Algorithmic Fairness in Computational Medicine

Jie Xu[1,2], Yunyu Xiao[2], Wendy Hui Wang[3], Yue Ning[3], Elizabeth A Shenkman[1], Jiang Bian[1], and Fei Wang[2,*]

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA
[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA
[3]Department of Computer Science, Stevens Institute of Technology, Hoboken, New Jersey, USA
[*]few2001@med.cornell.edu

ABSTRACT

Machine learning models are increasingly adopted for facilitating clinical decision-making. However, recent research has shown that machine learning techniques may result in potential biases when making decisions for people in different subgroups, which can lead to detrimental effects on the health and well-being of vulnerable groups such as ethnic minorities. This problem, termed algorithmic bias, has been extensive studied in theoretical machine learning recently. However, how it will impact medicine and how to effectively mitigate it still remains unclear. This paper presents a comprehensive review of algorithmic fairness in the context of computational medicine, which aims at improving medicine with computational approaches. Specifically, we overview the different types of algorithmic bias, fairness quantification metrics, and bias mitigation methods, and summarize popular software libraries and tools for bias evaluation and mitigation, with the goal of providing reference and insights to researchers and practitioners in computational medicine.

1 Introduction

The recent years have witnessed a surge of interests on development and deployment of machine learning algorithms in healthcare. These algorithms were learned from massive health data and have demonstrated promising performance in a diverse set of problems such as skin cancer detection from lesion images[1], prediction of the risk of acute kidney injury based on electronic health records (EHR)[2], adaptive learning of the optimal treatment regimes for sepsis patients in intensive care[3], and others[4].

Despite the promise, there is growing concern that machine learning algorithms may lead to unconscious bias when making decisions against ethnic minorities, both through the algorithms themselves and the data used to learn them. For example, associations between Framingham risk factors and cardiovascular events have been shown to be significantly different across different ethnicity groups[5]. Video stream analysis algorithms for measuring the body's spontaneous blink rate have been found to be particularly challenging for Asian individuals[6,7]. Undiagnosed silent hypoxemia, which can be detected by pulse oximetry using light to monitor vital signs, occurred approximately three times more frequently in Black people due to the fact that dark skin responds differently to those light wavelengths [8]. In these cases, the software system may bring in additional or exacerbate health equity issues[7].

With machine learning models gaining more and more attentions in medicine, it is crucial to be aware of the potential related bias and disparities, understand their causes, and mitigate them. This review will help achieve this goal by providing an overview of the existing literature studying the sources of bias and disparities in computational medicine, their quantification metrics, and mitigation strategies. We will also summarize outstanding questions and point out future directions.The PRISMA diagram of the literature reviewed in this paper is shown in Fig. 1.

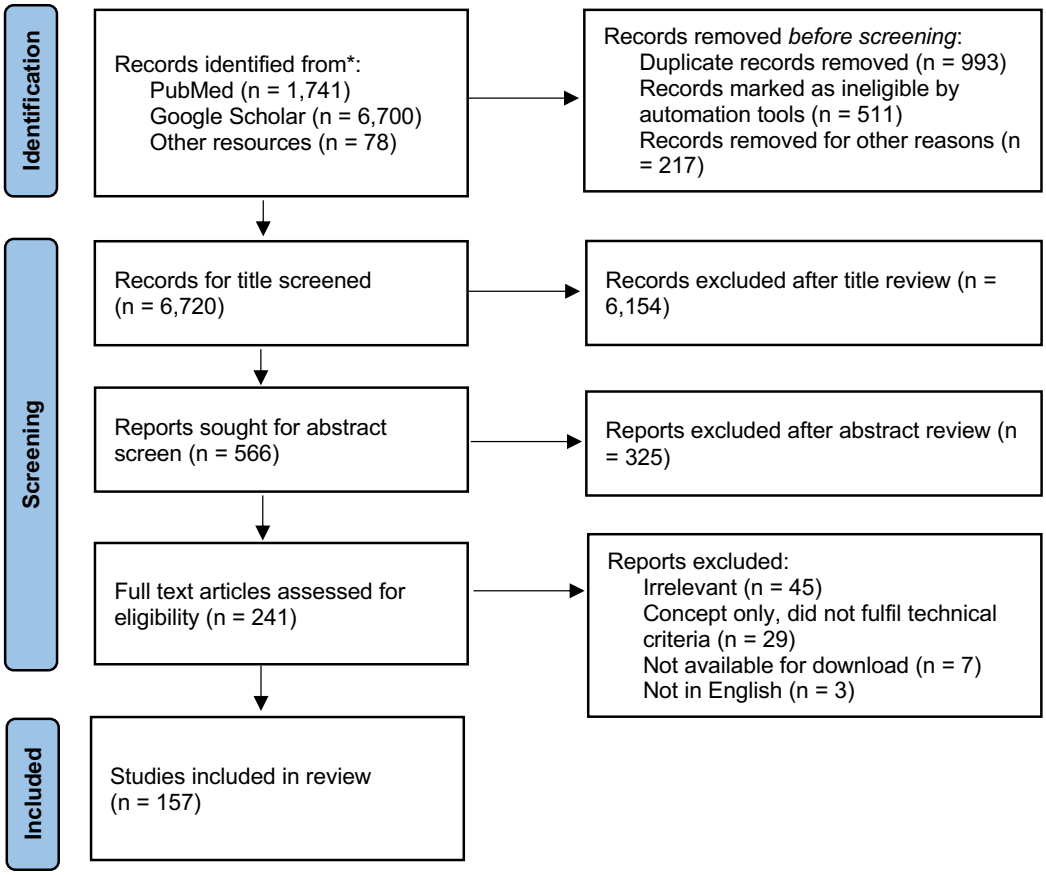

Figure 1. PRISMA flow diagram: disparity and fairness in computational medicine.

Difference with Existing Reviews. Mehrabi et al.[9] built a taxonomy of machine learning related fairness in different real world application contexts. Rajkomar et al.[10] introduced the principles of distributive justice and provided guidance to clinicians on how to prioritize each principle when facing with potential bias in model development and deployment. Gianfrancescogian et al.[11] summarized the potential bias sources for electronic health records (EHRs) and provided recommendations on appropriately mitigating them. Fletcher et al.[12] described three basic criteria (*i.e.*, Appropriateness, Fairness, and Bias) for evaluating machine learning and AI systems in the context of global health. Mhasawade et al.[13] focused on the interactions among different cultural, social, and environmental factors and their impact on individual and community health, how they will impact the fairness of machine learning algorithms and how machine learning, public and population health can work together to achieve health equity. Different from these existing works, this review summarizes sources and quantification methods for bias in computational medicine and how they will impact downstream machine learning models, as well as potential strategies to mitigate them through computational algorithms.

2 Computational Bias

We categorize computational biases into three different types according to their sources, *data bias*, *measurement bias*, and *algorithm bias*. We will introduce them in this section and provide examples in medical context.

2.1 Data Bias

Machine learning algorithms are all learned from data sets[14]. For example, classification models try to accurately map the sample input features to a set of pre-specified classes based on the observations from a set of training data. Clustering models aim at identifying grouping structures of a given data set. In this case, if the data set is over or under representing certain sample groups, the machine learning models learned from the data will be biased. For instance, studies found that patients of low socioeconomic status may have limited access to health care[15,16]. Consequently, compared to patients with better socioeconomic status, these patients may have less information in their EHRs or imaging and thus underrepresented in the data from which a machine learning model will be learned. This will lead to poorer model performance on this particular patient group. Below we list potential sources of data bias in medicine.

2.1.1 Sample bias

Sample bias, also known as selection bias, occurs when the selected data can not represent the real environment in which a model will be deployed[17]. For example, melanoma detection algorithms based on classification of skin lesion images[1] may perform poorly on colored skins if the training images are mostly with white skin[18]. For the same reason, Face2Gene, a machine learning algorithm to recognize Down syndrome based on facial images, performed much better in Caucasian (accuracy 80%) than in African (accuracy 36.8%)[19].

2.1.2 Allocation Bias

Allocation bias is relevant to clinical trials of interventions, which arises if there are systematic differences in how participants are allocated to treatment and control groups[20]. If researchers know or are able to predict which participants would benefit from an intervention, it would affect how they recruit participants and how they assign them to different groups so that they can select subjects with a good prognosis for trials. Allocation concealment could protect the randomization process, keep participants unaware of the intervention to be assigned before entering the study, and prevent prediction of subsequent allocations in actual clinical trials[20]. Recently there were studies trying to emulate clinical trials with real world data such as EHRs[21,22]. In this case, allocation bias could exist as the treatment and control groups are already observed in the data. This can lead to potentially bias estimations of treatment effects with machine learning models[23].

2.1.3 Attrition Bias

Attrition bias can occur if there are systematic differences in the way participants are dropped from the study, as different rates of losses to follow-up in the exposure groups may alter the characteristics of these groups[20]. Attrition bias will be more severe in observational data analysis as patients may move to another place or be transferred to another hospital, which will impact the machine learning model aiming at prediction of clinical events.

2.1.4 Publication bias

Publication bias occurs when a study is published and not depending on its own results[24]. Empirical studies consistently show that studies with positive or statistically significant results are easier and take less time to be published than studies without significant results[25,26]. This can make it difficult for decision makers to distinguish between sound evidence and overestimate the effectiveness of treatment or models[26]. For example, since the start of the COVID-19 pandemic, studies on COVID-19 is being published at a rapid rate. However, many peer-reviewed publications were with a limited number of patients included and showed a high risk of bias [27].

2.2 Measurement bias

Measurement bias is a systematic error that occurs when the data are labeled inconsistently, or study variables (*e.g.*, disease, exposure) are collected or measured inaccurately[28]. A recent example is there is a large disparity in the quality of COVID-19 data reporting across India[29]. Below we list several common causes of measurement bias.

2.2.1 Response bias

Response bias usually occurs in survey-based studies. When respondents tend to give inaccurate or even wrong answers on self-reported questions, the survey results will be affected[30]. For instance, people tend to paint the best picture of themselves, or feel pressured to provide socially acceptable answers[31]. In addition, misleading questions can lead to biased answers. Respondents may not have realized they weren't giving the answers in the way the investigator wanted them to[30]. In addition, people who are willing to answer survey questions are often different from those who are not[32]. Consequently, this will impact the machine learning algorithms trained on surveys or patient reported outcomes.

2.2.2 Recall Bias

Recall bias usually occurs during the data annotation phase of a project[33]. This happens when similar data are inconsistently labeled, thus leading to low accuracy. A participant may erroneously provide responses that rely on his/her ability to recall past events. However, recalling events of interest that happened long ago can be particularly difficult. For example, a publicity related to association between measles, mumps and rubella (MMR) vaccine and autism[34] influenced how often parents of autistic children recalled their child's regressive symptoms[35]. This may lead to the observation of a completely or partially untrue association between MMR and autism, which would subsequently impact the algorithms for inferring such associations.

2.2.3 Observer Bias

The observer bias occurs when the methods or procedures used to observe and record information for research leading to a systematic deviation from the facts[36], due to bad habits or lack of training in using measuring equipment or data sources[37]. Although some results of the diagnostic studies and physical examinations are objective, the symptoms and most findings of medical examinations are more or less subjective and prone to observer bias[38,39]. Hrobjartsson *et al.*[40] provided empirical evidence for observer bias in randomized clinical trials with results that involved subjective measurement scales. Consequently, these bias will be carried on to the machine learning models trained from these data.

2.3 Algorithm Bias

Another source of bias is from the algorithms themselves[41], which can be algorithm specific or agnostic. Algorithm specific bias is linked to their intrinsic hypotheses[42]. For example, logistic regression models assume the relationships between input and target variables are linear, but this may not be the reality. This bias makes the algorithm challenging to capture the actual input-output relationships in the data. We also list two types of algorithm agnostic bias as below.

2.3.1 Loss Function Bias

The loss function measures the difference between the output produced by the machine learning algorithm and the ground truth outcomes. It is used to evaluate how well the machine learning algorithm fits the data. Typical machine learning algorithms attempt to minimize such prediction loss on the training data, which is typically measured by adding up all prediction losses on individual samples. However, if certain group

of samples are more representative (e.g., white patients in a population[43]), the corresponding model will be better trained for this group.

2.3.2 Post-hoc Confirmation Bias

Although many machine learning models have been developed for binary classifications (e.g., disease diagnosis), they typically generated continues prediction scores and a cutoff threshold was needed for dichotomizing the predictions, and its optimal value is typically determined with post-hoc data driven analysis. The choice of cut points can introduce bias to diagnostic research. For example, Ewald analyzed the simulated data sets of test results from subjects with or without a particular disease and found that the use of data-driven cut points exaggerated test performance in many cases[44].

3 Fairness Metrics

The previous section has summarized the various potential sources of computational bias. Another important question is how can we quantify such bias given a specific healthcare context or data set. In this section, we will review the various bias quantification measures, which are referred to as fairness metrics. Mathematical notations that are used in this section are summarized in Table 1.

To facilitate the description of fairness metrics, we begin by a case study to build an alerting algorithm in ICU setting (e.g., for developing sepsis[45]) with machine learning based on EHR, and race (*i.e.*, Black and White is considered) as the protected attribute, which means we want to quantitatively examine if the algorithm behaves differently for black and white patients using various fairness metrics[46,47].

Table 1. Notations and Symbols.

Symbol	Description
$A \in \{0,1\}$	Binary protected attribute
$X \in \mathbb{R}^d$	Other observable attributes
U	Relevant latent attributes not observed
$Y \in \{0,1\}$	The outcome to be predicted
$\hat{Y} := f(X,A) \in \{0,1\}$	The prediction of Y
$\hat{Y}_{A \leftarrow a}$	Counterfactual value, *i.e.*, what would $\hat{Y}$ have been if A had been equal to a

3.1 Fairness through unawareness

Fairness through unawareness requires to not include the protected attribute (*e.g.*, race in our case study) as an independent variable in the model[51–53]. This method has been shown to be ineffective in situations where there are highly relevant features (*e.g.*, proxies for protected attributes). For example, race may be related to zip code, socioeconomic status or disease predisposition. Therefore, simply removing protected attribute is not enough.

3.2 Demographic parity

Demographic parity, also known as *statistical parity* or *independence*, requires that the overall proportion of individuals in a protected group predicted as positive (or negative) to be the same as that of the overall population[49]. Although it is intuitive to understand, prior studies[54] found that optimizing demographic parity may prevent the model from taking into account relevant clinical characteristics related to protected variables and outcomes, thereby reducing the performance of the model for all groups.

Table 2. Summary of Fairness Metrics

Type	Definition	In our case study
Fairness Through Unawareness[48]	No protected attribute A is explicitly used in the decision-making process: $\hat{Y} = f(X,A) = f(X)$	Train the model without race variable
Demographic Parity[49] / Statistical Parity / Independence	The outcomes must be equal: $\mathbb{P}(\hat{Y}\|A=0) = \mathbb{P}(\hat{Y}\|A=1)$	Blacks and Whites developed to sepsis at equal rates
Equalized Odds[50] / Separation	Different groups deal with similar odds, if $\hat{Y}$ and A are independent conditional on Y: $\mathbb{P}(\hat{Y}=1\|A=0,Y=y) = \mathbb{P}(\hat{Y}=1\|A=1,Y=y), y \in \{0,1\}$	The true positive rates (of those who actually developed sepsis, how many were correctly predicted to be positive) and false positive rates in Blacks and Whites are equal
Equal Opportunity[50]	The true positive rates in the unprivileged group and privileged group are equal. $\mathbb{P}\{\hat{Y}=1\|A=0,Y=1\} = \mathbb{P}\{\hat{Y}=1\|A=1,Y=1\}$	The true positive rates in Blacks and Whites are equal
Individual Fairness[51]	Similar individuals have similar predictions. Formally, given a metric $d(\cdot,\cdot)$, if individuals i and j are similar under this metric (*i.e.*, $d(i,j)$ is small), then their predictions should be similar: $\hat{Y}(X^{(i)},A^{(i)}) \approx \hat{Y}(X^{(j)},A^{(j)})$.	Similar patients have similar chance to develop sepsis
Counterfactual Fairness[48]	Predictor $\hat{Y}$ is counterfactually fair if under any context $X=x$ and $A=a$, $Pr(\hat{Y}_{A\leftarrow a}(U) = y\|X=x,A=a) = Pr(\hat{Y}_{A\leftarrow a'}(U) = y\|X=x,A=a)$, for all y and for any value a' attainable by A	The predicted outcome does not change if the values of the sensitive variables change

3.3 Equalized Odds

Unlike demographic parity, *equalized odds*[50] allows the prediction $\hat{Y}$ to depend on protected attribute A, but only through the target variable Y. This encourages the use of features that are directly related to Y, rather than through A[50]. To achieve equalized odds, both true positive rates (TPR) and true negative rates (TNR) of all groups defined by A are equal up to a fixed tolerance T. Compared to demographic parity, equalized odds is more flexible as it does not prevent learning a predictor where there is a real association between the protected attribute and the outcome[54].

3.4 Equal Opportunity

Equal opportunity checks whether the positive label is equally and accurately predicted by classifier for all values of the protected attribute[50]. In contrast to *equalized odds*, it is stronger because it means that all possible thresholds are equally likely to be met and therefore requires that all groups get the same ROC curve, but the decision threshold can be adjusted to satisfy *equalized odds*[54].

3.5 Individual Fairness

At a high level, individual fairness requires that any two individuals who are similar in the context of a given task should be treated similarly[51,55]. Clearly, individual fairness is more strict than group fairness defined by the protected attribute. The practical use of this concept is often limited due to the challenges of defining a appropriate similarity metric to encode the desired concept of fairness[51,54]. In addition, there

were also arguments that individual fairness is an inadequate as similar treatment is not enough to achieve fairness, thus it shouldn't be used alone to detect bias or evaluate whether algorithms are fair[56].

3.6 Counterfactual Measures

Counterfactual fairness is a potential way to explain why bias occurs. It states that a model is fair if its predictions about a particular individual in the real world is the same as it is in the counterfactual world, *i.e.* if the individuals is in a different protected group[48]. We list the mathematical definition of counterfactual fairness in the last row of Table 2, where $\hat{Y}_{A\leftarrow a}$ represents the the prediction $\hat{Y}$ if A had taken value a. This metric considers the predictor to be fair if its prediction remains unchanged when the protected attribute of each sample is flipped to its counterfactual value. A close concept of counterfactual fairness is counterfactual reasoning[57]. Some studies have shown that counterfactual reasoning is susceptible to biases such as outcome bias (that is, evaluating the quality of decisions when the outcome is known)[58]. In addition, it has been suggested that counterfactual reasoning may negatively affect the process of causality identification[59]. These concerns raise questions about the practical applicability of counterfactual measures.

Different metrics have different characteristics, and these aforementioned fairness metrics cannot be achieved at the same time, except in highly restricted special cases[60]. Both equalized odds and demographic parity focus on group fairness. Although their calculations and reasoning are simple and intuitive, the derived models may be discriminatory to structured subgroups with protected attributes, leading to fairness gerrymandering[54,61]. The concept of individual fairness potentially alleviate the issues of group fairness metrics by forcing any two individuals who are similar at a given task should be similarly classified. However, it is challenging to a domain-specific similarity measure, thus the practical use of individual fairness is often limited. Clinical prediction models may produce unfair results based on particular metrics. There is no clear consensus on what metric should be used in each scenario, researchers should choose the fairness metrics based on the given context.

4 Bias Mitigation

With the various sources of bias and different fairness metrics, in this section we will summarize different bias mitigation approaches for achieving algorithmic fairness. These methods can be categorized as pre-processing[62], in-processing[63–66], and post-processing methods[67], which are detailed below.

4.1 Pre-processing

Data pre-processing refers to the procedures of cleaning and preparing raw data for building machine learning models[68]. Pre-processing methods can potentially remove the bias from the data.

4.1.1 Sampling

Sampling is a popular preprocessing method to ensure the datasets are balanced across different groups[69]. If the data set is large, the majority group can be randomly sampled to the same size as the minority group without much information loss. However, if there is no redundant data, it is more common to oversample minority groups. Popular algorithms, like synthetic minority oversampling technique (SMOTE)[70] or its variations, such as SMOTE-ENC[71], Borderline-SMOTE[72]. However, healthcare data (such as EHRs or questionnaires) are typically complicated, and are thus challenging to be synthesized without overfitting[12]. In addition to sampling, collecting more data with good planning is also important to mitigate potential bias more objectively [43].

4.1.2 Reweighting

Reweighting is to impose different weights on each group-class combination based on the conditional probability of class by protected attribute, so that the protected attribute is independent of the outcome[62]. As a representative method, inverse propensity score weighting (IPW)[73] is often adopted to adjust poorly sampled data. It involves estimating the probability of individual participants in particular groups and then analysing the re-weighted samples of these participants[74]. However, IPW adjusts the distributions of all variables simultaneously, which may potentially increases imbalances and bias[75]. Borland *et al.*,[76] presented dynamic reweighting (DR) to correct selection bias with interactive visual analysis.

4.2 In-processing

In-processing methods aims at developing unbiased models directly from the data. A straightforward approach to achieve this goal is to remove the protected attribute from the model as we introduced in Section 3.1. However, if there are dependencies between the protected attribute and other covariates, the information of the sensitive attributes will "leak" into the decision.

4.2.1 Prejudice Remover

Prejudice refers to the fact that there is statistical dependence between the protected attribute and the predicted outcome or other independent variables[77]. Prejudice remover aims at learning a predictor whose predictions are independent of the protected attribute. For example, Kamiran and Calders *et al.*[78] proposed the concept of discrimination-aware classification and developed an algorithm to "clear away" such dependencies by "massaging the data" before applying traditional classification algorithms. Calders and Verwer [63] proposed a discrimination-free naive-Bayes through post-hoc processing, independent model training and balancing across different protected groups, or latent variable modeling. Kamishima et al.[65] proposed a prejudice remover regularization to enforce the prediction's independence on the protected attribute. Zafar et al.[64] proposed the concept of "disparate mistreatment" as different misclassification rates across different protected groups, and introduced a measure for decision boundary based classifiers, which further can be incorporated into the classifier optimization objectives as constraints to remove prejudice. With more and more machine learning models being developed for clinical risk prediction, there has been intense discussions on the ethical concerns[79,80]. These prejudice remover approaches can potentially make these algorithms fair.

4.2.2 Adversarial Learning

Adversarial learning[81] is a learning paradigm originally designed for generating fake samples to confuse the model. Typically there is a generator guaranteeing the generated fake samples which are close to real samples, and a discriminator to discriminate the fake samples from the real ones. The goal of adversarial learning is to learn a generator to generate samples that the discriminator cannot really tell they faked or no. Pfohl *et al.*[54] applied adversarial learning for developing an "equitable" risk prediction model for atherosclerotic cardiovascular disease (ASCVD) with EHR. They used the generator to build the risk predictor and discriminator to enforce equalized odds for the predicted risks across different protected groups.

4.2.3 Other Learning Strategies

Another closely related topic is interpretable learning[82], as interpretable models can allow the decision makers to better understand why certain predictions are made and make necessary modifications. Recent work at the FICO Data Science Lab has shown that interpretable neural networks can help uncover and eliminate data biases in models. Even in cases where the data is deliberately biased toward one subset of the population over another, the method minimizes the pickup of signals that are biased toward the

core relationship[83]. Similar argument has also been made by Rudin[84] that interpretable models are more preferred in high stakes decision making scenarios such as healthcare than black-box models.

Independent learning is another bias mitigation strategy which trains a machine learning model for each protected group[85]. However, this may sacrifice the training data sample sizes and reduce the model performance[85]. Gao and Cui[85] introduced a transfer learning approach to align the sample distributions across different protected groups. They demonstrated their method can achieve improved performance in underrepresented groups and effectively reduce disparity with cancer multiomics data.

4.3 Post-processing

The post-processing approach treats off-the-shelf predictors as black boxes and achieves fairness through adjustment of their predictions. For example, Hardt et al. [50] proposed equalized odds post-processing and calibrated equalization odds post-processing, which aims to solve for the probabilities of changing output labels to achieve the equalized odds objective. Kallus *et al.*[86] proposed to adjust the risk scores of the instances in the disadvantaged group with a parameterized monotonically increasing function to minimize the performance disparity. Cui et al.[87] proposed to adjust the ranking order of the samples across different protected groups according to their predicted scores with a dynamic programming procedure to achieve fairness without sacrificing prediction accuracy. One practical challenge for post-processing methods is that the involved adjustments are typically not explainable. Pan et al.[88] proposed a causal analysis approach that can quantitatively attribute algorithm performance disparity onto different causal decision paths, so that the paths with large contributions can be removed as post-processing.

In practice, these three types of methods work at different stages of a machine learning pipeline: pre-processing manipulates the data through sampling or weighting before building the model, in-processing enforces fairness constraints during model building, and post-processing makes adjustments after the model was built. Different strategies have different assumptions, therefore it is challenging to have a golden standard. A recent research from Park et al.[89] compared different risk mitigation methods in the context of postpartum risk prediction and found that these methods could indeed reduce bias but different methods can lead to different results. Therefore the practitioners should try to test different approaches and evaluate their impact in the particular context they were applied to.

5 Popular Software Libraries

We summarize existing popular algorithmic fairness research software libraries in Table 3.

6 Conclusions

In this review, we summarized the current research on algorithmic fairness in computational medicine. We first described the three types of computational bias: data bias, measurement bias, and model bias. Then we presented the fairness quantification metrics that are used in various literature. Additionally, we introduced three types of bias mitigation methods, namely, pre-processing, in-processing and post-processing, and listed the popular software libraries and tools for bias evaluation and mitigation. Fairness is not just the result of rigorous and thoughtful research, but rather the social and political processes needed to advance health equity[99]. With machine learning and artificial intelligence models gaining more and more attentions, we should be aware of these issues when designing the models and appropriately mitigate them. To help achieve this goal, we further list some probably encountered directions or open questions in Table 4.

Table 3. Popular library for fairness research

Project Name	Developer	Description
FairMLHealth[90]	KenSci	Tools and tutorials for evaluating bias in healthcare machine learning.
AIF360[91]	IBM	Fairness metrics for datasets and machine learning algorithms, interpretation of the metrics, and approaches for reducing bias in datasets and models. It is available in both Python and R.
Fairlearn[92]	Microsoft	A Python package to evaluate fairness and mitigate any observed inequities. Fairlearn includes mitigation algorithms and metrics for model evaluation. It also contains Jupyter notebooks with examples of Fairlearn usage.
Fairness-comparison[93]	Sorelle *et al.*	Compare fairness-aware machine learning techniques. It aims to facilitate benchmarking of fairness-aware machine learning algorithms.
MEASURES[94]	Cardoso *et al.*	A benchmark framework for assessing discrimination-aware models.
Fairness Indicators[95]	Google	A suite of tools built on top of TensorFlow Model Analysis that enable regular evaluation of fairness metrics in product pipelines.
ML-fairness-gym[96]	Google	A general framework for studying and exploring long-term equity effects in carefully constructed simulation scenarios where learning subjects interact with the environment over time.
themis-ml[97]	Niels Bantilan	A Python library built on top of pandasand sklearn that implements fairness-aware machine learning algorithms.
FairML[98]	Julius Adebayo	A Python toolkit for auditing machine learning model deviations.

Contributors

All authors read and approved the final version of the manuscript. JX drafted the manuscript. FW made thorough revisions to the draft. YX performed an initial literature review on computational bias. WW, YN and ES performed the literature review and data abstraction on bias mitigation methods. All authors contributed to the writing and editing of the manuscript. JB and FW conceived the idea.

Acknowledgements

FW is supported by NSF 1750326, NIH R01AG076234, R01MH124740 and RF1AG072449. YX is supported by CORONAVIRUSHUB-S-21-00188. YN is supported by NSF 1948432 and 2047843. WHW is supported by NSF 2029038 and 2135988. JB is supported by NIH R01AG076234, R01CA246418, R21CA253394, R21AG068717, and R21CA245858.

Declaration of Competing Interests

None declared.

References

1. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542**, 115–118 (2017).

2. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).

Table 4. Open questions and future directions.

Open Questions	Description
Data collection	As data is the source for building machine learning models, it is critical to be aware of the potential bias and improve the diversity and inclusiveness during the data collection process.
Multiform fairness	Different types of fairness are sometimes incompatible. For example, a model could be fair for equal positive and negative predictive values, but unfair for equalized odds (and vice versa). It is important to understand which types of fairness are achievable under which scenarios.
Algorithm explainability	Explainable models can reveal how a machine learning algorithm works and thus potentially alleviate decision bias. However, on the other hand, interacting with incorrect recommendations paired with explanations that contain limited but easily interpretable information can adversely affect the clinician's treatment choices[100]. Understanding such interaction between algorithm explainability and bias is important for medical machine learning.
Model generalization	Fairness in machine learning goes beyond preventing models from harming protected populations. It can also help focus care where it is really needed. The data used to develop the model may not be generalized to the data used during the deployment of the model (training-serving skew)[10]. Thus, besides model design and evaluation, fairness should also be incorporated into the scenario where the model is going to be deployed[101].

Search strategy and selection criteria

We searched PubMed and Google Scholar from inception of the database to Jul 30, 2021, for research articles using the search terms ("bias" OR "disparity" OR "fairness" OR "fair" OR "inequality" OR "equality") AND ("machine learning" OR "artificial intelligence") AND ("medical" OR "medicine" OR "healthcare") in English. We independently reviewed the title and abstracts for inclusion. We also reviewed the reference lists of eligible texts.

3. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. medicine* **24**, 1716–1720 (2018).
4. Wang, F. & Preininger, A. Ai in health: state of the art, challenges, and future directions. *Yearb. medical informatics* **28**, 016–026 (2019).
5. Gijsberts, C. M. *et al.* Race/ethnic differences in the associations of the framingham risk factors with carotid imt and cardiovascular events. *PLoS One* **10**, e0132321 (2015).
6. Zou, J. & Schiebinger, L. Ai can be sexist and racist—it's time to make it fair (2018).
7. Kadambi, A. Achieving fairness in medical devices. *Science* **372**, 30–31 (2021).
8. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. & Valley, T. S. Racial bias in pulse oximetry measurement. *New Engl. J. Medicine* **383**, 2477–2478 (2020).
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
10. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Annals internal medicine* **169**, 866–872 (2018).

11. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* **178**, 1544–1547 (2018).

12. Fletcher, R. R., Nakeshimana, A. & Olubeko, O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front. Artif. Intell.* **3**, 116 (2021).

13. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.* 1–8 (2021).

14. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).

15. Ng, J. H., Ye, F., Ward, L. M., Haffer, S. C. & Scholle, S. H. Data on race, ethnicity, and language largely incomplete for managed care plan members. *Heal. Aff.* **36**, 548–552 (2017).

16. Waite, S., Scott, J. & Colombo, D. Narrowing the gap: imaging disparities in radiology. *Radiology* **299**, 27–35 (2021).

17. Heckman, J. J. Sample selection bias as a specification error. *Econom. J. econometric society* 153–161 (1979).

18. Adamson, A. S. & Smith, A. Machine learning and health care disparities in dermatology. *JAMA dermatology* **154**, 1247–1248 (2018).

19. Lumaka, A. *et al.* Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clin. genetics* **92**, 166–171 (2017).

20. Nunan, D., Aronson, J. & Bankhead, C. Catalogue of bias: attrition bias. *BMJ evidence-based medicine* **23**, 21–22 (2018).

21. Chen, Z. *et al.* Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of alzheimer's disease. *NPJ digital medicine* **4**, 1–9 (2021).

22. Hernán, M. A. & Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *Am. journal epidemiology* **183**, 758–764 (2016).

23. Zang, C. *et al.* High-throughput clinical trial emulation with real world data and machine learning: A case study of drug repurposing for alzheimer's disease [preprint]. *medRxiv* (2022).

24. Jennions, M. D., Lortie, C. J., Rosenberg, M. S., Rothstein, H. R. *et al.* Publication and related biases. *Handb. Meta-analysis Ecol. Evol.* 207–236 (2013).

25. Dickersin, K. & Min, Y.-I. Nih clinical trials and publication bias. *The Online journal current clinical trials* 4967–words (1993).

26. Scherer, R. W. *et al.* Full publication of results initially presented in abstracts. *Cochrane Database Syst. Rev.* (2018).

27. Raynaud, M. *et al.* Covid-19-related medical research: a meta-research and critical appraisal. *BMC Med. Res. Methodol.* **21**, 1–11 (2021).

28. Coggon, D., Barker, D. & Rose, G. *Epidemiology for the Uninitiated* (John Wiley & Sons, 2009).

29. Vasudevan, V., Gnanasekaran, A., Sankar, V., Vasudevan, S. A. & Zou, J. Disparity in the quality of covid-19 data reporting across india. *BMC public health* **21**, 1–12 (2021).

30. Glen, S. Response bias: Definition and examples. *From StatisticsHowTo.com: Elementary Statistics for the rest of us!* https://www.statisticshowto.com/response-bias/.

31. Paulhus, D. L. Measurement and control of response bias. *Meas. personality social psychological attitudes* (1991).

32. van den Akker, M., Buntinx, F., Metsemakers, J. & Knottnerus, J. Morbidity in responders and non-responders in a register-based population survey. *Fam. practice* **15**, 261–263 (1998).

33. of Bias Collaboration, C., Spencer, E., Brassey, J., Mahtani, K. *et al.* Recall bias [catalogue of bias 2017]. *Retrieved Oct.* **20**, 2019 (2017).

34. Wakefield, A. J. *et al.* Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (1998).

35. Andrews, N. *et al.* Recall bias, mmr, and autism. *Arch. disease childhood* **87**, 493–494 (2002).

36. Mahtani, K., Spencer, E. A., Brassey, J. & Heneghan, C. Catalogue of bias: observer bias. *BMJ evidence-based medicine* **23**, 23–24 (2018).

37. Hróbjartsson, A. *et al.* Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *Bmj* **344** (2012).

38. Brooks, C. N., Talmage, J. B. & Mueller, K. Subjective, objective, or both? In *Guides Newsletter*, vol. 17 (2012).

39. Morgenstern, J. Bias in medical research. https://first10em.com/bias/.

40. Hróbjartsson, A. *et al.* Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *Cmaj* **185**, E201–E211 (2013).

41. Hooker, S. Moving beyond "algorithmic bias is a data problem". *Patterns* **2**, 100241 (2021).

42. Carbonell, J. G., Michalski, R. S. & Mitchell, T. M. An overview of machine learning. *Mach. learning* 3–23 (1983).

43. Chen, I. Y., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3543–3554 (2018).

44. Ewald, B. Post hoc choice of cut points introduced bias to diagnostic research. *J. clinical epidemiology* **59**, 798–801 (2006).

45. Wong, A. *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Medicine* **181**, 1065–1070 (2021).

46. Ahmad, M. A., Patel, A., Eckert, C., Kumar, V. & Teredesai, A. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3529–3530 (2020).

47. Verma, S. & Rubin, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7 (IEEE, 2018).

48. Kusner, M., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. *Adv. Neural Inf. Process. Syst. 30 (NIPS 2017) pre-proceedings* **30** (2017).

49. Calders, T., Kamiran, F. & Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18 (IEEE, 2009).

50. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Adv. neural information processing systems* **29** (2016).

51. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226 (2012).

52. Luong, B. T., Ruggieri, S. & Turini, F. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510 (2011).

53. Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P. & Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, vol. 1, 2 (2016).

54. Pfohl, S. *et al.* Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 271–278 (2019).

55. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In *International conference on machine learning*, 325–333 (PMLR, 2013).

56. Will Fleisher, W. What's fair about individual fairness? In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (ACM, 2021).

57. Lewis, D. Causation. *The journal philosophy* **70**, 556–567 (1974).

58. Baron, J. & Hershey, J. C. Outcome bias in decision evaluation. *J. personality social psychology* **54**, 569 (1988).

59. Dawid, A. P. Causal inference without counterfactuals. *J. Am. statistical Assoc.* **95**, 407–424 (2000).

60. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

61. Kearns, M., Neel, S., Roth, A. & Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572 (PMLR, 2018).

62. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **33**, 1–33 (2012).

63. Calders, T. & Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data mining knowledge discovery* **21**, 277–292 (2010).

64. Zafar, M. B., Valera, I., Gomez Rodriguez, M. & Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180 (2017).

65. Kamishima, T., Akaho, S. & Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650 (IEEE, 2011).

66. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. neural information processing systems* **27** (2014).

67. Tang, Z. & Zhang, K. Attainability and optimality: The equalized-odds fairness revisited. (2020).

68. Zhang, S., Zhang, C. & Yang, Q. Data preparation for data mining. *Appl. artificial intelligence* **17**, 375–381 (2003).

69. Kamiran, F. & Calders, T. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, 1–6 (Citeseer, 2010).

70. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. artificial intelligence research* **16**, 321–357 (2002).

71. Mukherjee, M. & Khushi, M. Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **4**, 18 (2021).

72. Han, H., Wang, W.-Y. & Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887 (Springer, 2005).

73. Caliendo, M. & Kopeinig, S. Some practical guidance for the implementation of propensity score matching. *J. economic surveys* **22**, 31–72 (2008).

74. Nilsson, A. *et al.* Reweighting a swedish health questionnaire survey using extensive population register and self-reported data for assessing and improving the validity of longitudinal associations. *Plos one* **16**, e0253969 (2021).

75. King, G. & Nielsen, R. Why propensity scores should not be used for matching. *Polit. Analysis* **27**, 435–454 (2019).

76. Borland, D., Zhang, J., Kaul, S. & Gotz, D. Selection-bias-corrected visualization via dynamic reweighting. *IEEE Transactions on Vis. Comput. Graph.* **27**, 1481–1491 (2020).

77. Kamishima, T., Akaho, S., Asoh, H. & Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, 35–50 (Springer, 2012).

78. Kamiran, F. & Calders, T. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, 1–6 (IEEE, 2009).

79. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *The New Engl. journal medicine* **378**, 981 (2018).

80. Cohen, I. G., Amarasingham, R., Shah, A., Xie, B. & Lo, B. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Heal. affairs* **33**, 1139–1147 (2014).

81. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. & Tygar, J. D. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 43–58 (2011).

82. Wang, F., Kaushal, R. & Khullar, D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Annals internal medicine* **172**, 59–60 (2020).

83. Zoldi, S. Fighting bias: How interpretable latent features remove bias in neural networks. https://www.fico.com/blogs/fighting-bias-how-interpretable-latent-features-remove-bias-neural-networks (2021).

84. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

85. Gao, Y. & Cui, Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat. communications* **11**, 1–8 (2020).

86. Kallus, N. & Zhou, A. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Adv. neural information processing systems* **32** (2019).

87. Cui, S., Pan, W., Zhang, C. & Wang, F. Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 207–217 (2021).

88. Pan, W., Cui, S., Bian, J., Zhang, C. & Wang, F. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1287–1297 (2021).

89. Park, Y. *et al.* Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA network open* **4**, e213909–e213909 (2021).

90. Allen, C. *et al.* fairMLHealth: Tools and tutorials for fairness evaluation in healthcare machine learning. https://github.com/KenSciResearch/fairMLHealth (2020).

91. Bellamy, R. K. E. *et al.* AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (2018).

92. Bird, S. *et al.* Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (2020).

93. Friedler, S. A. *et al.* A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 329–338 (2019).

94. L. Cardoso, R., Meira Jr, W., Almeida, V. & J. Zaki, M. A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 437–444 (2019).

95. Google. Tensorflow fairness indicators. https://www.tensorflow.org/responsible_ai/fairness_indicators/tutorials/Fairness_Indicators_Example_Colab.

96. Google. Ml-fairness-gym: A tool for exploring long-term impacts of machine learning systems. https://ai.googleblog.com/2020/02/ml-fairness-gym-tool-for-exploring-long.html (2020).

97. Bantilan, N. A library that implements fairness-aware machine learning algorithms. https://themis-ml.readthedocs.io/en/latest/.

98. Adebayo, J. FairML - is a python toolbox auditing the machine learning models for bias. https://github.com/adebayoj/fairml.

99. Sikstrom, L. *et al.* Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Heal. & Care Informatics* **29** (2022).

100. Jacobs, M. *et al.* How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl. psychiatry* **11**, 1–9 (2021).

101. Cui, S., Pan, W., Liang, J., Zhang, C. & Wang, F. Addressing algorithmic disparity and performance inconsistency in federated learning. *Adv. Neural Inf. Process. Syst.* **34** (2021).