0.3/
18. Singh VK, Long T. Automatic assessment of mental health using phone metadata. Proc Assoc Info Sci Technol 2018;55(1):450-459 [FREE Full text [doi:10.1002/pra2.2018.14505501049]
19. Perez Alday EA, Rad AB, Reyna MA, et al. Age, sex and race bias in automated arrhythmia detectors. J Electrocardiol. 2022; 74:5-9. doi: 10.1016/j.jelectrocard.2022.07.007
20. Park J, Arunachalam R, Silenzio V, Singh VK. Fairness in Mobile Phone-Based Mental Health Assessment Algorithms: Exploratory Study. JMIR Form Res. 2022;6(6):e34366. Published 2022 Jun 14. doi:10.2196/34366
21. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27(12):2011-2015. doi:10.1093/jamia/ocaa088
22. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf. 1 janv 2018;34(4):285-91.
23. Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, et al. Mixed Methods Appraisal Tool (MMAT) version 2018: user guide. Mixed Methods Appraisal Tool. 2018. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf [accessed 2024-04-01]
24. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. *JBI Evid Synth*. 2023;21(3):520-532. Published 2023 Mar 1. doi:10.11124/JBIES-22-00123
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473. doi:10.7326/M18-0850
26. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The Problem of Fairness in Synthetic Healthcare Data. Entropy (Basel). 2021;23(9):1165. Published 2021 Sep 4. doi:10.3390/e23091165
27. Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. Front Artif Intell. 2021; 3:561802. Published 2021 Apr 15. doi:10.3389/frai.2020.561802
28. Foryciarz A, Pfohl SR, Patel B, Shah N. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. BMJ Health Care Inform. 2022;29(1):e100460. doi:10.1136/bmjhci-2021-100460
29. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias | IEEE Journals & Magazine | IEEE Xplore. ieeexplore.ieee.org. Accessed January 11, 2024. https://ieeexplore.ieee.org/document/9903601/authors#authors
30. Hane CA, Wasserman M. Designing Equitable Health Care Outreach Programs From Machine Learning Patient Risk Scores. Med Care Res Rev. 2023;80(2):216-227. doi:10.1177/10775587221098831
31. Juhn YJ, Ryu E, Wi CI, et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. J Am Med Inform Assoc. 2022;29(7):1142-1151. doi:10.1093/jamia/ocac052

32. Khurshid S, Reeder C, Harrington LX, et al. Cohort design and natural language processing to reduce bias in electronic health records research. NPJ Digit Med. 2022;5(1):47. Published 2022 Apr 8. doi:10.1038/s41746-022-00590-0
33. Martinez-Martin N, Greely HT, Cho MK. Ethical Development of Digital Phenotyping Tools for Mental Health Applications: Delphi Study. JMIR Mhealth Uhealth. 2021;9(7): e27343. Published 2021 Jul 28. doi:10.2196/27343
34. Nong P, Raj M, Platt J. Integrating predictive models into care: facilitating informed decision-making and communicating equity issues. Am J Manag Care. 2022;28(1):18-24. doi:10.37765/ajmc.2022.88812
35. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. doi:10.1126/science.aax2342
36. Panigutti C, Perotti A, Panisson A, Bajardi P, Pedreschi D. FairLens: Auditing black-box clinical decision support systems. Information Processing & Management. 2021;58(5):102657. doi:10.1016/j.ipm.2021.102657
37. Park Y, Hu J, Singh M, et al. Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. JAMA Netw Open. 2021;4(4): e213909. Published 2021 Apr 1. doi:10.1001/jamanetworkopen.2021.3909
38. Seker E, Talburt JR, Greer ML. Preprocessing to Address Bias in Healthcare Data. Stud Health Technol Inform. 2022; 294:327-331. doi:10.3233/SHTI220468
39. Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health Care Inform [Internet] 2022; 29. doi:10.1136/bmjhci-2021-100457.
40. Yan M, Pencina MJ, Boulware LE, Goldstein BA. Observability and its impact on differential bias for clinical prediction models. J Am Med Inform Assoc. 2022;29(5):937–43. doi:10.1093/jamia/ocac019
41. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health*. 2023;2(6): e0000278. Published 2023 Jun 22. doi: 10.1371/journal.pdig.0000278
42. Delgado J, de Manuel A, Parra I, et al. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J Bioeth Inq*. 2022;19(3):407-419. doi:10.1007/s11673-022-10200-z
43. Liu, X., Cruz Rivera, S., Moher, D. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 26, 1364–1374 (2020). https://doi.org/10.1038/s41591-020-1034-x
44. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
45. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res. Sep 03, 2021;23(9): e29839
46. Wang H, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc. Jul 12, 2022;29(8):1323-1333
47. Wang JX, Somani S, Chen JH, Murray S, Sarkar U. Health equity in artificial intelligence and primary care research: protocol for a scoping review. JMIR Res Protoc. Sep 17,

2021;10(9): e27799.
48. Chu, C.H., Donato-Woodger, S., Khan, S.S. *et al.* Age-related bias and artificial intelligence: a scoping review. *Humanit Soc Sci Commun* **10**, 510 (2023). https://doi.org/10.1057/s41599-023-01999-y
49. Rodgers S, et al. Artificial intelligence and health inequities in primary care: a systematic scoping review and framework. Fam Med Com Health 2022;10: e001670. doi:10.1136/fmch-2022-001670

Abbreviations

AI	artificial intelligence
BER	balanced error rate
CBPHC	community-based primary health care
CIUSSS	centre intégré universitaire de santé et de services sociaux
CPM	clinical prediction model
DIR	disparate impact remover
ECG	electrocardiogram
EHR	electronic health record
JBI	Joanna Briggs Institute
ML	machine learning
MMAT	Mixed-Methods Appraisal Tool
NLP	natural language processing
PCC	Population [or Participant], Concept, and Context
PICO	Participant, Intervention, Comparator, and Outcome
PREMIA	Protecting and Engaging Vulnerable Populations in the Development of Predictive Models in Primary Health Care for Inclusive, Diverse and Equitable AI
PRISMA-ScR	Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PROGRESS	place of residence, race/ethnicity/culture/language, occupation, gender/sex, religion, education, socioeconomic status, and social capital
SES	socioeconomic status

Supplementary Files

Multimedia Appendixes

The Databases search strategies.
URL: http://asset.jmir.pub/assets/75e739adf9db8d7d1b19efdd48913675.docx

Quality assessment/The MMAT scores.
URL: http://asset.jmir.pub/assets/6ed7278a7706134e12aceb4bff00f1ef.xlsx