Reporting of Fairness Metrics in Clinical Risk Prediction Models: A Call for Change

Running title: Reporting Fairness in Clinical Risk Prediction Models

Word Count: 3300 words

Author List: Lillian Rountree, BA; Yi-Ting Lin, ScM; Chuyu Liu, BS; Maxwell Salvatore, MPH; Andrew Admon, MD, MPH, MSc; Brahmajee K Nallamothu, MD, MPH; Karandeep Singh, MD, MMSc; Anirban Basu, PhD; Bhramar Mukherjee, PhD

Affiliations:

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (L.R.)

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (Y.L.)

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (C.L.)

Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan (M.S.)

Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan; Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan; and VA Center for Clinical Management Research VA Ann Arbor Healthcare System, Ann Arbor (A.A.)

Institute for Healthcare Policy & Innovation, University of Michigan, and Veterans' Affairs Center for Clinical Management Research, University of Michigan, and Michigan Center for Health Analytics and Medical Prediction, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor (B.K.N.)

Department of Medicine, University of California San Diego School of Medicine, San Diego, California (K.S.)

The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, School of Pharmacy, the Department of Health Systems and Population Health, and the Department of Economics, University of Washington, Seattle (A.B.)

Department of Biostatistics, School of Public Health, University of Michigan, Department of Epidemiology, School of Public Health, University of Michigan Ann Arbor, Michigan, USA (B.M.)

Corresponding Author: Bhramar Mukherjee, PhD, Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA (bhramar@umich.edu).

Abstract

Clinical risk prediction models integrated in digitized healthcare systems hold promise for personalized primary prevention and care. Fairness metrics are important tools for evaluating potential disparities across sensitive features in the field of prediction modeling. In this paper, we seek to assess the uptake of fairness metrics in clinical risk prediction modeling by conducting a scoping literature review of recent high impact publications in the areas of cardiovascular disease and COVID-19. Our review shows that fairness metrics have rarely been used in clinical risk prediction modeling despite their ability to identify inequality and flag potential discrimination. We also find that the data used in clinical risk prediction models remain largely demographically homogeneous, demonstrating an urgent need for collecting and using data from diverse populations. To address these issues, we suggest specific strategies for increasing the use of fairness metrics while developing clinical risk prediction models.

Introduction and Background

Prediction models are increasingly prevalent in research and decision making across a wide swath of fields, from finance and criminal justice to public health and healthcare (1,2). However, potential statistical and historical biases ingrained in these models can impact their accuracy and ethical viability, particularly for populations at risk of discrimination due to sensitive features like race/ethnicity, age, and sex (1,3–5). Are these seemingly objective and data-driven prediction models furthering existing inequities? As prediction models often inform who receives an intervention (e.g., a loan, a release on bail, a healthcare treatment) or identify who is at a higher disease risk when designing targeted prevention strategies, it is critical that they are fair, providing both accurate and nondiscriminatory predictions.

Algorithmic fairness is closely related to but distinct from algorithmic bias, another concern when assessing model performance (6). Algorithmic bias refers to the systematic and unfair discrimination that can occur when algorithmic models perpetuate existing biases present in the data they are trained on or the way they are designed. This bias can manifest in various ways, such as favoring one group over another due to race/ethnicity, sex, age, or other sensitive characteristics, or reinforcing stereotypes present in the training data. Algorithmic fairness, on the other hand, is the goal of designing algorithms and artificial intelligence models in a way that minimizes or mitigates bias and ensures fair treatment for all individuals or groups affected by

the model (2,4,6,7). Achieving algorithmic fairness often requires careful consideration of the design, development, and deployment of models, including the selection of appropriate training data, the use of fairness-aware models, and the incorporation of fairness metrics to evaluate the performance of the model.

In recent years, several metrics have been introduced to evaluate the fairness of prediction models (1,8), alongside various packages and toolboxes for their implementation (see (9–12)). These metrics differ from common prediction metrics of discrimination and calibration, which measure overall model performance (6). Some of the most cited and used fairness metrics are enumerated in Table 1, all of which assess differences in predictions ($\hat{Y}$) for a binary decision outcome Y (e.g., treatment or no treatment, disease, or no disease) for different values of a sensitive variable S ($S=a$ or $S=b$). These metrics give a numerical sense of the differences in predicted model outcomes across different demographic groups such as race/ethnicity, age, and sex, and thus how fair or unfair a model may be. The common practice is to use these metrics to provide only point estimates without associated uncertainty quantifications; while fair inferential methods have been proposed (13), statistical inference and interval estimation is not yet widely adopted in fairness research.

Metric	**Equation**	**Definition**
Equalized Odds	$P(\hat{Y}=1 \mid S=a, Y=y) = P(\hat{Y}=1 \mid S=b, Y=y)$	Both groups should have equal true positive and false positive rates.
Equal Opportunity	$P(\hat{Y}=1 \mid S=a, Y=1) = P(\hat{Y}=1 \mid S=b, Y=1)$	Both groups should have equal true positive rates. A relaxed version of equalized odds.
Predictive Equality	$P(\hat{Y}=1 \mid Y=0, S=a) = P(\hat{Y}=1 \mid Y=0, S=b)$	The rate of false positives (negative events categorized as positives) should be independent of the sensitive feature.
False Negative Rate Parity	$P(\hat{Y}=0 \mid Y=1, S=a) = P(\hat{Y}=0 \mid Y=1, S=b)$	The rate of false negatives (positive events categorized as negatives) should be independent of the sensitive feature.
Predictive Parity	$P(Y=1 \mid \hat{Y}=1, S=a) = P(Y=1 \mid \hat{Y}=1, S=b)$	Model precision should be the same for both groups.
Demographic Parity	$P(\hat{Y}=1 \mid S=a) = P(\hat{Y}=1 \mid S=b)$	The prediction or decision should be independent of the sensitive feature.

Table 1. Commonly used fairness metrics, adapted from (1,8). Y-hat is our prediction/decision, Y is our observed data, and S is our observed feature, in the case of S being a multi-group variable where a comparison with a reference or privileged group is meaningful. Y-hat and Y are binary variables.

Though quite simple, such metrics can shed light on otherwise unseen disparities in a model or source dataset. No metrics, however, are without their limitations or challenges. For these metrics to produce meaningful results, a clear interpretation of the predictor variables used in a model is required in addition to how they affect the outcome of interest. For some variables, this is straightforward. For the sensitive features that fairness metrics usually seek to assess, however, it can be much more complex. Sensitive features often have multiple definitions and interpretations, particularly when it comes to what is biological versus what is socially constructed. This ambiguity is particularly important for sex and race/ethnicity, two of the most used variables in outcome regression models. Sex is a biological variable, but it will likely include impacts caused by individual's gender—information that might be the actual cause behind the sex variable's recorded influence on an outcome of interest. In the case of race and ethnicity, these variables' self-reported—and thus exclusively social nature—is not always made clear, allowing for the inaccurate and harmful conclusion that differences observed with these race and/or ethnicity variables have a biological basis. Table 2 offers definitions of different measures related to these two specific sensitive variables and highlights how interpretations can be conflated and may not align with intended use.

Sensitive variable	**Definition**
Social	Variables based on an individual's lived experience. Frequently self-reported.
Race	A social construct based on perceived physical differences (13–15) that often acts as a proxy for various social and health consequences of racism in modeling (16). Often paired with ethnicity, though they are not strictly equivalent. Includes no biological information.
Ethnicity	A social construct based on shared culture, language, geography, religion, and history (14,17). Often paired with race, though they are not strictly equivalent. Includes no biological information.
Gender	A social construct referring to an individual's self-presentation in society, informed by culture, psychology, and society (18). Though distinct from sex, it often indirectly captures sex's impact.
Biological	Variables based on an individual's biology.
Genetic ancestry	Genetic similarities between people due to common ancestors (17). Distinct from race. Strictly biological.
Sex	The biological factors used to categorize individuals as male, female, or intersex (19). Since 2016, it has been a required covariate in NIH-funded research (18). Though distinct from gender, it often indirectly captures gender's impact.

Table 2. Definitions of commonly used sensitive variables. Note that the definition of these variables is frequently left ambiguous in practice, raising the possibility for harmfully conflating the social and the biological.

The risk of including sensitive variables only to misinterpret them in a way that furthers existing inequities has led to an ongoing debate about the place of these sensitive features in clinical disease risk prediction models. Exclusion of such sensitive features is one possible solution (one that has been particularly brought up regarding self-reported race (21,22)); the use of fairness metrics is another, allowing for the quantification of possible harm or help a sensitive variable might bring within the model. As prediction models are increasingly embedded into electronic health records and become more and more important for precision medicine (23,24), determining the place of sensitive features in prediction modeling is a crucial issue. Debates around best practice are ongoing, particularly regarding the inclusion of a race variable. Many other excellent papers explore the complexities of the issue at length (see (1,3,4,6,17,25)).

While assessing fairness metrics may be a key mechanism to ensure that models are equitable, how widely they are used or reported in the clinical risk prediction literature is not always objectively quantified. Thus, we sought to examine the usage of fairness metrics in clinical risk prediction research through a scoping review of recently published risk prediction models in high-impact journals for two diseases: cardiovascular disease ("CVD," a long-studied chronic and non-communicable disease) and COVID-19 (a newly emerged infectious disease). We hypothesized that there would be little reporting of fairness metrics in CVD research, where many studies span years and began long before discussions of fairness metrics, but that the emergence of COVID-19 would pose an opportunity to more frequently incorporate modern advances in fairness metrics into predicting disease outcomes.

Methods

A literature review was conducted for each of the two diseases of interest, CVD and COVID-19. Our outcomes of interest differed slightly between CVD and COVID-19: the clinical risk prediction models for CVD focused on fatal and non-fatal risk of CVD, while the models for COVID-19 focused on both risk of mortality and risk of severe disease (see Figure 1 for specific search terms). We did not use a classic systematic literature review approach, as our priority was to expediently capture only the highest impact papers. Figure 1 details the reproducible steps taken for data collection. This review began with searches on PubMed conducted in November 2023 to gain a high-level understanding of the use of sensitive features in risk prediction for CVD and COVID-19. Google Scholar was then used for actual article collection, focusing on

extensively cited publications in high impact journals.

The criteria for highly impactful publications differed between CVD and COVID-19, reflecting the difference between a long-studied disease and an emerging area of study: CVD papers from the last ten years (2013–2023) were reviewed and selected if they exceeded 100 citations and were from journals with impact factors exceeding 5, while COVID-19 papers from 2021 to 2023 were reviewed and selected if they exceeded 50 citations for 2021 papers or exceeded 10 citations for 2022 to 2023 papers, both from journals with impact factors over 5. Systematic reviews and meta-analysis papers were excluded from the search results. Figure 1 is a flow diagram representing the search process.

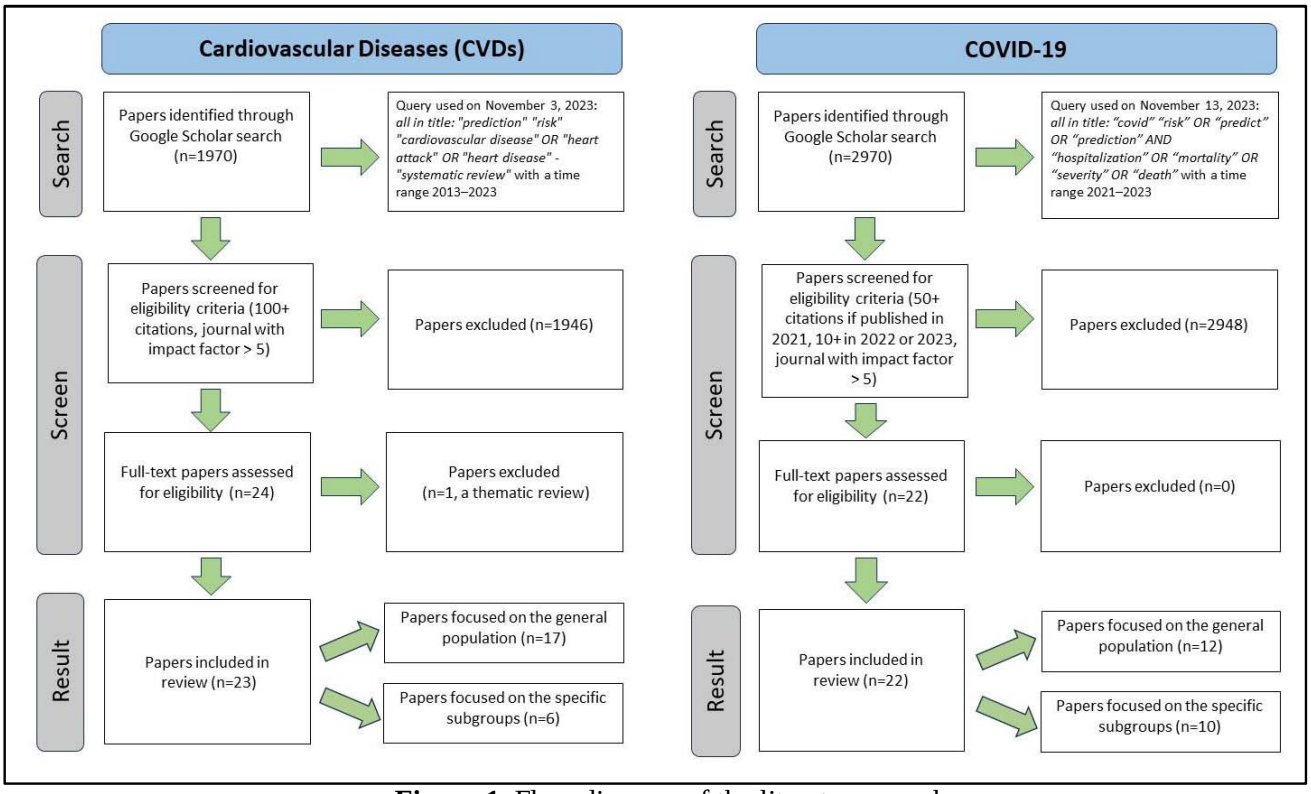


Figure 1. Flow diagram of the literature search process.

Results

CVD

A PubMed search query *("cardiovascular disease"[All Fields] OR "heart disease"[All Fields] OR "heart attack"[All Fields]) AND "prediction"[All Fields] AND "risk"[All Fields] AND (y_10[Filter])* conducted on November 13, 2023, returned 5,107 results. Further specification with the term “sex” returned 817 results (16% of papers). The addition of the term

"race" (but not "sex") returned 145 results (2.8% of papers); specifying only one race/ethnicity returned 145 results ("Black"), 56 results ("Hispanic"), and 186 results ("Asian"). The Google Scholar search query: *allintitle: "prediction" "risk" "cardiovascular disease" OR "heart attack" OR "heart disease" OR "mortality" OR "death" -"systematic review"*, with a time range of 2013–2023 on November 3, 2023, returned 1970 results, 1000 of which were accessible for review. The yielded results were selected if they exceeded 100 citations and were from journals with impact factors exceeding 5. This provided a shortlist of 23 articles detailing models predicting the risk of a fatal or non-fatal CVD event. These 23 articles were then divided into groups based on their target population (general population or a specific subpopulation). Sections S.1 and S.2 of the supplementary material provide additional details.

Of the 17 CVD papers focusing on a general population that met the criteria of this review (supplementary material section S.1), none discussed fairness metrics. Of these 17 papers, five (29%) stratified their models by sex, (i.e., built different models for each sex), and 11 (65%) included sex as a covariate. Nine of the 17 (53%) included data on race/ethnicity. As many of the papers paired race and ethnicity together or used them interchangeably, we will refer to any discussion of either as race/ethnicity jointly, despite this being an imprecise practice. Seven papers included race/ethnicity by self-reporting, and two by genetic ancestry (in genetically homogeneous populations). The eight studies that did not include race/ethnicity data were all based in the United States and Northern or Western Europe. Of the nine papers with recorded race/ethnicity data, five (55%) were multiracial/multiethnic (more than one racial/ethnic group identified). Four of the five multiracial/multiethnic studies included race/ethnicity as a covariate; no study stratified its model by race/ethnicity (see Figure 2). Other sensitive features considered in the studies include a covariate for area-based measures of deprivation (26), a covariate for body mass index (27), and stratification by risk region, a grouping of countries in Europe by their CVD mortality rates (28).

Similar results were observed in the six CVD papers that focused on specific subpopulations (supplementary material section S.2). Subpopulations considered in these studies range from those with chronic conditions (29–32) to those from specific ethnic (33) or age groups (34). Three of the six studies (50%) included sex as a covariate; only one stratified its model by sex (see Figure 2). All but one study (34) included race/ethnicity data (all self-reported), but only three (50%) were multiracial/multiethnic, and only two (33%) consider

race/ethnicity as a risk factor. Other sensitive features considered in the studies include covariates for geographic region and level of urbanization (33) and covariates measuring obesity (30).

COVID-19

For COVID-19, the PubMed search query for all fields *(((covid) AND (risk)) AND (prediction) AND (2021/1/1:2023/12/31[pdat])) AND ((((hospitalization) OR (death)) OR (severity)) OR (mortality) AND (2021/1/1:2023/12/31[pdat])) NOT (((long) AND (meta)) AND (review) AND (2021/1/1:2023/12/31[pdat]))* conducted on November 13, 2023, returned 5722 results. The addition of the term "race" returned 141 results (2.46% of papers). The addition of the term "sex" (but not "race") returned 614 results (10.73% of papers). Specifying only one race/ethnicity returned 103 results ("Black"), 62 results ("Hispanic"), and 71 results ("Asian"). The Google Scholar search query: *allintitle: "covid" "risk" "predict" OR "prediction" AND "hospitalization" OR "mortality" OR "severity" OR "death"* with a time range of 2021–2023 on November 13, 2023, returned 2970 results, 1000 of which were accessible for review. These results were then selected if they exceeded 50 citations of 2021 papers and exceeded 10 citations of 2022 to 2023 papers, both from journals with impact factors over 5. This yielded a shortlist of 22 articles detailing models predicting the risk of COVID-19 hospitalization or death. These 22 were then divided into groups based on their target population (general population or a specific subpopulation). Sections S.3 and S.4 of the supplementary material provide details.

Of the 12 COVID-19 papers (supplementary material section S.3) focusing on a general population that met the inclusion criteria, none mention fairness metrics. Most papers (10 out of 12 papers, 83.33%) have sex as a covariate included in the prediction model; only one paper (8.33%) stratified by sex. Five out of 12 papers (41.67%) considered race/ethnicity, all five of which included it as a covariate—none stratified by race/ethnicity (see Figure 2). Of the remaining seven papers, three cited a lack of diverse data as the reason for excluding race/ethnicity information; one assumed the population to be entirely white (35); and the other three made no comments on race/ethnicity. Other potentially sensitive covariates explored include patient-level socioeconomic index (35) and the patient's geographical region or hospital region (4 out of the 12 papers).

Similar results were observed in the list of ten COVID-19 papers (supplementary material section S.4) that focused on specific subpopulations. Subpopulations considered in the list range from those with pre-existing chronic diseases (36,37) to papers focused on the elderly (38) and infants (39). All ten papers included age as a risk factor; similarly, all ten included sex in their models. Only one paper implemented stratification by sex, with the remaining nine out of 10 (90%) including sex as a covariate. In all five papers (50%) that include race/ethnicity data, the information is used as a covariate rather than a means of stratification. Though there was very little stratification by sex and none by race/ethnicity (see Figure 2), stratification was undertaken for various other risk factors, such as the type of medication used (40) and geographical location (41).

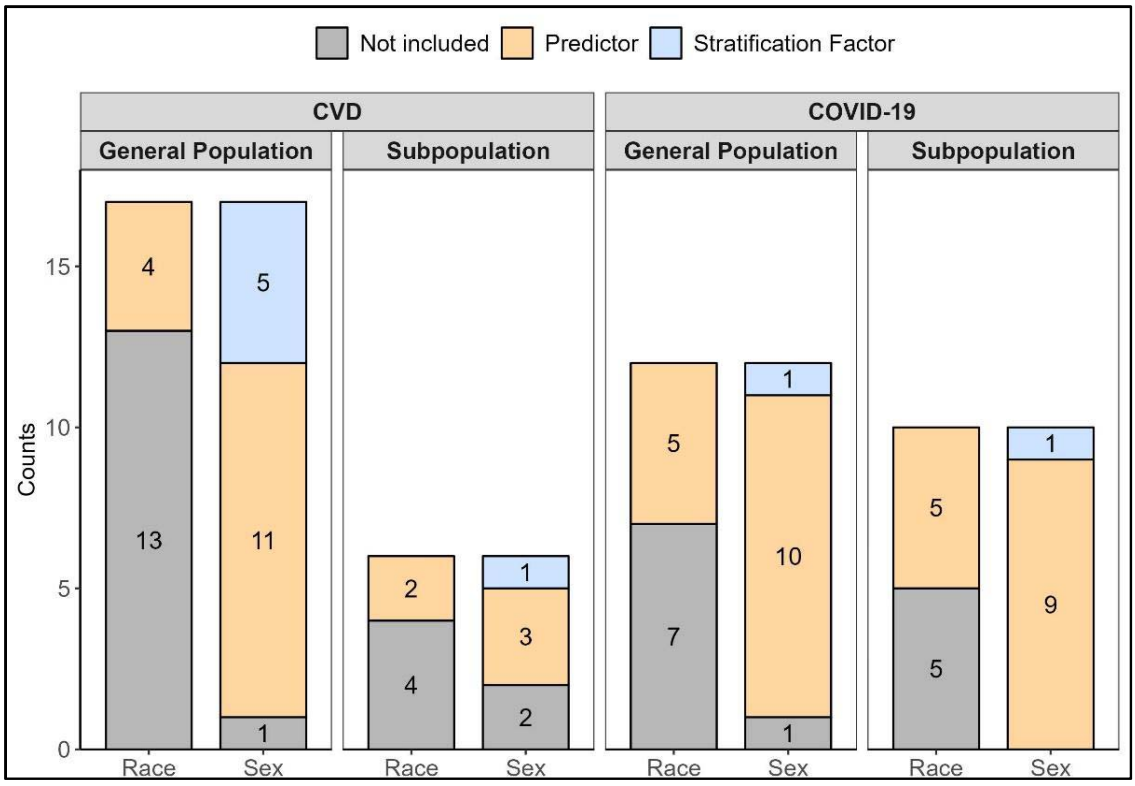


Figure 2. Counts depicting how many of the reviewed articles include race/ethnicity and/or sex as either predictors or stratification in their clinical risk prediction models, separated by disease of interest and population focus of the article. For cardiovascular diseases (CVD) the outcome considered was risk of cardiovascular disease, heart attack or heart failure, whereas for COVID-19 it was hospitalization and death.

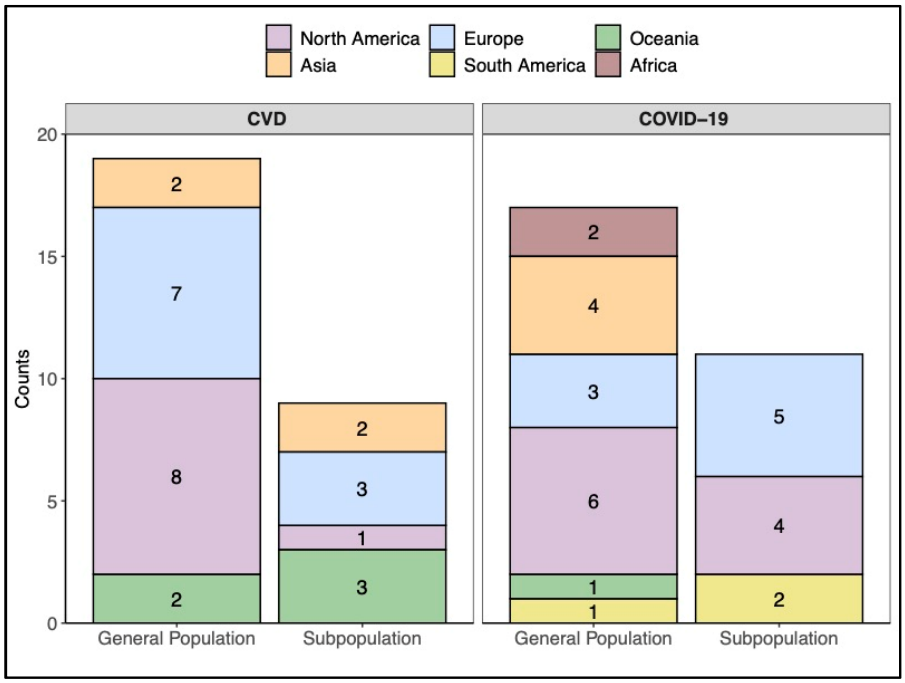


Figure 3. Counts depicting the geographic origin of data used in the reviewed articles, separated by disease of interest and population focus of the article. Studies with study regions covering multiple continents are double counted.

Discussion and Conclusion

We found that while the practice of assessing differences in model performance for sensitive features like sex and race/ethnicity was common, the use of fairness metrics to evaluate clinical risk prediction models was rare. Even though COVID-19 model constructions began well after the field of fairness metrics was first developed, fairness metric usage and similar considerations of sensitive features were as absent among COVID-19 as they were among CVD studies. Though some studies (7 of 23 of the CVD papers, 9 of 22 of the COVID-19 ones) used various discrimination and calibration metrics to assess model fit across different subgroups (such as across race/ethnicity), such calibration analysis was not routinely carried out; regardless, high discrimination and calibration alone do not guarantee model fairness (42,43). It is clear that, as the field of algorithmic fairness has grown and other disciplines have begun integrating fairness metrics into their own predictive modeling (44–48), clinical risk prediction models have not kept up with the progress.

Our reviews suggest that one major reason slowing the uptake of fairness metrics in clinical risk prediction models is the lack of data from diverse populations, particularly in a racial/ethnic and geographical sense. Though over half of the studies in the identified high

impact CVD papers were multiracial/multiethnic, the data used in these studies was still over 50% one race/ethnicity. While there was more racially/ethnically diverse data in the COVID-19 papers—of the ten papers with multiracial/multiethnic data, only five (50%) were still a majority one race/ethnicity—a lack of geographical diversity remained (see Figure 3). Most studies (22 out of the 23 CVD papers, 15 out of 22 of the COVID-19 ones) were from the Global North. Though important work has been done to improve clinical risk prediction by focusing on data from underrepresented subgroups, such as with African Americans in the Jackson Heart Study (49), the results from our review suggest that such equity-focused research remains far from the norm in this field. Our search was limited to Google Scholar's database and could have missed important articles, but high-level PubMed search queries support this conclusion: the search query of *"fairness"[All Fields] AND "risk prediction"[All Fields]* on November 16, 2023 returned only 15 papers total, none of which met our criteria for inclusion in our review.

The largely homogenous data observed in our review implies that, for many studies, fairness assessment is infeasible to begin with. Lacking sufficient data to include meaningful covariates for different sensitive features, the inclusion of fairness metrics in analyses will be absent by default. Yet even when there clearly was the opportunity, such as the articles whose models were stratified by sex or study region, fairness analyses were not done. The opportunity to use fairness metrics is there; it simply has not been adopted as a part of the assessment routine.

It is possible that a lack of clarity as to how to best approach model fairness contributed to the dearth of fairness considerations seen in our review, as even when a clear opportunity for fairness metrics arises, the question of how to properly leverage them remains complex. There is no "one size fits all" method or universal fairness metric: instead, the specific context of the model—whether it is a preventative intervention or a limited-supply treatment, for example—must inform how relevant concerns of fairness are, and what metrics can address the concerns (6,7). Many fairness criteria are in fact mutually incompatible in practical settings (for example, demographic parity and equalized odds (2)), requiring a case-by-case decision on what kind of fairness (and thus fairness metric) will be most meaningful for the data and situation at hand. There is also the matter of the limitations of many of the most common fairness metrics that make their use inapplicable or unappealing for certain models. For example, most metrics are designed to assess only dichotomous outcomes and would require recalculation if a model's predictions involved different cutoffs of the underlying continuous measures for different

decisions. The "polarity" of a predicted outcome (where a "polar" outcome is one that is always preferred (6)) also impacts the importance of fairness, and identifying said polarity is not always straightforward (6,7). These limitations could be further contributing to the slow uptake in a clinical setting.

No critical number of developed fairness metrics will address this problem—it is not an issue of lack of methods, but of implementation. The methods exist and have already been adopted in a variety of other predictive modeling fields, including criminal justice, finance, and computational linguistics (46,48,50). There are signs of progress in clinical risk prediction, on both the applied (42) and theoretical (51) sides, but more work like these papers is needed. Our own recommendations for how this can happen in the field of clinical prediction are listed below and illustrated in the roadmap of Figure 4.

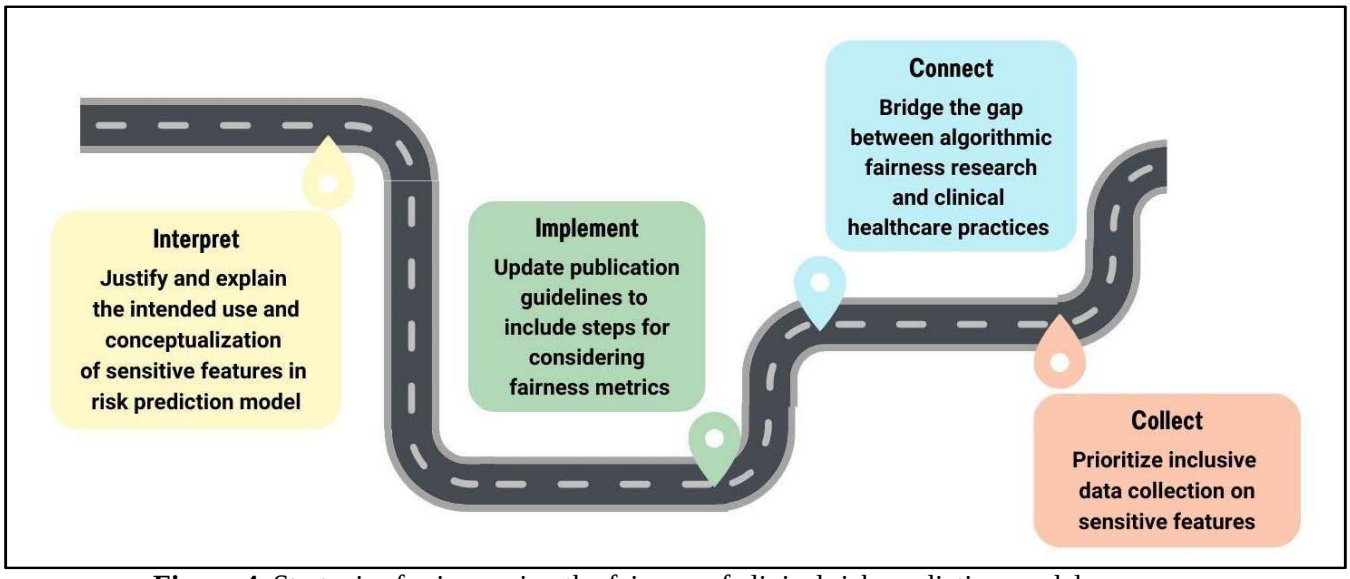


Figure 4. Strategies for increasing the fairness of clinical risk prediction models: Interpret, Implement, Connect and Collect (I2C2).

Strategies for increasing the fairness of clinical risk prediction models across sensitive variables (I2C2)

- **Interpret:** In line with the NIH's requirement of including (or justifying the exclusion) of a specifically biological sex variable, papers should interpret, justify, and explain their intended use and conceptualization of sensitive features in risk prediction models.
- **Implement:** Influential guidelines like EQUATOR's TRIPOD guidelines (47, 48) should

include steps on considering algorithmic fairness as part of implementation and application of clinical risk prediction models.

- **Connect:** The community of methods research in algorithmic fairness should ensure that the methods and tools developed are well broadcast to those in the community of practice in clinical healthcare, connecting theory to practice.
- **Collect:** The field of clinical risk prediction should highly prioritize collecting inclusive data across race, geographic region, and a variety of other sensitive or historically underrepresented features.

To understand the current barriers in the practice community, we have developed a short questionnaire for the authors of our selected studies or those like them to help the field identify key challenges in implementing fairness metrics. This questionnaire will help elucidate where resources should be most concentrated for this I2C2 roadmap. It can be accessed here, and Table 3 below shows the summary questions. We hope by educating and enabling practitioners around use of fairness metrics, we will create a more equitable prediction world for all.

A 10-question survey for understanding investigators' barriers for using fairness metrics in clinical risk prediction models	
Question 1	What is the overall goal of the clinical risk prediction model that you are constructing?
Question 2	What were the sensitive features included in your prediction model?
Question 3	What is the intended use of the sensitive features in your prediction models?
Question 4	Are the measures (and how they were collected) consistent with their intended use?
Question 5	How were the sensitive features you selected in Q2 included in your model?
Question 6	What were the criteria used for model evaluation and performance?
Question 7	Were the model evaluation criteria focused on overall performance or the performance within specific subgroups of the data defined by the sensitive variables you chose in Q2?
Question 8	Was model fairness considered and/or assessed?
Question 9	If you answered yes to Q7, please explain how you considered model fairness. If you answered no, please explain what prevented such consideration.
Question 10	If there were any other challenges or insights concerning the use (or non-use) of model fairness in your developed prediction model that you would like to share, please do so below.

Table 3. The ten questions included in the questionnaire. All but questions 9 and 10 are multiple choice, with the option to elaborate in a free response.

All materials used for this review can be accessed via GitHub.

References

1. Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. Scientific Reports. 2022 Mar 10;12(1):4209.

2. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. Association for Computing Machinery. 2021;54(6).

3. Bærøe, Torbjørn Gundersen, Edmund Henden, Kjetil Rommetveit. Can medical algorithms be fair? Three ethical quandaries and one dilemma. BMJ Health Care Inform. 2022 Apr 1;29(1):e100445.

4. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annu Rev Stat Appl. 2021 Mar 7;8(1):141–63.

5. Suresh H, Guttag J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In New York, NY, USA; 2021.

6. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. npj Digital Medicine. 2020 Jul 30;3(1):99.

7. Evans CV, Johnson ES, Lin JS. Assessing Algorithmic Bias and Fairness in Clinical Prediction Models for Preventive Services. Agency for Healthcare Research and Quality [Internet]. 2023; Available from: https://www.uspreventiveservicestaskforce.org/uspstf/sites/default/files/inline-files/assessing-algorithmic-bias-fairness_0.pdf

8. Verma S, Rubin J. Fairness Definitions Explained. Association for Computing Machinery. 2018;1–7.

9. Adebayo JA. FairML□: ToolBox for diagnosing bias in predictive modeling [Internet]. Massachusetts Institute of Technology; 2016. Available from: http://hdl.handle.net/1721.1/108212

10. Ashryaagr. Julia. 2020. Fairness.jl. Available from: https://www.juliapackages.com/p/fairness

11. Kozodoi N, Varga TV. Nikita Kozodoi. 2020. Algorithmic Fairness in R. Available from: https://kozodoi.me/r/fairness/packages/2020/05/01/fairness-tutorial.html

12. Fairlearn [Internet]. 2023. Available from: https://fairlearn.org/

13. Nabi R, Shpitser I. Fair Inference on Outcomes. AAAI [Internet]. 2018 Apr 25 [cited 2023 Sep 19];32(1). Available from: https://ojs.aaai.org/index.php/AAAI/article/view/11553

14. Fuentes A, Ackermann RR, Athreya S, Bolnick D, Lasisi T, Lee SH, et al. AAPA Statement on Race and Racism. American Journal of Physical Anthropology. 2019 Jul 1;169(3):400–2.

15. Lu C, Ahmed R, Lamri A, Anand SS. Use of race, ethnicity, and ancestry data in health research. PLOS Global Public Health. 2022 Sep 15;2(9):e0001060.

16. Smedley A, Smedley BD. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. American Psychologist. 2005;60(1):16–26.

17. Paulus JK, Kent DM. Race and Ethnicity: A Part of the Equation for Personalized Clinical Decision Making? Circulation: Cardiovascular Quality and Outcomes. 2017 Jul 1;10(7):e003823.

18. Sirugo G, Tishkoff SA, Williams SM. The quagmire of race, genetic ancestry, and health disparities. J Clin Invest [Internet]. 2021 Jun 1;131(11). Available from: https://doi.org/10.1172/JCI150255

19. Arnegard ME, Whitten LA, Hunter C, Clayton JA. Sex as a Biological Variable: A 5-Year Progress Report and Call to Action. Journal of Women's Health. 2020 Jun 1;29(6):858–64.

20. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. Nature. 2019 Nov 1;575(7781):137–46.

21. Khan SS, Coresh J, Pencina MJ, Ndumele CE, Rangaswami J, Chow SL, et al. Novel Prediction Equations for Absolute Risk Assessment of Total Cardiovascular Disease Incorporating Cardiovascular-Kidney-Metabolic Health: A Scientific Statement From the American Heart Association. Circulation [Internet]. 2023 Nov [cited 2023 Nov 15];0(0). Available from: https://doi.org/10.1161/CIR.0000000000001191

22. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. N Engl J Med. 2020 Aug 27;383(9):874–82.

23. Videha Sharma, Ibrahim Ali, Sabine van der Veer, Glen Martin, John Ainsworth, Titus Augustine. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. BMJ Health Care Inform. 2021 Feb 1;28(1):e100253.

24. Lee TC, Shah NU, Haack A, Baxter SL. Clinical Implementation of Predictive Models Embedded within Electronic Health Record Systems: A Systematic Review. Informatics. 2020;7(3).

25. Basu A. Use of race in clinical algorithms. Science Advances. 2023;9(21):eadd2704.

26. Pylypchuk R, Wells S, Kerr A, Poppe K, Riddell T, Harwood M, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. The Lancet. 2018 May 12;391(10133):1897–907.

27. McClelland RL, Jorgensen Neal W., Budoff Matthew, Blaha Michael J., Post Wendy S., Kronmal Richard A., et al. 10-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors. Journal of the American College of Cardiology. 2015 Oct 13;66(15):1643–53.

28. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal. 2021 Jul 1;42(25):2439–54.

29. Friis-Møller N, Ryom L, Smith C, Weber R, Reiss P, Dabis F, et al. An updated prediction model of the global risk of cardiovascular disease in HIV-positive persons: The Data-collection on Adverse Effects of Anti-HIV Drugs (D:A:D) study. Eur J Prev Cardiolog. 2016 Jan 1;23(2):214–23.

30. Goh LG, Satvinder S Dhaliwal, Timothy A Welborn, Andy H Lee, Phillip R Della. Anthropometric measurements of general and central obesity and the prediction of cardiovascular disease risk in women: a cross-sectional study. BMJ Open. 2014 Feb 1;4(2):e004138.

31. Thompson-Paul AM, Lichtenstein KA, Armon C, Palella FJ Jr, Skarbinski J, Chmiel JS, et al. Cardiovascular Disease Risk Prediction in the HIV Outpatient Study. Clinical Infectious Diseases. 2016 Dec 1;63(11):1508–16.

32. Vistisen D, Andersen GS, Hansen CS, Hulman A, Henriksen JE, Bech-Nielsen H, et al. Prediction of First Cardiovascular Disease Event in Type 1 Diabetes Mellitus. Circulation. 2016 Mar 15;133(11):1058–66.

33. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population. Circulation. 2016 Nov 8;134(19):1430–40.

34. Norrish G, Ding T, Field E, Ziółkowska L, Olivotto I, Limongelli G, et al. Development of a Novel Risk Prediction Model for Sudden Cardiac Death in Childhood Hypertrophic Cardiomyopathy (HCM Risk-Kids). JAMA Cardiology. 2019 Sep 1;4(9):918–27.

35. Simpson CR, Chris Robertson, Steven Kerr, Ting Shi, Eleftheria Vasileiou, Emily Moore, et al. External validation of the QCovid risk prediction algorithm for risk of COVID-19 hospitalisation and mortality in adults: national validation cohort study in Scotland. Thorax. 2022 May 1;77(5):497.

36. Bajaj JS, Guadalupe Garcia-Tsao, Scott W Biggins, Patrick S Kamath, Florence Wong, Sara McGeorge, et al. Comparison of mortality risk in patients with cirrhosis and COVID-19 compared with patients with cirrhosis alone and COVID-19 alone: multicentre matched cohort. Gut. 2021 Mar 1;70(3):531.

37. Peña JE de la, Rascón-Pacheco RA, Ascencio-Montiel I de J, González-Figueroa E, Fernández-Gárate JE, Medina-Gómez OS, et al. Hypertension, Diabetes and Obesity, Major

Risk Factors for Death in Patients with COVID-19 in Mexico. Archives of Medical Research. 2021 May 1;52(4):443–9.

38. Ramos-Rincon JM, Buonaiuto V, Ricci M, Martín-Carmona J, Paredes-Ruíz D, Calderón-Moreno M, et al. Clinical Characteristics and Risk Factors for Mortality in Very Old Patients Hospitalized With COVID-19 in Spain. The Journals of Gerontology: Series A. 2021 Mar 1;76(3):e28–37.

39. Halasa NB, Olson SM, Staat MA, Newhams MM, Price AM, Pannaraj PS, et al. Maternal Vaccination and Risk of Hospitalization for Covid-19 among Infants. N Engl J Med. 2022 Jul 14;387(2):109–19.

40. Bramante CT, Ingraham NE, Murray TA, Marmor S, Hovertsen S, Gronski J, et al. Metformin and risk of mortality in patients hospitalised with COVID-19: a retrospective cohort analysis. The Lancet Healthy Longevity. 2021 Jan 1;2(1):e34–41.

41. Oliveira EA, Colosimo EA, Simões e Silva AC, Mak RH, Martelli DB, Silva LR, et al. Clinical characteristics and risk factors for death among hospitalised children and adolescents with COVID-19 in Brazil: an analysis of a nationwide database. The Lancet Child & Adolescent Health. 2021 Aug 1;5(8):559–68.

42. Kartoun U, Khurshid S, Kwon BC, Patel AP, Batra P, Philippakis A, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. Scientific Reports. 2022 Jul 22;12(1):12542.

43. Loi M, Heitz C. Is Calibration a Fairness Requirement? An Argument from the Point of View of Moral Philosophy and Decision Theory. Association for Computing Machinery [Internet]. 2022;Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Available from: https://doi.org/10.1145/3531146.3533245

44. Halevy M, Harris C, Bruckman A, Yang D, Howard A. Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. In: Equity and Access in Algorithms, Mechanisms, and Optimization [Internet]. New York, NY, USA: Association for Computing Machinery; 2021. p. 1–11. (EAAMO '21). Available from: https://doi.org/10.1145/3465416.3483299

45. Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency [Internet]. PMLR; 2018. p. 134–48. (Proceedings of Machine Learning Research; vol. 81). Available from: https://proceedings.mlr.press/v81/chouldechova18a.html

46. Wang C, Han B, Patel B, Rudin C. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. Journal of Quantitative Criminology. 2023 Jun 1;39(2):519–81.

47. Wu C, Wu F, Wang X, Huang Y, Xie X. Fairness-aware News Recommendation with Decomposed Adversarial Learning. AAAI. 2021 May 18;35(5):4462–9.

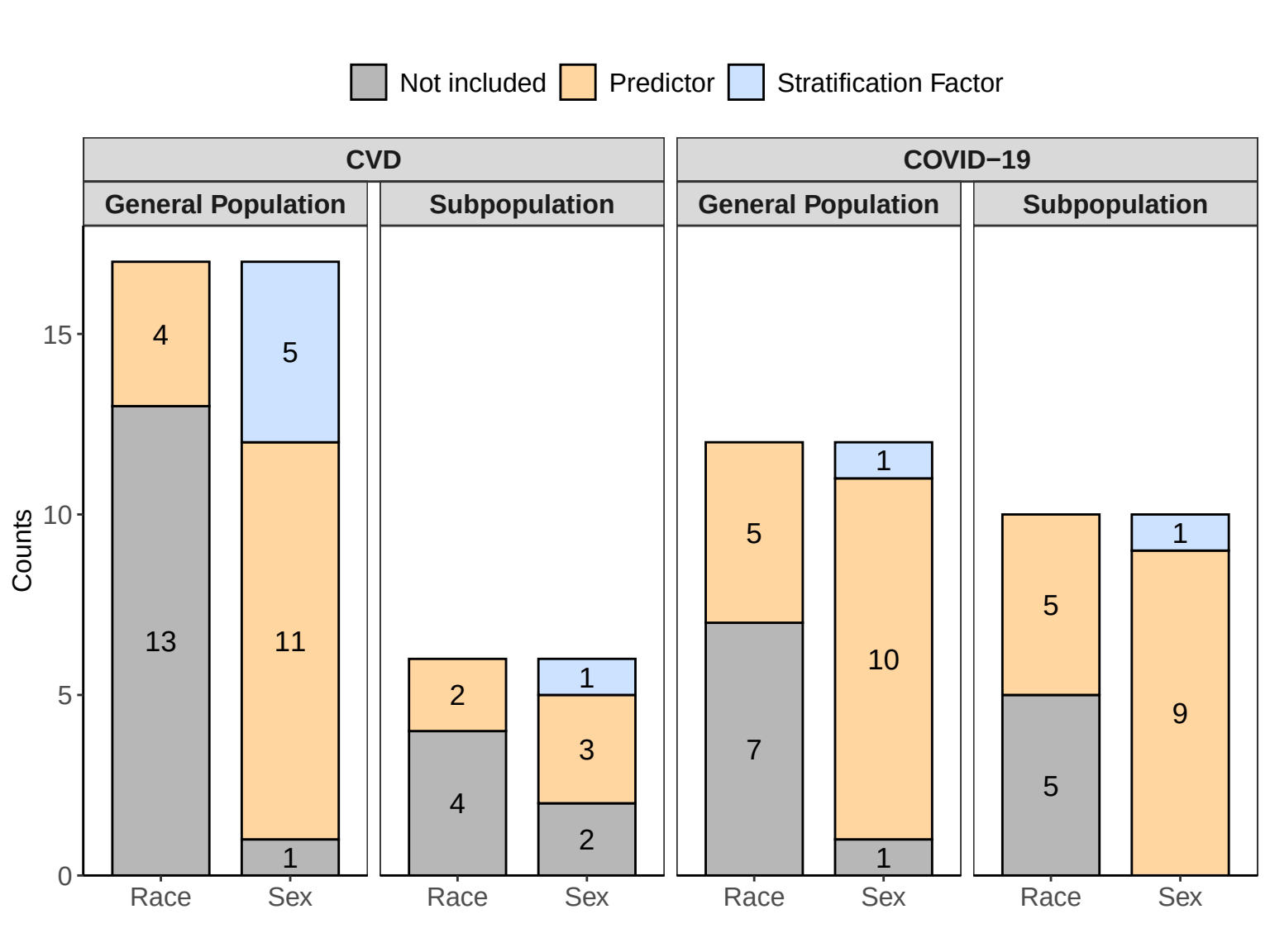
48. Y. -F. Te, M. Wieland, M. Frey, H. Grabner. Mitigating Discriminatory Biases in Success Prediction Models for Venture Capitals. In: 2023 10th IEEE Swiss Conference on Data Science (SDS). 2023. p. 26–33.

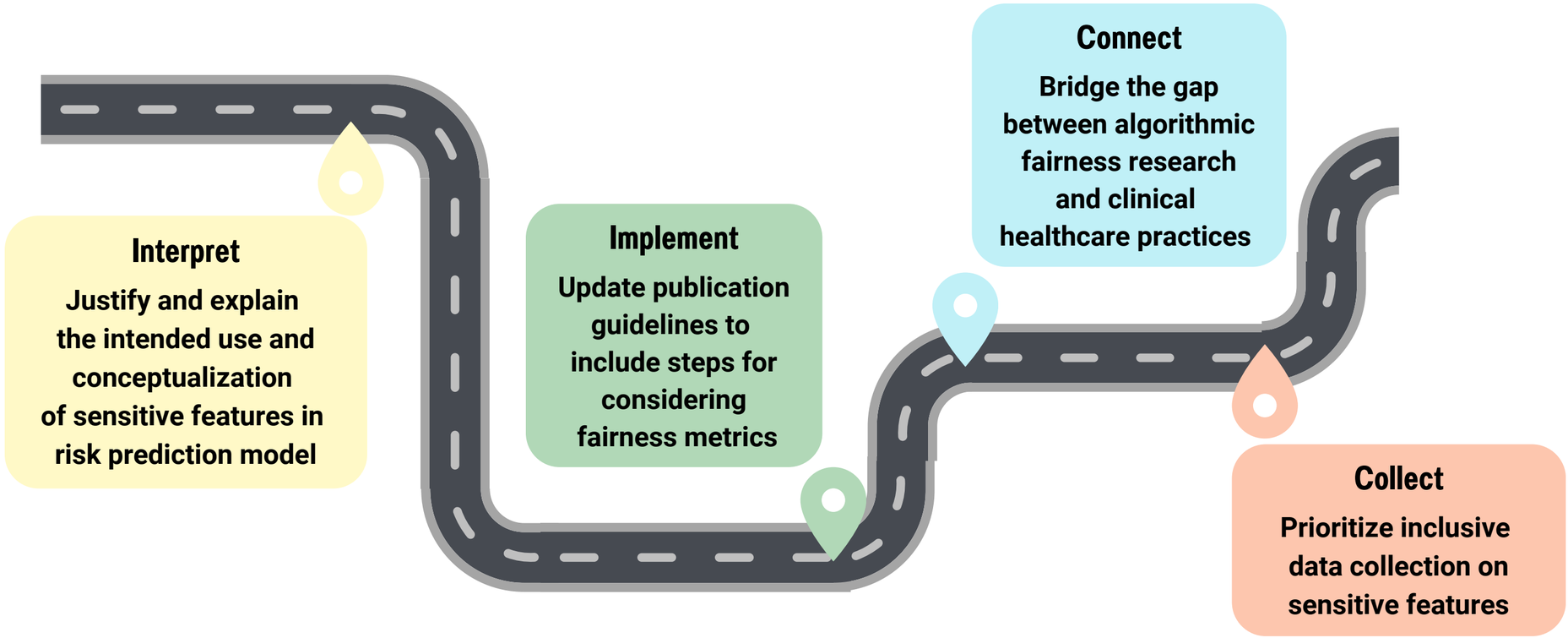
49. Addison C, Campbell Jenkins BW, White M, Thigpen Odom D, Fortenberry M, Wilson G, et al. Twenty Years of Leading the Way among Cohort Studies in Community-Driven Outreach and Engagement: Jackson State University/Jackson Heart Study. International Journal of Environmental Research and Public Health. 2021;18(2).

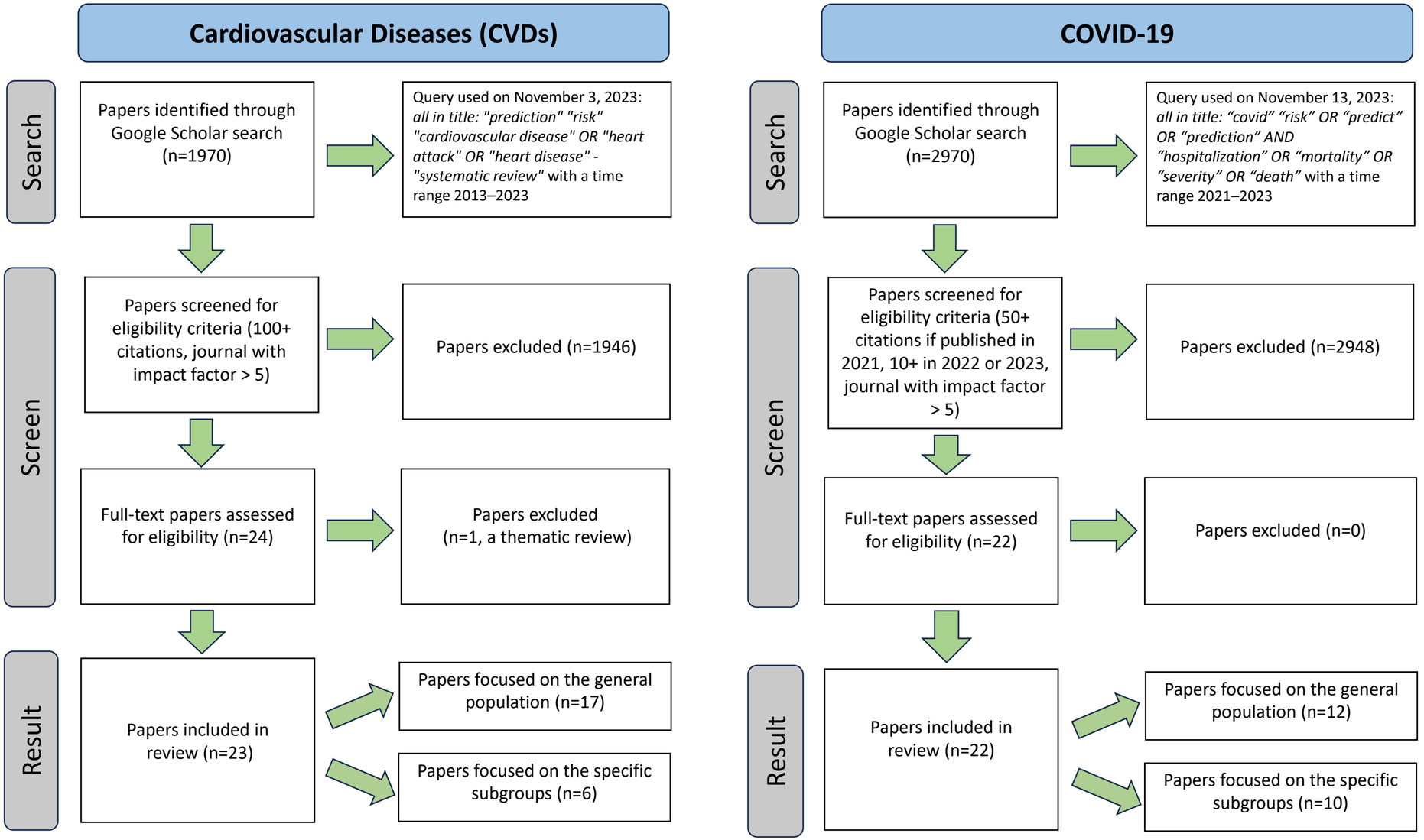
50. Schwertmann L, Kannan Ravi MP, de Melo G. Model-Agnostic Bias Measurement in Link Prediction. In: Findings of the Association for Computational Linguistics: EACL 2023 [Internet]. Dubrovnik, Croatia: Association for Computational Linguistics; 2023. p. 1632–48. Available from: https://aclanthology.org/2023.findings-eacl.121

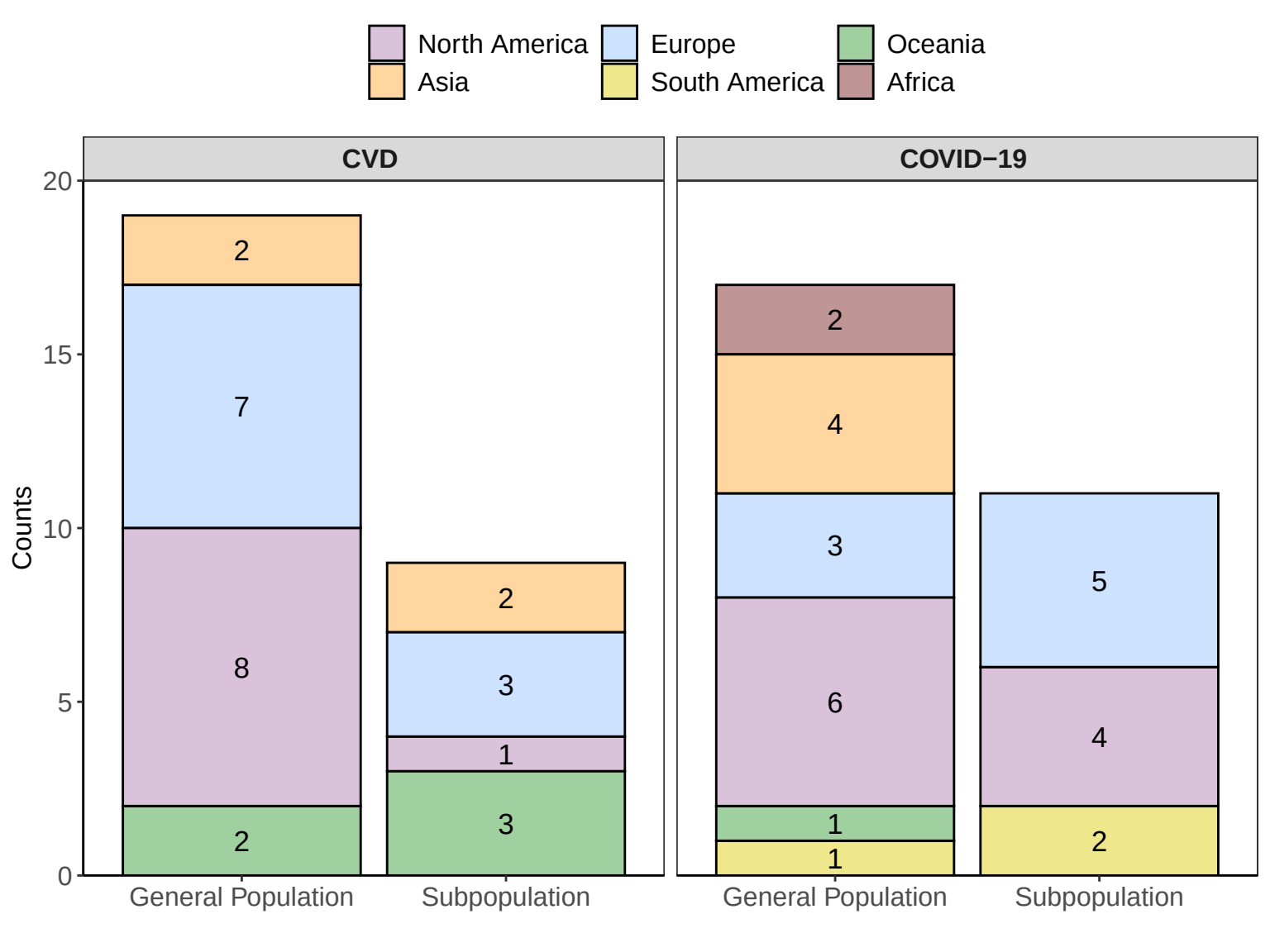
51. Wastvedt S, Huling JD, Wolfson J. An intersectional framework for counterfactual fairness in risk prediction. Biostatistics. 2023 Aug 31;kxad021.

Metric	Equation	Definition
Equalized Odds	$P(\hat{Y} = 1 \mid S = a, Y = y) = P(\hat{Y} = 1 \mid S = b, Y = y)$	Both groups have equal true positive and false positive rates.
Equal Opportunity	$P(\hat{Y} = 1 \mid S = a, Y = 1) = P(\hat{Y} = 1 \mid S = b, Y = 1)$	Both groups have equal true positive rates. A relaxed version of equalized odds.
Predictive Equality	$P(\hat{Y} = 1 \mid Y = 0, S = a) = P(\hat{Y} = 1 \mid Y = 0, S = b)$	The rate of false positives (negative events categorized as positives) is independent of the sensitive feature.
False Negative Rate Parity	$P(\hat{Y} = 0 \mid Y = 1, S = a) = P(\hat{Y} = 0 \mid Y = 1, S = b)$	The rate of false negatives (positive events categorized as negatives) is independent of the sensitive feature.
Predictive Parity	$P(Y = 1 \mid \hat{Y} = 1, S = a) = P(Y = 1 \mid \hat{Y} = 1, S = b)$	Model precision is the same for both groups.
Demographic Parity	$P(\hat{Y} = 1 \mid S = a) = P(\hat{Y} = 1 \mid S = b)$	The prediction or decision is independent of the sensitive feature.


Not included
Predictor
Stratification Factor
CVD
COVID−19
General Population
Subpopulation
General Population
Subpopulation
Counts
0
5
10
15
4
13
5
11
1
2
4
1
3
2
5
7
1
10
1
5
5
1
9
Race
Sex
Race
Sex
Race
Sex
Race
Sex




Cardiovascular Diseases (CVDs)
Search
Papers identified through Google Scholar search (n=1970)
Query used on November 3, 2023: all in title: "prediction" "risk" "cardiovascular disease" OR "heart attack" OR "heart disease" - "systematic review" with a time range 2013–2023
Screen
Papers screened for eligibility criteria (100+ citations, journal with impact factor > 5)
Papers excluded (n=1946)
Full-text papers assessed for eligibility (n=24)
Papers excluded (n=1, a thematic review)
Result
Papers included in review (n=23)
Papers focused on the general population (n=17)
Papers focused on the specific subgroups (n=6)
COVID-19
Search
Papers identified through Google Scholar search (n=2970)
Query used on November 13, 2023: all in title: “covid” “risk” OR “predict” OR “prediction” AND “hospitalization” OR “mortality” OR “severity” OR “death” with a time range 2021–2023
Screen
Papers screened for eligibility criteria (50+ citations if published in 2021, 10+ in 2022 or 2023, journal with impact factor > 5)
Papers excluded (n=2948)
Full-text papers assessed for eligibility (n=22)
Papers excluded (n=0)
Result
Papers included in review (n=22)
Papers focused on the general population (n=12)
Papers focused on the specific subgroups (n=10)


North America
Europe
Oceania
Asia
South America
Africa
CVD
COVID−19
Counts
20
15
10
5
0
2
7
8
2
2
3
1
3
2
4
3
6
1
1
5
4
2
General Population
Subpopulation
General Population
Subpopulation

A 10-question survey for understanding investigators' barriers for using fairness metrics in clinical risk prediction models	
Question 1	What is the overall goal of the clinical risk prediction model that you are constructing?
Question 2	What were the sensitive features included in your prediction model?
Question 3	What is the intended use of the sensitive features in your prediction models?
Question 4	Are the measures (and how they were collected) consistent with their intended use?
Question 5	How were the sensitive features you selected in Q2 included in your model?
Question 6	What were the criteria used for model evaluation and performance?
Question 7	Were the model evaluation criteria focused on overall performance or the performance within specific subgroups of the data defined by the sensitive variables you chose in Q2?
Question 8	Was model fairness considered and/or assessed?
Question 9	If you answered yes to Q7, please explain how you considered model fairness. If you answered no, please explain what prevented such consideration.
Question 10	If there were any other challenges or insights concerning the use (or non-use) of model fairness in your developed prediction model that you would like to share, please do so below.

Sensitive variable	Definition
Social	Variables based on an individual's lived experience. Frequently self-reported.
Race	A social construct based on perceived physical differences (14–16) that often acts as a proxy for various social and health consequences of racism in modeling (17). Often paired with ethnicity, though they are not strictly equivalent. Includes no biological information.
Ethnicity	A social construct based on shared culture, language, geography, religion, and history (15,18). Often paired with race, though they are not strictly equivalent. Includes no biological information.
Gender	A social construct referring to an individual's self-presentation in society, informed by culture, psychology, and society (19). Though distinct from sex, it often indirectly captures sex's impact.
Biological	Variables based on an individual's biology.
Genetic ancestry	Genetic similarities between people due to common ancestors (18). Distinct from race. Strictly biological.
Sex	The biological factors used to categorize individuals as male, female, or intersex (20). Since 2016, it has been a required covariate in NIH-funded research (19). Though distinct from gender, it often indirectly captures gender's impact.