JAMIA Open, 2025, **8(1)**, ooae149
https://doi.org/10.1093/jamiaopen/ooae149
Review

Review

Algorithmic individual fairness and healthcare: a scoping review

Joshua W. Anderson, MS*[,1] and Shyam Visweswaran, MD, PhD[1,2]

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15213, United States, [2]Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15213, United States

*Corresponding author: Joshua W. Anderson, MS, Intelligent Systems Program, University of Pittsburgh, 4741 Baum Blvd, Pittsburgh, PA 15213, United States (jwa45@pitt.edu)

Abstract

Objectives: Statistical and artificial intelligence algorithms are increasingly being developed for use in healthcare. These algorithms may reflect biases that magnify disparities in clinical care, and there is a growing need for understanding how algorithmic biases can be mitigated in pursuit of algorithmic fairness. We conducted a scoping review on algorithmic individual fairness (IF) to understand the current state of research in the metrics and methods developed to achieve IF and their applications in healthcare.

Materials and Methods: We searched four databases: PubMed, ACM Digital Library, IEEE Xplore, and medRxiv for algorithmic IF metrics, algorithmic bias mitigation, and healthcare applications. Our search was restricted to articles published between January 2013 and November 2024. We identified 2498 articles through database searches and seven additional articles, of which 32 articles were included in the review. Data from the selected articles were extracted, and the findings were synthesized.

Results: Based on the 32 articles in the review, we identified several themes, including philosophical underpinnings of fairness, IF metrics, mitigation methods for achieving IF, implications of achieving IF on group fairness and vice versa, and applications of IF in healthcare.

Discussion: We find that research of IF is still in their early stages, particularly in healthcare, as evidenced by the limited number of relevant articles published between 2013 and 2024. While healthcare applications of IF remain sparse, growth has been steady in number of publications since 2012. The limitations of group fairness further emphasize the need for alternative approaches like IF. However, IF itself is not without challenges, including subjective definitions of similarity and potential bias encoding from data-driven methods. These findings, coupled with the limitations of the review process, underscore the need for more comprehensive research on the evolution of IF metrics and definitions to advance this promising field.

Conclusion: While significant work has been done on algorithmic IF in recent years, the definition, use, and study of IF remain in their infancy, especially in healthcare. Future research is needed to comprehensively apply and evaluate IF in healthcare.

Lay Summary

The use of algorithms in healthcare holds the potential to improve care delivery and reduce costs. However, these algorithms can sometimes reflect biases, leading to unfair treatment of individuals, particularly those from marginalized groups. This study reviews the concept of algorithmic individual fairness (IF), which ensures that similar individuals are treated similarly. The review identifies various philosophies and methods used to achieve IF and highlights how they can address biases in healthcare. While IF approaches are still in their early stages, they show promise in reducing disparities in healthcare. The findings emphasize the need for further research to enhance fairness in healthcare algorithms and ensure equitable treatment for individuals.

Key words: algorithmic fairness; individual fairness; health disparities; healthcare.

Introduction

Statistical and artificial intelligence (AI) algorithms have improved clinicians' ability to provide quality healthcare (in the biomedical literature, models are frequently referred to as algorithms. In this article, we use the terms algorithm and model interchangeably and preferably use the term algorithm). Such algorithms have accelerated healthcare discoveries, improved clinical decision-making, and lowered healthcare costs.[1] However, ethical concerns have been raised about the potential for such algorithms to exacerbate already-existing disparities among marginalized populations.[2]

Algorithmic fairness in healthcare is critical for ensuring equitable assessment and treatment of all individuals, regardless of their background. Various biases can creep into algorithmic development and application, affecting the fairness of such algorithms.[3] A range of protected attributes, factors that should not influence health, have been chosen because of legal mandates or organizational values.[4] Some common protected attributes include race, ethnicity, religion, national origin, gender, marital status, age, and socioeconomic status. Yet, several healthcare algorithms have been shown to be unfair, particularly across racial categories[5] and nearly 50 clinical algorithms are in use that include race, a key protected attribute, as an input variable.[6]

Unfairness in healthcare algorithms

Broadly speaking, biases in statistical and AI algorithms are caused by three factors: (1) unrepresentative data used for

Received: October 7, 2024; **Revised:** December 5, 2024; **Editorial Decision:** December 10, 2024; **Accepted:** December 12, 2024

algorithm development (data bias), (2) poor design in algorithm development (development bias), and (3) improper user—clinician or patient—interactions with the algorithm (interaction bias).[7] Biases in data are problems that arise from a variety of issues related to data collection and organization, including minority bias, missing data bias, informativeness bias, and selection bias.[8] Minority bias occurs when there are insufficient data from minority groups to develop an accurate algorithm (eg, the data includes far too few members of racial minority groups). Missing data bias occurs when data from minority groups are missing systematically, making it difficult to learn accurate statistical patterns (eg, members of racial minorities with limited access to healthcare have fewer electronic health record [EHR] data). Informativeness bias occurs when data and features used by an algorithm are less useful in a minority group (eg, detecting melanoma in patients with dark skin is more difficult than in those with light skin). Selection bias occurs when the data used to develop an algorithm is not representative of the population it will be deployed (eg, data from a single healthcare system may not be representative of other healthcare systems). Observational data, such as from EHRs that are increasingly used in developing algorithms, likely introduce more biases than carefully curated data from research studies due to inadequate documentation, ambiguous or varying definitions, and other systematic issues.[9–12]

Despite utilizing unbiased and representative data, algorithms may still manifest bias due to poor design in algorithm development. An example of such development bias issues is label bias.[7] Label bias occurs when algorithm development employs inconsistent labels, which do not mean the same thing for all individuals because they are an imperfect proxy. For example, racial bias was identified in an algorithm that predicted the future healthcare needs of patients because the data that was used in development employed medical cost as a surrogate for healthcare utilization.[13]

Interaction biases can occur when healthcare providers or patients interact with algorithms in ways that affect the algorithm's performance and fairness.[8] Automation bias is an example of clinician-interaction bias in which clinicians are unaware that an algorithm is less accurate for a specific group and place too much trust in it, accepting incorrect recommendations.[14] Privilege bias is a type of patient-interaction bias that occurs when algorithms are not available in settings where protected groups receive care, resulting in unequal distribution of algorithmic healthcare benefits.[15]

Measuring algorithmic fairness

To characterize algorithmic fairness, measures to assess fairness or, equivalently, bias are needed. Broadly speaking, 2 types of fairness metrics have been described: group and individual fairness (IF).[16]

Most of the literature focuses on the first notion of fairness, which is based on parity of statistical metrics across groups that differ in a protected attribute (eg, male and female groups). Compared to group fairness (GF), IF is less frequently described in the literature. Dwork et al[17] was the first to propose that "similar individuals should be treated similarly," with similarity between pairs of individuals defined in terms of a task-specific metric. According to Joseph et al, "less qualified individuals should not be favored over more qualified individuals," where quality is defined with respect to the true underlying label that the algorithm does not know.[18] Kusner et al[19] proposed a type of IF called counterfactual fairness. Counterfactual fairness is a principle for ensuring fairness that states that a decision is fair if it would be the same for an individual even if their protected attributes (eg, race, gender) were different in a counterfactual world. This means the algorithm's decision is not affected by group membership but only by the relevant characteristics of an individual. A glossary of terms is provided in Appendix S1.

Motivation

Our examination of IF was prompted in part by a rough parallel in the domain of predictive modeling. Statistical and AI approaches for training predictive algorithms can be broadly categorized into population-wide and patient-specific modeling that have rough parallels to GF and IF. The conventional predictive modeling approach in healthcare (and other areas) consists of learning a single algorithm from a database of individuals, which is then applied to decisions for each future individual. Such a model is called a population-wide model since it is intended to be applied to an entire population of future individuals and is optimized to have good predictive performance on average on all members of the population.[20] Patient-specific modeling, on the other hand, focuses on learning models that are tuned to the characteristics of the individual at hand, and such models are optimized to perform well for a specific individual.[20] Many patient-specific methods depend on assessing the similarity between individuals and hence use a similarity method. The canonical technique is the *k*-nearest neighbor method, which predicts the outcome in an individual based on a group of *k*-nearest individuals in the data. Other patient-specific methods train a model that is influenced by the characteristics of the patient at hand without using a similarity measure.[20–23]

While Dwork et al, Joseph et al, and Kusner et al provide notions of IF, ambiguity and heterogeneity persist, which continues to deter the deployment of real-world applications of IF in healthcare.[17–19] Furthermore, we uncovered no existing literature reviews focused on IF, which was the primary motivation for conducting this review.

Methods

We determined that a scoping review was appropriate due to the lack of existing literature reviews on this topic, our desire to broadly summarize the approaches to IF, and the potential role of IF in mitigating algorithmic bias, especially in healthcare. We followed the methodological framework by Arksey and O'Malley[24] and the Preferred Reporting Items for Systematic review and Meta-Analyses extension for Scoping Reviews.[24,25] We performed this scoping review in the following steps: (1) identify the research questions, (2) find relevant articles, (3) select articles, (4) extract data and themes, and (5) report the findings. We describe the first 4 steps here and report the findings in the "Results" section.

Identifying the research questions

The purpose of this study was to conduct a scoping review of the literature on IF to describe current approaches to IF and explore the potential role of IF methods in mitigating algorithmic bias. IF methods relevant to this review were defined according to the descriptions presented by Dwork et al, "similar individuals being treated similarly," Joseph et al, "less qualified individuals should not be favored over more

Table 1. Database queries for identifying relevant articles.

Database	Query date	Query	Filter	Number of records
PubMed	November 26, 2024	Algorithmic individual fairness [All fields]	2013-2024	192
ACM DL	November 26, 2024	[Title: algorithmic individual fairness] AND [Title: individual model fair- ness] AND [Title: fair ai] AND [E-Publication Date: (01/01/2013 TO 11/26/2024)]	N/A	1523
IEEE explore	November 26, 2024	((“All Metadata”:individual fairness) AND ((“All Metadata”:algorithmic) OR (“All Metadata”:model) OR (“All Metadata”:machine learning) OR (“All Metadata”: Artificial Intelligence)))	2013-2024	527
medRxiv	November 26, 2024	algorithmic individual fairness [All fields] [Subject area: Health Informatics]	2013-2024	256

qualified individuals,” and by Kusner et al, that individual decisions should “remain unchanged in a world where an individual’s protected attributes had been different in a causal sense.” Specifically, the review was conducted to address the gap in understanding of the characteristics of IF methods and their scope of use by addressing the following 2 research questions:

1) What notions of similarity are used in IF?
2) How is IF used to mitigate bias in algorithms?

Finding relevant articles

We searched for relevant articles and conference proceedings in four databases: PubMed, ACM Digital Library (DL), IEEE Xplore, and medRxiv. Because we wanted to retrieve as many relevant articles as possible, we devised a search strategy that prioritized recall over precision. Since the term “fairness” spans many disciplines in forms that are not algorithmic, we developed distinct search queries for each database to adjust for their relative sensitivities concerning these non-algorithmic notions of fairness. The search query for PubMed included the term “algorithmic individual fairness” appearing in the title or abstract. The search query for ACM DL included the term “algorithmic individual fairness” appearing in the title or abstract along with one or more of the following terms: “fair AI,” “individual model fairness.” The search query for IEEE Xplore included the term “individual fairness” appearing in the title or abstract along with one or more of the following terms: “algorithmic,” “model,” “machine learning,” or “artificial intelligence.” The search query for medRxiv included the term “algorithmic individual fairness” in all fields with the filter for papers classified in “Health Informatics.” The database-specific fields and queries we used are summarized in Table 1.

Selection of articles

We reviewed the titles and abstracts of unique articles obtained in the first step to identify articles for a full-text review (see Figure 1). We selected articles that studied IF methods and their uses based on the inclusion and exclusion criteria shown in Figure 1.

Because of the broad search criteria, many of the articles returned were not specifically about algorithmic IF. A number of articles that were found discussed differential privacy. Differential privacy and algorithmic fairness are closely related, and methods from differential privacy have been used to develop notions of algorithmic IF.[17] Despite this connection, we decided to leave these articles out of the final list because we preferred explicit notions of algorithmic IF. We identified a group of relevant articles for the review using the inclusion and exclusion criteria. Reasons for exclusion were recorded for the excluded articles. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-ScR flow diagram displays the number of excluded papers as well as the reasons for exclusion (see Figure 1).

Data extraction

We extracted information from articles and entered it into a spreadsheet for analysis. We recorded the year of publication, the similarity metric or methodology used, fairness mitigation methods, and the notion of IF for each article. In addition, we included a summary of each article’s findings. The extracted data were grouped into categories based on notions of similarity, types of IF, and types of mitigation; the categorizations and themes were iteratively refined based on discussions by the authors. A compilation of this information for each article is provided in Appendix S2.

Results

We identified 2498 articles through database searches and 7 articles through citations (see Figure 1). After de-duplication and title and abstract screening, 1591 articles were excluded, and due to the unavailability of full text, 7 more articles were excluded. This resulted in 893 articles for in-depth review, of which 868 were excluded after full-text review. A total of 32 articles (including 7 identified manually) were studied and analyzed in this review. Appendix S2 lists and summarizes the 32 articles chosen for inclusion in this scoping review.

Based on the 32 articles in the review, we identified several themes, including philosophical underpinnings of fairness, IF metrics, mitigation methods for achieving IF, implications of achieving IF on GF and vice versa, and applications of IF in healthcare.

Study characteristics

The publication years for the most articles (both with $n = 7$) were 2022 and 2023. Since the seminal article by Dwork et al[17] was published in 2012, the rate of publication has increased, indicating that this field is still in its infancy, and growth is expected to continue. In contradiction to this trend, publications in 2024 trended towards GF with a smaller number of publications on IF ($n = 2$), although we noticed a higher prevalence of articles applying GF specifically in healthcare. This indicates a growing awareness of fairness in health informatics, but the focus is still primarily on GF.

Similarity in counterfactuals ($n = 10$) was the most common type of similarity. The literature has begun to deviate

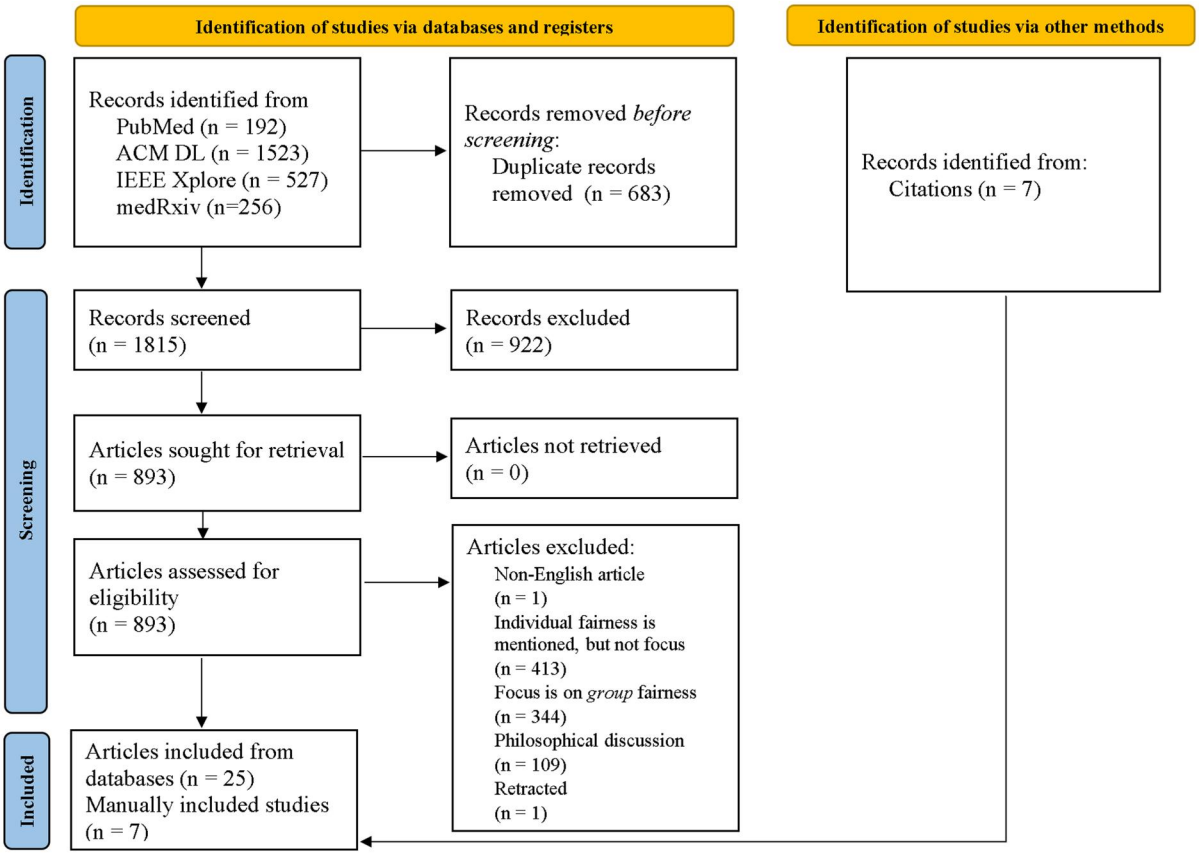

Figure 1. PRISMA diagram of the article screening process.[25]

from a domain-specific distance metric in favor of alternative methods for measuring fairness. Recent publications favor learned distance metrics ($n = 6$) and similarity relating to counterfactuals ($n = 10$). Being the original and most intuitive goal for IF, consistency ($n = 22$) was the most prevalent type of IF implied by authors. In-processing methods ($n = 21$) were the popular mitigation type. Our review only found a single mitigation method that we considered post-processing[26] (Table 2).

Philosophical corollaries of fairness

IF is motivated by the notion that similar individuals are treated similarly, which has been linked to achieving consistency in fairness literature. This notion is linked with Aristotle's conception of justice as consistency.[29] It is a desirable aspect of justice for judges to render accurate and consistent judgments for every individual and arrive at the same conclusion in identical cases. In the context of algorithms, consistency ensures that an algorithm's decisions are similar for similar individuals, regardless of group membership. Similarity-based or distance-based measures are commonly used to assess and achieve consistency-based fairness (see next section).

GF is motivated by the notion that groups of individuals should be treated similarly on average when they differ only in protected attributes (in this article, we use the terms protected and sensitive interchangeably and preferably use the term protected). This notion is linked to anti-discrimination laws, which prohibit discrimination against certain groups of people based on protected attributes such as race, sex, and age. Anti-discrimination in the context of algorithms ensures that an algorithm's decisions for an underprivileged group are similar, on average, to decisions for a privileged group. To assess and achieve anti-discrimination fairness, discrimination statistics that measure the average similarity of decisions across groups are used.[30]

In addition to consistency and anti-discrimination, a third concept is counterfactual fairness, which ensures an algorithm's decisions remain consistent across hypothetical scenarios where individuals' protected attributes are altered.[19] Typically, causal models that describe how changes in protected attributes affect decisions and other attributes of individuals are used to assess and achieve counterfactual fairness.

Fairness metrics

Measuring IF is typically based on a metric that measures similarity between individuals, and a common way to calculate similarity is by a metric or distance function that defines the distance between individuals as a non-negative real number. Dwork et al[17] defined an IF metric that assesses the fairness of an algorithm based on if it assigns the same decision to individuals with similar characteristics. The distance between two individuals, say a and b, is quantified by a distance measure $d(a, b)$, and IF is satisfied when

$$\sum_i |P(i|a) - P(i|b)| \leq d(a, b) \quad (1)$$

where $P(i|a)$ and $P(i|b)$ are the probabilities of decision i for individuals a and b, respectively. Similarly, Zemel et al[31] defined an IF metric called the consistency index which assesses the disparity between the decision assigned by an

Table 2. Philosophical corollaries of fairness.

Philosophical corollary	Description	Example article
(Anti) Discrimination	Ensures that algorithms treat different groups similarly. Discrimination statistics based on demographic parity or equal opportunity, eg, can be used to accomplish this.	Zhang et al[27]
Consistency	Ensures that algorithms treat similar individuals similarly, regardless of group membership. Similarity-based or distance-based measures can be used to accomplish this.	Bechavod et al[28]
Counterfactual	Ensures that algorithms would have made the same decision for an individual, regardless of their group membership, even if their attributes had been different. Causal models that explain how protected attributes affect decisions can be used to accomplish this.	Kusner et al[19]

algorithm to an individual and that individual's k-nearest neighbors. The consistency index is expressed as

$$1 - \frac{1}{n}\sum_i \left| \widehat{Y}_i - \frac{1}{k}\sum_{j \in \mathrm{kNN}(x_i)} \widehat{Y}_j \right|, \tag{2}$$

where n is the total number of individuals, $\widehat{Y}_i$ is the predicted output for individual i, and x_i is the feature vector of individual i.

Distance metrics are also used to measure counterfactual fairness. The counterfactual of an individual is a hypothetical scenario in which that individual's sensitive attributes differ.[32] Kusner et al[19] originally defined the counterfactual fairness metric and compared an algorithm's decision for an individual to their counterfactual. Counterfactual fairness is satisfied when

$$P\left(\widehat{Y}_a(U) = y | X = x, A = a\right) = P(\widehat{Y}_{a'}(U) = y | X = x, A = a) \tag{3}$$

where $\widehat{Y}_a$ and $\widehat{Y}_{a'}$ are the predicted decisions for an individual and their counterfactual, respectively, defined by sensitive attributes a, $a' \in$ A, latent variables U, and feature vector $x \in X$. Rather than simply flipping the value of the sensitive attribute(s) to represent the counterfactual, the causal effect of $A \rightarrow X$ is distributed across X_a to derive the features of the counterfactual, $X_{a'}$. Under this definition, predictions of $P\left(\widehat{Y}_a\right)$ are counterfactually fair if A is not a cause of $\widehat{Y}$.

Several types of IF metrics appear in the literature that iterate on the work of Dwork et al, Zemel et al, and Kusner et al. Most simply, generally defined similarity metrics from mathematics, such as Euclidean distance, cosine similarity, and Pearson correlation coefficient, are widely applicable across various domains and types of data. Domain-specific distance metrics are designed for specific types of data or fields, and they may not be widely applicable outside of their intended domain. For example, Rahman and Purushotham[33] use a derivative of cosine similarity adjusted specifically for hazard-based survival models by Keya et al.[34] Learned distance metrics are derived from the dataset on which they will be applied. Unlike pre-defined distance metrics, learned metrics adapt to the specific characteristics of the dataset (see Table 3). Additionally, counterfactual methods use a variety of unique methods to measure the difference between an individual and their counterfactual. Methods for counterfactual distance vary across each article.

Mitigation methods

In the context of creating fair algorithms, pre-processing, in-processing, and post-processing are three categories of methods to mitigate bias. Pre-processing methods adjust or transform the data to ensure balanced representation and remove discrimination. Resampling (adjusting the data to balance the representation of different groups), reweighting (assigning different weights to samples to counteract imbalances), and removing protected attributes (removing features like race, gender, or age that are protected and could lead to biased decisions) are examples some examples of pre-processing.[32,34,38,39] Articles related to pre-processing evaluated the effectiveness of methods in a wide range of metrics, from increased explainability[38] to problem-specific distance metrics.[32]

In-processing methods modify the training of the algorithm to incorporate fairness constraints or objectives directly. Regularization techniques (adding a fairness constraint or regularization term to the learning objective) and adversarial debiasing (using adversarial networks to learn representations that do not contain biased information about protected attributes). Post-processing methods adjust or transform the decisions or outputs of an algorithm after its training. Examples include calibration (adjusting predicted probabilities of decisions to reflect the true likelihoods of those decisions accurately) and threshold adjustments (changing decision thresholds for different groups to balance performance metrics) are some examples.[39–41] Most articles discussing in-processing for IF compare mitigated models to baseline models using IF fairness metrics (eg, consistency, discrimination, etc) and discuss the impact on relevant performance metrics (eg, true negative rate, accuracy, etc).

Although pre-processing and in-processing techniques were frequently employed in the articles we reviewed, post-processing for IF was only investigated by Petersen et al.[26] Petersen et al evaluated the post-processing by observing the trade-offs between IF and accuracy, IF and GF, and the distribution of violations of the IF constraint. The results emphasized the large disparities in GF because of the IF method, reiterating the consensus in the literature that these two types of fairness are orthogonal (Table 4).

The relationship between the two kinds of fairness

Several GF metrics are incompatible in that fairness cannot be achieved simultaneously on those metrics. The incompatibility of IF and GF metrics has received less attention. GF does not imply IF, and IF implies GF if the Wasserstein distance (distance between probability distributions) is small, that is, the distributions of similar individuals are relatively uniform across groups, which is uncommon in practice.[17,44]

Binns[45] discusses the trade-offs that arise when one type of fairness is preferred over another. When ignoring IF favoring GF, algorithms may make different decisions for identical individuals. Furthermore, emphasizing IF alone can lead to significant differences in group decisions.[45] According to Fleisher, optimizing IF alone does not guarantee GF. For

Table 3. Types of IF metrics.

Type of IF metric	Description	Example article
Generally defined distance metric	Predefined distance metrics that are widely applicable across various domains and types of data.	Ghadage et al[35]
Domain-specific distance metric	Predefined distance metrics that are designed for specific types of data or fields.	Rahman and Purushotham[33]
Learned distance metric	Distance metrics that are derived from the dataset on which they will be applied.	Hu and Rangwala[36]
Distance in counterfactuals	Various distance metrics are used to compare an individual to their counterfactuals.	Ma et al[37]

Table 4. Types of mitigation methods.

Type of mitigation	Description	Example article
Pre-processing	Adjust, transform, reweight, or augment data to ensure balanced representation and remove discrimination.	Zhang et al[42]
In-processing	Modify the training of the algorithm to incorporate fairness constraints or objectives directly.	Sharifi-Malvajerdi et al[43]
Post-processing	Adjust or transform the decisions or outputs of an algorithm after its training.	Petersen et al[26]

example, an algorithm that assigns a negative decision to every individual will satisfy IF but not GF.[46]

Applications in healthcare

Rahman and Purushotham, Cheng et al, Chien et al, Tal, Jun et al, and Zhou et al discussed the applications of IF in healthcare.[33,47–51] Rahman and Purushotham[33] describe an IF method for survival analysis to address the problem of censoring in clinical trials, particularly in underprivileged groups.[52] The authors demonstrated that their IF-based deep survival algorithms significantly reduced unfairness in censoring.

Cheng et al[47] created a framework for interviewing stakeholders to understand better their interpretations and notions of fairness in clinical predictive systems. Twelve participants were polled, and many of them were skeptical of IF. For example, one participant remarked, "I think it's tricky to compare things this way… It's hard to say." Although more participants favored GF, they disagreed on which GF measures were appropriate.

Chien et al[48] suggest that the traditional fixed-clinical trial method prevents beneficial modifications after trials begin, and AI methods can be employed to make trials fairer. According to the authors, optimizing for GF is less useful than optimizing for IF or counterfactual fairness for the problem of fairness in clinical trials, despite the advantage of GF methods being task-agnostic and less complex.

Tal[49] argues that an important cause of bias in healthcare algorithms is due to conflicting notions of problem definitions. For example, a statistical notion of bias and accuracy would claim that the two are orthogonal, allowing a model to be both biased and accurate. On the other hand, a clinician would argue that bias and accuracy are contradictory and cannot co-occur. Target specification bias, a particular case of this divergence in definitions, occurs when the notions of the decision variable by analysts and clinicians differ.[49,53] This occurs because a clinician expects to predict a decision for a patient if they were treated differently all else being equal (counterfactual), whereas most models predict similarly observed individuals with measured decisions. This issue is closely related to IF, implying that counterfactual fairness is a more accurate representation of the problem from the standpoint of a clinician.

Jun et al[50] applies fairness-aware causal analysis that links social determinants of health (SDoH) to EHRs to evaluate unfairness in methicillin-resistant staphylococcus aureus infection-related 30-day mortality. While not explicitly framed as counterfactual fairness, we include the paper since comparisons are made across baseline characteristics to discover unfairness related to SDoH, similar to methods for consistency. It is a rare example of an IF method applied in real-world EHR data.

Zhou et al[51] propose a novel rank similarity regularization method, Joint Correlation Learning with Rank Similarity Regularization, which improves fairness by enforcing consistency in predictions for both common and rare fetal brain age cases in highly imbalanced magnetic resonance imaging data. The proposed framework enhances the gestational age prediction model, with fairness adjustments specifically targeting under-represented cases.

Discussion

Given the relatively small number of articles we found (32 articles from 2013 to 2024), the first implication of our findings is that the definition, use, and study of IF remain in their infancy, especially in healthcare. However, since the seminal article on IF was published in 2012,[17] the rate of IF article publication has steadily increased, indicating that this field is likely to grow in the future. Only 6 articles described the use of IF in healthcare,[33,47–51] despite evidence that there is intense interest in measuring and mitigating bias in clinical risk calculators based on race, differential laboratory test reference ranges are recommended based on race, and differential therapy is recommended based on race.[6] This is most likely due to the infancy of the field of IF in general.

There is mounting evidence that due to the limitations of GF, alternative approaches to fairness, such as IF and counterfactual fairness, are needed.[17,19] One limitation of GF is that it may mask individual differences within a group,

whereas IF is more flexible and adaptable and can take individual features other than protected attributes into account.[45] Second, GF may disregard relevant features that are not protected at- tributes, whereas IF may lead to more accurate and fair outcomes by considering all relevant features.[31] Third, defining and measuring GF can be challenging when dealing with multiple groups with overlapping memberships or complex relationships.[48] IF is not limited in such circumstances.[17] While IF offers several advantages over GF, it also has limitations that need to be considered. One limitation of IF is that determining what constitutes "similar individuals" can be complex and subjective. Different contexts and tasks may require different definitions of similarity, making it challenging to achieve universal applicability. Second, IF methods that rely on learning similarity metrics from data are susceptible to encoding existing biases present in the data, which can perpetuate existing inequities.[46]

Our study had some limitations. By limiting our search to four databases, it is possible that articles relevant to this topic were not found and included in this review. It is also possible that relevant articles that used keywords other than those that were used were missed in our search. Our search results show that the number of articles in IF has increased steadily over the last decade. However, we did not investigate how IF metrics and definitions have evolved over this period of time.

Conclusion

This scoping review explored the breadth of algorithmic IF metrics and methods developed to achieve IF. We provide preliminary structure and grouping of varying ideas and strategies and describe current research relating to applications in healthcare. The articles that explored this topic showed that the definition, use, and study of IF remain in their infancy. Future research is needed to evaluate and apply IF to continue to have a real-world impact on reducing disparities in assessment and treatment in healthcare.

Author contributions

All authors made substantial contributions to the concept, design, and execution of this study.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

Research reported in this publication was supported by the National Institutes of Health under award number T15 LM007059 from the National Library of Medicine and under award number UL1 TR001857 from the National Center for Advancing Translational Sciences. It was also supported by a School of Computing and Information Predoctoral Fellowship to J.W.A.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

No new data are associated with this study.

References

1. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: *Artificial Intelligence in Healthcare*. NIH; 2020.
2. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4:31.
3. Bærøe K, Gundersen T, Henden E, Rommetveit K. Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health Care Inform*. 2022;29:1-6.
4. World Medical Association. Declaration of Geneva. 1983. https://www.wma.net/what-we-do/medical-ethics/declaration-of-geneva/decl-of-geneva-v1983/
5. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight- reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383:874-882.
6. Visweswaran S, Sadhu EM, Morris MM, Samayamuthu MJ. Clinical algorithms with race: an online database. medRxiv 2023, preprint: not peer reviewed.
7. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169:866-872.
8. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol*. 2024;42:3-15.
9. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. 2018;154:1247-1248.
10. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. *J Biomed Inform*. 2023;139:104269.
11. Konkel L. Racial and ethnic disparities in research studies: the challenge of creating more diverse cohorts. *Environ Health Perspect*. 2015;123:A297-A302.
12. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15:e1002683.
13. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453.
14. Anderson M, Anderson SL. How should AI be developed, validated, and implemented in patient care? *AMA J Ethics*. 2019;21:125-130.
15. Fiscella K, Williams DR. Health disparities based on socioeconomic inequities: implications for urban health care. *Acad Med*. 2004;79:1139-1147.
16. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. *Commun ACM*. 2020;63:82-89.
17. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Association for Computing Machinery; 2012:214-226.
18. Joseph M, Kearns M, Morgenstern JH, Roth A. Fairness in learning: classic and contextual bandits. *Adv Neural Inf Process Syst*. 2016;29:1-9.
19. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. *Adv Neural Inf Process Syst*. 2017;30:1-11.
20. Visweswaran S, Angus DC, Hsieh M, Weissfeld L, Yealy D, Cooper GF. Learning patient-specific predictive models from clinical data. *J Biomed Inform*. 2010;43:669-685.
21. Johnson A, Cooper GF, Visweswaran S. Patient-specific modeling with personalized decision paths. In: *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association; 2020:602.
22. Lengerich B, Aragam B, Xing EP. Learning sample-specific models with low-rank personalized regression. *Adv Neural Inf Process Syst*. 2019;32:1-11.

23. Visweswaran S, Ferreira A, Ribeiro GA, Oliveira AC, Cooper GF. Personalized modeling for prediction with decision-path models. *PLoS One*. 2015;10:e0131022.
24. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Social Res Methodol*. 2005;8:19-32.
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-SCR): checklist and explanation. *Ann Intern Med*. 2018;169:467-473.
26. Petersen F, Mukherjee D, Sun Y, Yurochkin M. Post-processing for individual fairness. *Adv Neural Inf Process Syst*. 2021;34:25944-25955.
27. Zhang P, Wang J, Sun J, et al. Automatic fairness testing of neural classifiers through adversarial sampling. *IEEE Trans Software Eng*. 2022;48:3593-3612.
28. Bechavod Y, Jung C, Wu SZ. Metric-free individual fairness in online learning. *Adv Neural Inf Process Syst*. 2020;33:11214-11225.
29. Schauer F. On Treating Unlike Cases Alike. 2018. Accessed November 2, 2024. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3183939
30. Thomas PS, Castro da Silva B, Barto AG, Giguere S, Brun Y, Brunskill E. Preventing undesirable behavior of intelligent machines. *Science*. 2019;366:999-1004.
31. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: *International Conference on Machine Learning*, PMLR; 2013:325-333.
32. Garg S, Perot V, Limtiaco N, Taly A, Chi EH, Beutel A. Counterfactual fairness in text classification through robustness. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery; 2019:219-226.
33. Rahman MM, Purushotham S. Fair and interpretable models for survival analysis. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2022:1452-1462.
34. Keya KN, Islam R, Pan S, Stockwell I, Foulds J. Equitable allocation of healthcare resources with fair survival models. In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM; 2021:190-198.
35. Ghadage A, Yi D, Coghill G, Pang W. Multi-stage bias mitiga- tion for individual fairness in algorithmic decisions. In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer; 2022:40-52.
36. Hu Q, Rangwala H. Metric-free individual fairness with cooperative con- textual bandits. In: *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE; 2020:182-191.
37. Ma J, Guo R, Zhang A, Li J. Learning for counterfactual fairness from observational data. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023:1620-1630.
38. Aggarwal A, Lohia P, Nagar S, Dey K, Saha D. Black box fairness testing of machine learning models. In: *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Association for Computing Machinery; 2019:625-635.
39. Caton S, Haas C. Fairness in machine learning: a survey. *ACM Comput Surv*. 2024;56:1-38.
40. Kang J, He J, Maciejewski R, Tong H. Inform: individual fairness on graph mining. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery; 2020:379-389.
41. Biswas S, Rajan H. Fairify: fairness verification of neural networks. In: *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE; 2023:1546-1558.
42. Zhang C, Huiyi Cen, S, Shah D. Matrix estimation for individual fairness. In: *International Conference on Machine Learning*. PMLR; 2023:40871-40887.
43. Sharifi-Malvajerdi S, Kearns M, Roth A. Average individual fairness: algorithms, generalization, and experiments In: Wallach H, Larochelle H, Beygelzimer A, d'Alch'e-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.; 2019.
44. Zhou W. *Group vs Individual Algorithmic Fairness*. PhD Thesis. University of Southampton; 2022.
45. Binns R. On the apparent conflict between individual and group fairness. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2020:514-524.
46. Fleisher W. What's fair about individual fairness? In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery; 2021:480-490.
47. Cheng H-F, Stapleton L, Wang R, et al. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2021:1-17.
48. Chien I, Deliu N, Turner R, Weller A, Villar S, Kilbertus N. Multidisciplinary fairness considerations in machine learning for clinical trials. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2022:906-924.
49. Tal E. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery; 2023:312-321.
50. Jun I, Ser SE, Cohen SA, et al. Quantifying health outcome disparity in invasive methicillin-resistant staphylococcus aureus infection using fairness algorithms on real-world data. In: *Pacific Symposium on Biocomputing 2024*. World Scientific; 2023:419-432.
51. Zhou R, Liu Y, Xia W, et al. Jocorank: joint correlation learning with ranking similarity regularization for imbalanced fetal brain age regression. *Comput Biol Med*. 2024;171:108111.
52. Schulman KA, Berlin JA, Harless W, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. *N Engl J Med*. 1999;340:618-626.
53. Li X, Wu P, Su J. Accurate fairness: improving individual fairness without trading accuracy. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. Association for Computing Machinery; 2023:14312-14320.

JAMIA Open, 2025, 8, 1–8
https://doi.org/10.1093/jamiaopen/ooae149
Review