Algorithmic Fairness in Computational Medicine

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

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

ABSTRACT

Machine learning models are increasingly adopted for facilitating clinical decision-making. However, recent research has shown that techniques may result in unintentional biases towards gender, race, and/or ethnicity. Such bias can lead to unexpected social implications, especially detrimental effects on the health and well-being of disadvantaged groups such as racial-ethnic minorities. In recent years, extensive studies on such algorithmic fairness have been conducted in machine learning in general. Yet, the understanding of how algorithmic bias will impact medicine and how to mitigate such bias still remains limited. The goal of this survey is to fill in this gap by providing a comprehensive review of algorithmic fairness within the scope of computational medicine, which aims at improving medicine with computational approaches. We first overview the three types of computational bias: data bias, measurement bias, and model bias. Then we summarize the fairness quantification metrics that are used in various literature. Finally, we present two types of bias mitigation methods, namely, data-based and model-based ones. We also summarize the popular software libraries and tools for bias evaluation and mitigation, followed by open questions and future directions.

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.

1 Introduction

The recent years have witnessed a surge of interests on development and deployment of machine learning algorithms in healthcare. These models can derive insights from massive health data and have demonstrated promising performance in a diverse set of settings 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], etc.

There have been growing concerns that machine learning models may result in unintended decision disparities on racial and ethnic minorities through the model itself and the data used to train them. For example, associations between Framingham risk factors and cardiovascular events have been shown to be significantly different across race and ethnicity groups[4]. Algorithms that can help diagnose various neurological diseases through video streams analysis to measure the body’s spontaneous blink rate, *e.g.*,

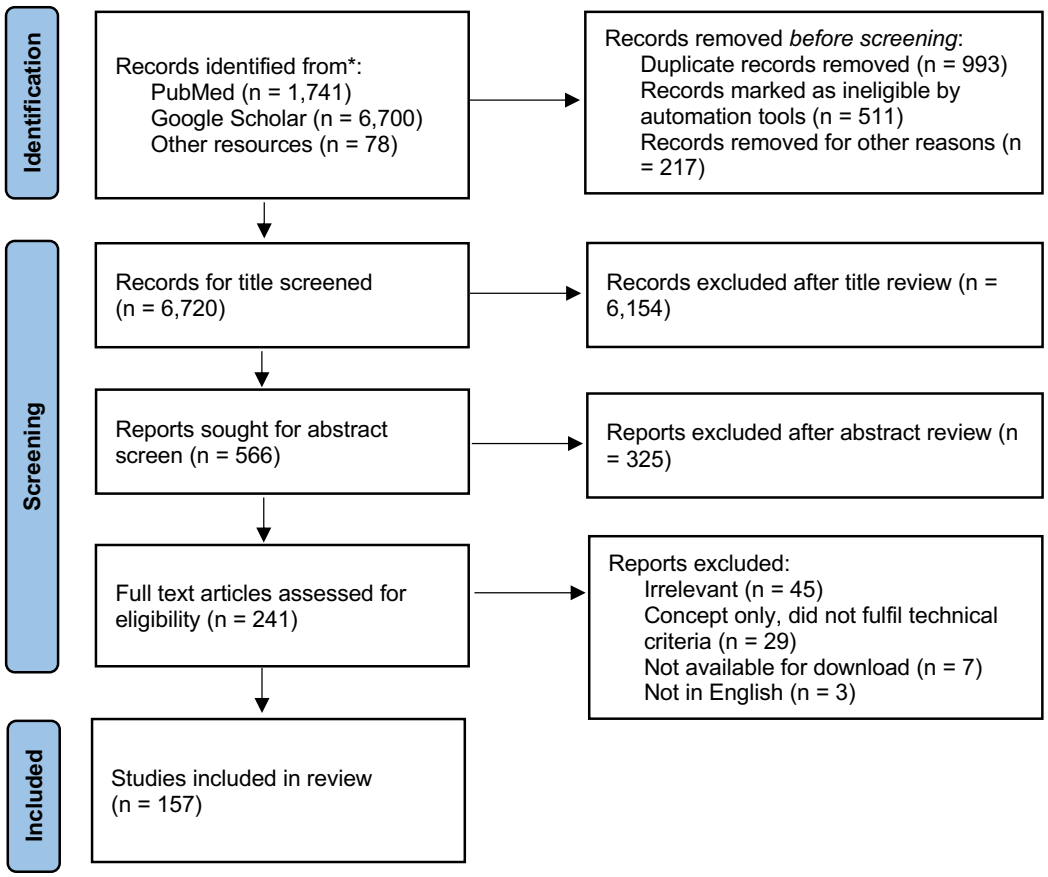

Figure 1. PRISMA flow diagram: disparity and fairness in computational medicine.

Parkinson's disease[5] and Tourette syndrome[6], have been found to be particularly challenging in detecting blinks for Asian individuals[7]. In another example, undiagnosed silent hypoxemia, which can be detected by pulse oximetry using light to monitor vital signs, occurred approximately three times more frequently in black patients due to the fact that dark skin responds differently to those light wavelengths [8]. In these cases, the implicit bias in the software system may create or exacerbate health disparities[7].

Therefore, to achieve the goal of building "a society in which all people live long, healthy lives," and "achieve health equality, eliminate inequalities and improve health for all groups" as has been envisioned in the Healthy People 2020 Vision Statement[9], it is crucial to be aware of the potential bias and disparity when developing machine learning models for helping with healthcare, understand their causes, and mitigate them. This review aims to achieve this goal by providing an overview of the existing literature studying the sources of bias and disparity in computational medicine, their quantification metrics, and mitigation strategies. The PRISMA diagram of the literature reviewed in this paper is shown in Fig. 1.

Difference with Existing Reviews. There have been several existing reviews on bias and fairness in machine learning. Specifically, Mehrabi et al.[10] investigated different real-world applications that have demonstrated biases in various ways, listed different sources of biases, and built a taxonomy of machine learning related fairness definitions for researchers to refer to. Rajkomar et al.[11] described how model design, biases in data, the interactions of model predictions with clinicians and patients may exacerbate healthcare disparities, as well as principles of distributive justice, and guided clinicians as to when they should prioritize each principle. Verma et al.[12] collected popular definitions of fairness for classification problems, and summarized the rationales behind these definitions. Chen et al. [13] outlined

how algorithmic biases (*e.g.*, image acquisition, genetic variation) arose in current clinical workflows, what healthcare disparities did they cause, and further reviewed strategies to mitigate bias. Fletcher et al.[14] described three basic criteria (*i.e.*, Appropriateness, Fairness, and Bias) that can help with the evaluations of machine learning and AI systems, with the goal of providing proper guidance in the context of global health. Mhasawade et al.[15] 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. Quy et al.[16] summarized the popular datasets used for the research of fairness-aware machine learning. Jones et al.[17] systematically compared the popular fair machine learning algorithms for supervised classification. Different from above works which focus either on bias and fairness in machine learning techniques or on eliminating disparities in health care through broad approaches such as adjusting policy, this survey examines where and how biases arise in computational medicine and provides specific examples, and how to mitigate them by machine learning engineers.

2 Computational Bias

In this section, we summarize common computational biases in the field and categorize them into *dataset bias*, *measurement bias*, and *algorithm bias*, as well as present examples of where they occur.

2.1 Dataset Bias

Machine learning models are learned from training data. Ideally, such training data should be representative of the real world situations. For example, some studies found that individuals from vulnerable groups, including those with lower socioeconomic status, with psychosocial problems, and immigrants, were more likely to go to multiple institutions or healthcare systems for treatment[18,19]. In addition, patients of low socioeconomic status may receive fewer diagnostic tests and medications for chronic diseases and have limited access to health care[18]. One consequence is that there may not be enough information in these patients' EHRs to meet the disease definition in clinical decision support tools, which could trigger early intervention or be identified only if their disease becomes more severe. The EHRs may also fail to collect data on factors related to improving the health of these subgroups, such as housing or transportation difficulties[19]. Below we've listed several possible and common causes of dataset bias in the medicine field.

2.1.1 Sample bias

Sample bias, also known as selection bias, occurs when a dataset does not reflect the realities of the environment in which a model will run[20]. Individuals, groups, or data are selected for analysis, but appropriate randomization cannot be achieved to ensure that the sample obtained is representative of the population to be analyzed. Or the cohort is default to traditional or easily measured groups, regardless of other potential protected groups or granularity levels (for example, whether sex is recorded as male, female, or other or more fine-grained categories). Another example of this is certain facial recognition systems trained primarily on images of white men. These models have considerably lower levels of accuracy with women and people of different ethnicity. For example, Face2Gene which can be used to recognize Down syndrome, showed a better recognition rate for Down syndrome in Caucasian (80%) compared to African (36.8%)[21]. The results predicted by the model do not take into account the correct representation of features in the population[22].

2.1.2 Allocation Bias

Allocation bias is a type of selection bias and is relevant to clinical trials of interventions[23]. If there are systematic differences between participants in how they are assigned to treatment groups and comparison

groups in a clinical trial, then this is called allocation bias[?]. Allocation biases can arise if researchers know or predict which intervention the next eligible participant should receive. This knowledge may influence how researchers approach potential eligible participants and how they assign them to different groups so that subjects with a good prognosis (*i.e.*, expected good outcomes and treatment responses) are selected for one of these groups[23]. The random allocation is usually used to minimize the allocation bias. 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[23].

2.1.3 Attrition Bias

Attrition bias is a type of selection bias that results from systematic differences in the number of study groups and the way participants are dropped from the study[23]. Regardless of the exposure or intervention of the study, different rates of loss to follow-up in the exposure groups may change the characteristics of these groups, or the loss may be affected by the positive or adverse events of the exposures[23]. Thus it will increase the chances of Type I (false positive) and Type II (false negative) errors[24].

2.1.4 Publication bias

Study publication bias occurs when a study is published or not depending on its results[25]. Empirical studies consistently show that published studies are more likely to be positive or statistically significant than unpublished studies[26]. There is also other evidence that studies with no statistically significant results take longer to be published than studies with significant results, further skewering the evidence over time[27]. Publication bias may lead to an overestimation of treatment effectiveness; it is considered a threat to the validity of meta-analysis and may make existing evidence unreliable for decision-making[27].

2.2 Measurement bias

Measurement bias is a systematic measurement error (e.g. offset baseline) that skews all data[10]. This type of bias occurs when the data collected for training differs from that collected in the real world, or when faulty measurements result in data distortion. A good example of this bias occurs in image recognition datasets, where the training data is collected with one type of camera, but the production data is collected with a different camera. Measurement bias can also occur due to inconsistent annotation during the data labeling stage of a project. A recent example is the quality of COVID-19 data reporting across India. Vasudevan *et al.*,[28] presented a comprehensive assessment of the quality of COVID-19 data reporting done by the Indian state governments between 19 May and 1 June, 2020. Their results indicated a large disparity in the quality of COVID-19 data reporting across India. The COVID-19 Data Reporting Score (CDRS, ranging from 0–1) varies from 0.61 (good) in Karnataka to 0.0 (poor) in Bihar and Uttar Pradesh, with a median value of 0.26[28]. Below we list several common causes of measurement bias.

2.2.1 Response bias

Response bias, also known as survey bias, refers to a person's tendency to answer untrue or misleading questions in a survey[?]. For example, they may feel pressure to give socially acceptable answers. People tend to paint the best picture of themselves, which can affect the survey results. Response bias is common in psychology, especially when it comes to self-reporting such as personal traits, attitudes like racism or sexism, behavior like alcohol use or unusual sexual behaviors[29]. This is also sometimes known as self-reporting bias or social desirability bias. In addition, misleading questions can lead to biased answers. Respondents may not be aware that they are not answering the questions the way the researcher wants them to. The format of the questions or the nature of the previous questions may have an unnecessary influence on how a person responds to the survey[30].

The term "response bias" is used differently by different people. Some people defines response bias as a type of selection bias[31]. That's what happens when patients themselves take part in trials that end up with no representative sample. For example, the people who respond to surveys are often different from those who choose to ignore them.

2.2.2 Recall Bias

Recall bias is a measurement bias that is common during the data annotation phase of a project[32]. It occurs when similar types of data are labeled inconsistently, thus resulting in low accuracy. A participant may erroneously provide responses that rely on his/her ability to recall past events. For example, parents of autistic children with regressive symptoms who were diagnosed after the publicity alleging a link with measles, mumps, and rubella (MMR) vaccine tended to recall the onset shortly after MMR more often than parents of similar children who were diagnosed prior to the publicity[33]. This may lead to observing an entirely partially untrue association between MMR and autism. Recalling events of interest can be particularly difficult when they happened a long time ago.

2.2.3 Observer Bias

The observer bias is a broader bias that occurs when the method or process of observing and recording information for a study leads to systematic discrepancy from the truth[34]. Biases in recording objective data may result from a lack of training in the use of measuring equipment or data sources, or unchecked bad habits[35]. In fact, this often happens because the study "sees what she wants to see". For example, in a chart review, it is common to have multiple observations recorded in a note. If researchers looked at aortic dissection, they might choose one description that said the pain was "severe" and ignore the others because that's what they expected to see. The interpretation of almost all medical tests is somewhat subjective, so observer bias is almost always possible[31].

2.3 Algorithm Bias

Another source of bias can be traced back to the algorithms themselves, where the mathematics of data processing against certain groups. Algorithmic bias is systematic and repeatable errors that create unfair outcomes. The algorithm bias is related to rigidity. High bias can cause the algorithm to adhere too strictly to the rules, ignoring the complexity of the data. Conversely, the high variance causes the algorithm to focus too much on data points that may be noisy. How to find the right balance between these two properties for a given model in a given environment is a key.

Preexisting is codification of already-present biases. For example, typical machine learning programs attempt to maximize the overall predictive accuracy of training data. If a particular set of individuals appear more frequently in the training data than others, the program will be optimized for those individuals because this improves overall accuracy. In addition, computer scientists evaluate algorithms based on "test" data sets, but these data sets are often random subsamples of the original training set and therefore may contain the same biases. Another one is technical which occurs due to the technical limitations of actually presenting the data. For example, the names at the top of a list will get preference, even if scored equally to those lower down. The third one is emergent which is development of new biases or new understandings of biases as technology develops, *e.g.*, growing popularity of audio-books in favor of print disadvantages the deaf population.

3 Fairness Metrics

The laws often use two different concepts to assess the fairness of the decision-making process: disparate treatment and disparate impact[36]. A decision process suffers from disparate treatment if it is (in part)

based on the sensitive nature of the subject. If its results disproportionately hurt (or benefit) people with certain sensitive attribute values (e.g., women, blacks), then it has a disparate impact. In this section, we first discuss a clinical application where machine learning may harm protected groups by being inaccurate, diverting resources, or worsening outcomes, especially if these patients are not considered in modeling[11]. We then use this application to explain the commonly used fairness metrics, as shown in Table 2. Notations are explained in Table 1.

3.1 A Case Study: Intensive Care Unit Monitoring

A common area of predictive modelling research has focused on creating a surveillance system – for example, alerting rapid response teams to high-risk inpatients who have deteriorated and need to be transferred to the intensive care unit (ICU) within 6 hours[37]. How could such a system inadvertently harm protected groups? In this study, we think of race (*i.e.*, Black and White) as a protected attribute.

To build the model, our hypothetical investigator collected historical records of patients with and without clinical deterioration. The model serves as a "diagnostic test" for the risk of transfer to the ICU. However, if too few black patients were included in the training data, the model may be inaccurate for them. For example, it may be less sensitive and miss more patients at risk of deterioration. Black patients may be harmed if clinical teams come to rely on alerts to identify high-risk patients without realizing that the prediction system underdetects this group of patients[38]. If the model has a lower positive predictive value for Blacks, this may also hurt them because it triggers bias generalization alarms, and fatigue clinicians may learn to discount or "dismiss" alarms for Blacks because they are more likely to be false positives[11].

3.2 Fairness Metrics

Algorithmic fairness has begun to attract the attention of artificial intelligence, software engineering and legal researchers, with more than twenty different fairness metrics proposed in the past few years. In this section, while not exhaustive, we've listed the most common ones used in computational medical research. For various other definitions and in-depth discussion, we refer the reader to a recent survey[12].

Table 1. Notations used in this article.

Symbol	Description
$A \in \{0,1\}$	Binary sensitive attribute (*e.g.*, majority/minority)
$X \in \mathbb{R}^d$	Other observable attributes of any particular individual
U	Relevant latent attributes which are 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

3.2.1 Fairness through unawareness

Fairness through unawareness is consistent with disparate treatment, which requires to not include the sensitive attributes as features during training[41]. It is intuitive and easy to use. The basic limitation is that there may be many highly related characteristics (*e.g.*, proxies for sensitive attributes). Therefore, simply removing sensitive attributes is not enough.

3.2.2 Demographic parity

*D*emographic parity, also called *independence* or *statistical parity*, ensures that the overall proportion of members of a protected group who receive a positive (negative) classification is the same as that of the

Table 2. Summary of Fairness Metrics

Type	Definition	In our case study
Fairness Through Unawareness	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[39] / Statistical Parity / Independence	The outcomes must be equal: $\mathbb{P}(\hat{Y}\|A=0) = \mathbb{P}(\hat{Y}\|A=1)$	Blacks and Whites must be transferred to ICU at equal rates
Equalized Odds[40] / 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 transferred to ICU, how many were correctly predicted to be positive) and false positive rates in Blacks and Whites are equal
Equal Opportunity[40]	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[41]	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 of transfer to ICU
Counterfactual Fairness[42]	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

population as a whole[39]. Although it is an important property, optimizing for demographic parity has a limited effect on clinical risk prediction, as it may prevent the model from taking into account relevant clinical characteristics related to sensitivities and outcomes, thereby reducing the performance of the model for all groups[43].

3.2.3 Equalized Odds

Unlike demographic parity, *equalized odds*[40] allows $\hat{Y}$ to depend on A, but only through the target variable Y. This encourages the use of features that are directly related to Y, rather than through A[40]. It can be seen from this that the false positive rates (FPR) and false negative rates (FNR) of all groups defined by A are equal under a fixed threshold T. Equalized odds is more appropriate in clinical settings than demographic parity because it does not necessarily preclude learning the best predictor where there is a real relationship between sensitive attributes and outcomes[40,43].

Both demographic parity and equality odds are known as group fairness measures because they focus on encouraging invariance of some property of a classifier relative to groups of sensitive attributes. While the calculations and reasoning are simple, optimizing these metrics may produce models that are discriminatory to structuralized subgroups within and across groups with sensitive attributes, constituting a fair form of *fairness gerrymandering*[43].

3.2.4 Equal Opportunity

In this case, the probability distribution of the predicted results depends on whether the event occurs or not and should be matched between groups of sensitive variables. Definition of *equal opportunity*[40] is stronger than definition of *equalized odds* because it means that all possible thresholds are equally likely to be met and therefore requires that all groups get the same ROC curve. This is desirable because it gives the end users the ability to freely adjust the model decision threshold without violating equalized odds[43].

3.2.5 Individual Fairness

At a high level, individual fairness requires that any two people who are similar at a given task should be classified similarly[41,44]. Although individual fairness may be able to address the issues that models are discriminatory over structured subgroups due to optimizing group fairness, the practical use of this concept is often limited due to the challenges of developing domain-specific similarity measures to encode the desired concept of fairness[41,43].

3.2.6 Counterfactual Measures

Counterfactual fairness is a possible way to interpret the causes of bias[42]. The definition is shown in Table 2. $\hat{Y}_{A \leftarrow a}$ is to be interpreted as the outcome of the predictor $\hat{Y}$ if A had taken value a. For the mathematical details and in-depth discussion, please refer to Kusner *et al.*[42]. This metric considers the predictor to be fair if its output remains unchanged when the protected attribute is flipped to its counterfactual value. In the social science literature, the closest counterpart is the counterfactual reasoning theory proposed by Lewis[45]. 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)[46]. In addition, it has been suggested that counterfactual reasoning may negatively affect the process of determining causation[47]. These critiques call into question the applicability of counterfactual measures in potential areas (*e.g.,* healthcare system) of prediction using machine learning, where the biases described above are often observed[39].

It turns out that any two of the three criteria in above definitions are mutually exclusive, except in non-degenerate cases. Clinical predictive models trained on potentially biased data may produce unfair results based on selected metrics[48]. There is no clear consensus on what definition should be used in each case. Researchers should choose the most appropriate method to quantify bias in a given research context.

4 Bias Mitigation

There are several well-known methods for correcting the model. The most common ones can be summarized as preprocessing[49], in-processing[50–53] and post-processing methods[?]. Next, we will present the state-of-the-art methods for each category.

4.1 Preprocessing

Preprocessing methods allow discrimination to be removed from data sets more effectively than simple methods, such as removing sensitive attributes from training data. The classifier is then learned on this cleaned data set.

4.1.1 Sampling

Sampling is a common method to ensure the datasets are balanced[54]. If the data set is large, the majority classes can be randomly sampled without losing important information to produce the same class size. However, if there is no redundant data, it is more common to oversample minority groups. Popular

algorithms, like synthetic minority oversampling technique (SMOTE)[55] or its variations, such as SMOTE-ENC[56], Borderline-SMOTE[57], adaptive synthetic sampling (ADASYN)[58], can be used for oversampling the imbalance data. However, in health applications, the data typically consists of questionnaire data and medical record fields, which may be binary and are difficult to synthesize without overfitting[14]. Chen *et al.*[59] argue that the fairness of the prediction should be evaluated in the context of the data, and that unfairness caused by inadequate sample sizes or unmeasured predictive variables should be addressed by data collection rather than by constraining the model. Based on the results of case studies that predicted income, mortality and review ratings, they found that data collection was often a way to reduce discrimination without sacrificing accuracy.

4.1.2 Reweighing

Reweighing is one of the pre-processing methods to mitigate the algorithm bias[49]. It is a simple but effective tool for minimizing bias. It applies different weights to each group-label combination according to conditional probability of label by protect attribute, making sensitive attribute and outcome label statistically independent. In more traditional statistical settings, techniques such as propensity score matching (PSM)[60] are used to adjust poorly sampled data. The method, sometimes called inverse probability of participation weighting, involves estimating the probability of an individual's participation, followed by an analysis of a re-weighted sample of participants[61]. For example, Nilsson *et al.*, reweighted the Swedish health questionnaire survey using a wide range of population-registered and self-reported data to assess and improve the validity of longitudinal associations[61]. However, because PSM aims to adjust the distribution of all variables simultaneously, it tends to increase imbalances and bias[62]. Borland *et al.*,[63] presented dynamic reweighting (DR) which goes beyond the solely communicating selection bias effects and enables selection bias correction in interactive visual analysis of high-dimensional data. They developed a prototype implementation of the DR approach with a visual analysis platform for cohort selection from medical data. The results suggest the DR approach is useful and understandable.

4.2 In-processing

A naive approach to fair decision making is simply to remove sensitive covariates from the model. However, if there is a dependency between the sensitive variable and other variables, the information of the sensitive variable will "leak" back into the decision. For example, there may be many highly correlated features (*e.g.*, neighborhood) that are proxies of the sensitive attribute (*e.g.*, race). Instead of concentrating mainly on identifying the discriminatory rules that are present in a dataset, and the specific subset of the data where they hold, some works focus on adjusting models through regularization or constraints[50–53].

4.2.1 Prejudice Remover

Prejudice remover is one of the in-processing methods to mitigate algorithmic bias. By adding a regularization term to the objective function, the protect attribute becomes less influential in the final prediction. In the learning process, the model tries to learn a classifier with independent constraints for future predictions, or minimizes the objective function and reduces the influence of sensitive attributes on the determination of the result class. For example, Kamiran and Calders *et al.*[64] first proposed the discrimination-aware classification and its extension to independent constraints to deal with unwanted dependency problems by "clearing away" dependencies on the data set before applying traditional classification algorithms. Some works have proposed variants of classifiers that enhance fairness, including discrimination-free naive-Bayes[50], a regularized version of logistic regression[52], and introducing constraints into the objective function of logistic regression, hinge loss classifier and support vector machine[51]. Although there is considerable interest in the ethical implications of implementing machine learning in healthcare[65], there is

relatively little research on the extent to which risk prediction models developed with EHR data satisfy formal fairness constraints.

4.2.2 Adversarial Learning

Satisfying equality of odds for a continuous risk score can be reduced to the problem of minimizing a divergence over each pair (A_i, A_j) of distributions. Adversarial learning[53] is suitable for this problem, and they provide a flexible framework for minimizing the divergence distribution parameterized by neural networks. Some related works [43] have demonstrated the benefits of using adversarial discriminators to enhance the classifier in order to align the distribution of predictions for satisfying fairness constraints. Specifically, Pfohl *et al.*[43] developed an "equitable" risk prediction model for atherosclerotic cardiovascular disease (ASCVD) using adversarial learning and a large observational cohort extracted from the EHR, reducing variability in error rates between groups. This is a very early effort to explore the extent to which risk prediction models constructed using the high-dimensional data of EHRs can achieve such formal fairness metrics, *e.g.*, equality of odds.

4.2.3 Other Learning Strategies

Another approach is ensemble methods. Two or more types of models can be combined to generate a hybrid classifier that performs well on unbalanced data sets. Approaches include bagging techniques (*e.g.*, bootstrap aggregation) to reduce variance and overfitting, and boosting methods to reduce bias. Popular ensemble methods include Adaboost and Extreme Gradient Boosting[66]. Although the ensemble method can handle challenging tasks well, such as cancer detection in radiology, it is less popular in diagnostic tasks due to its lack of interpretability and increased complexity[14].

Besides the aforementioned methods, another scheme is independent learning, which uses data from different ethnic groups separately to train an independent model for each ethnic group. Due to insufficient training data, this learning scheme is also prone to generate models with low prediction accuracy for minority groups with unfavorable data. Transfer learning can be used as a step toward a longer-term solution. Transfer learning could repurpose design parameters from task A (based on a balanced dataset) to task B (with an unbalanced dataset).

4.3 Post-processing

The post-processing approach treats off-the-shelf predictors as unexplained black boxes and implements fairness through outputting a function of the original prediction[67]. The main idea is to assess the classifier's confidence in each prediction and give favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups[68]. Classification with rejection is a typical way to prevent critical misclassification[69]. Basically, it involves retaining automatic categorization of items in cases where decisions are not considered fair enough. The rejected items are then processed by different classifiers or human experts. Recently, there are two popular methods proposed, including equalized odds post-processing and calibrated equalization odds post-processing[40]. Their main idea is to find the probabilities of changing output labels with equalized odds objective, where the former tries to solve a linear program and the latter to optimize calibrated classifier score outputs.

In practice, there is no doubt that the best way to strike a balance between accuracy and fairness is through optimization at training times, *i.e.*, in-processing. However, preprocessing and post-processing methods can maintain fairness without modifying the classifier. Such a feature is desirable when we do not have the ability to modify the classifier[70].

5 Popular Library

As concerns about bias have grown, a number of companies or research groups have launched various tools for scientific research and product development. Several commonly used ones are listed in Table 3.

Table 3. Popular library for fairness research

Project Name	Developer	Description
FairMLHealth	Allen *et al.*	Tools and tutorials for evaluating bias in healthcare machine learning.
AIF360[71]	IBM	A comprehensive set of fairness metrics for datasets and machine learning models, interpretation of these metrics, and algorithms for reducing bias in datasets and models. It is available in both Python and R.
Fairlearn[72]	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[73]	Sorelle *et al.*	Compare fairness-aware machine learning techniques. It aims to facilitate benchmarking of fairness-aware machine learning algorithms.
MEASURES[74]	Cardoso *et al.*	A benchmark framework for assessing discrimination-aware models.
Fairness Indicators	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	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	Niels Bantilan	A Python library built on top of pandasand sklearnthat implements fairness-aware machine learning algorithms.
FairML	Julius Adebayo	A Python toolkit for auditing machine learning model deviations.

6 Conclusions and Open Questions

In this survey, we review the current progress on bias in computational medicine including, but not limited to the healthcare field. We summarize where and how biases arise, most popular fairness metrics, bias mitigation strategies, popular libraries, and hope to provide useful resources for researchers to refer to. Besides the summarized general issues in fairness, we list some probably encountered directions or open questions in reducing bias in the healthcare area in the following.

- Diversity and inclusion: Because data is the driving force behind technology, and that unfairness caused by inadequate sample sizes or unmeasured predictive variables should be addressed by data collection rather than by constraining the model[59]. This begins with including a diverse and well-balanced study population in medical research.

- Tradeoffs: Different types of fairness are sometimes incompatible. A model may be fair for equal positive and negative predictive values, but unfair for equalized odds (and vice versa), but no model can satisfy both. The fairness of machine learning needs to be understood not only by machine

learning experts. It requires clinical and ethical reasoning to decide what type of fairness to use in each case, and what level of fairness is satisfactory.

- Transparency: For decisions involving human input, such as patient diagnosis, interpretable models are recommended to support human review and consensus. They allow users and certain parts of the internal system to be more transparent, explaining their decisions with a degree of detail. These interpretations are important to ensure the fairness of the algorithm, to identify potential biases/problems in the training data, and to ensure that the algorithm works as expected[75].

- Model deployment: 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)[11]. Thus, besides model design and evaluation, principles of distributive justice must also be incorporated into model deployment.

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).
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. 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).
5. Sandyk, R. The significance of eye blink rate in parkinsonism: a hypothesis. *Int. journal neuroscience* **51**, 99–103 (1990).
6. Karson, C. N., Kaufmann, C. A., Shapiro, A. K. & Shapiro, E. Eye-blink rate in tourette's syndrome. *J. Nerv. Mental Dis.* (1985).
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. Healthy people. *About healthy people* (2018).
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
11. 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).
12. Verma, S. & Rubin, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7 (IEEE, 2018).
13. Chen, R. J. *et al.* Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603* (2021).

14. 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).

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

16. Quy, T. L., Roy, A., Iosifidis, V. & Ntoutsi, E. A survey on datasets for fairness-aware machine learning. *arXiv preprint arXiv:2110.00530* (2021).

17. Jones, G. P. *et al.* Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986* (2020).

18. 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).

19. 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).

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

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

22. Alderton, G. Bias in medical devices (2021).

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

24. Gray, L. The importance of post hoc approaches for overcoming non-response and attrition bias in population-sampled studies. *Soc. psychiatry psychiatric epidemiology* **51**, 155–157 (2016).

25. 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).

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

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

28. Vasudevan, V., Gnanasekaran, A., Sankar, V., Vasudevan, S. A. & Zou, J. Disparity in the quality of covid-19 data reporting across india. *medRxiv* (2020).

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

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

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

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

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

35. 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).

36. Barocas, S. & Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.* **104**, 671 (2016).

37. Escobar, G. J. *et al.* Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. *J. hospital medicine* **11**, S18–S24 (2016).

38. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Informatics Assoc.* **24**, 423–431 (2017).

39. Gajane, P. & Pechenizkiy, M. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

40. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

41. 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).

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

43. 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).

44. Joseph, M., Kearns, M., Morgenstern, J., Neel, S. & Roth, A. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559* **1** (2016).

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

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

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

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

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

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

51. 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).

52. 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).

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

54. 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).

55. 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).

56. 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).

57. 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).

58. He, H., Bai, Y., Garcia, E. A. & Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 1322–1328 (IEEE, 2008).

59. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* (2018).

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

61. 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).

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

63. 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).

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

65. 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).

66. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Syst. Man, Cybern. Part C (Applications Rev.* **42**, 463–484 (2011).

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

68. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W. & Davis, J. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277* (2021).

69. Ni, C., Charoenphakdee, N., Honda, J. & Sugiyama, M. On the calibration of multiclass classification with rejection. *arXiv preprint arXiv:1901.10655* (2019).

70. Zhong, Z. A tutorial on fairness in machine learning. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

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

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

73. 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).

74. 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).

75. Gilpin, L. H. *et al.* Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89 (IEEE, 2018).