Title Page

A Survey on Optimization and Machine Learning-Based Fair Decision Making in Healthcare

Zequn Chen[a]; Wesley J. Marrero[a*]

[a] Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

***Corresponding author:** Wesley J. Marrero

wesley.marrero@dartmouth.edu

Phone: (603) 646-3457

Address: 15 Thayer Dr, Hanover, NH 03755

https://orcid.org/0000-0002-7092-2292

Zequn Chen, M.S.; zequn.chen.th@dartmouth.edu; (603) 646-3457 ; 15 Thayer Dr, Hanover, NH 03755; ORCID: 0009-0000-0693-1250

Wesley J. Marrero, Ph.D.; wesley.marrero@dartmouth.edu; (603) 646-3457; 15 Thayer Dr, Hanover, NH 03755; ORCID: 0000-0002-7092-2292

Statements and Declarations

Acknowledgments The financial support for this study was provided in part by the Thayer Tuition Fellowship at Dartmouth College. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Conflicts of interest All authors declare they have no conflicts of interest. This manuscript is being submitted only to Health Care Management Science and will not be submitted elsewhere while under consideration. A preprint version of this manuscript is available on medRxiv but has not been published elsewhere, and should it be published in Health Care Management Science, it will not be published elsewhere without the permission of the editors.

Abstract

The unintended biases introduced by optimization and machine learning (ML) models are a topic of great interest to medical researchers and professionals. Bias in healthcare decisions can cause patients from vulnerable populations (e.g., racially minoritized, low-income, or living in rural areas) to have lower access to resources and inferior outcomes, thus exacerbating societal unfairness. In this systematic literature review, we present a structured overview of the literature regarding fair decision making in healthcare until April 2024. After screening 801 unique references, we identified 114 articles within the scope of our review. In our review, we comprehensively examine fair decision-making methodologies in healthcare by systematically identifying and categorizing biases within both data and models. Initially, we elucidate existing bias within healthcare decision making. Then, we present a range of fairness metrics drawn from different use cases, followed by analyzing and classifying bias mitigation strategies into pre-processing, in-processing, and post-processing techniques. We provide a broad conceptual overview and practical illustrations of each approach. Additionally, we examine emerging bias mitigation technologies that, though not yet applied in healthcare, show substantial promise for future integration. Our review aims to increase awareness of fairness in healthcare decision making and facilitate the selection of appropriate approaches under varying scenarios.

Keywords: Fairness, Decision making, Optimization, Machine learning, Systematic literature review

1. Introduction

An increasing number of healthcare researchers and practitioners are leveraging quantitative methodologies to improve decision making. These methodologies aim to achieve desirable resource allocation strategies, testing procedures, or treatment protocols. Optimization and machine learning (ML) models and algorithms have been proven to be extremely helpful in medical decision making and health policy [1–7]. For example, researchers have been applying optimization models to schedule patients based on their predicted no-show rates; and reinforcement learning has been used for treatment recommendations to ensure effective prescription and low mortality rates [8, 9]. However, such advanced decision-making methods can lead to inequitable outcomes as they may not give sufficient attention to underrepresented groups [10], such as those who are racially minoritized or low-income. Our survey presents a review of fair decision-making techniques, revealing how we can leverage optimization and ML approaches to ensure equitable access to healthcare.

The unintended biases introduced by optimization and ML are of special interest to practitioners and researchers [11–14]. For example, compared to White veterans, the medical policies generated by reinforcement learning often offer Black veterans fewer opportunities to receive cardiovascular screenings. Hispanic patients are disproportionately underdiagnosed by convolutional neural networks because they tend to have limited access to healthcare resources and the ML approach cannot perform satisfyingly with inadequate data [15]. Clinicians have utilized ML models in Warfarin dosing, which shows superior performance in European patients but unsatisfying outcomes in Asian patients [16, 17]. Such bias may cause decision makers to distribute fewer medical resources to racially minoritized subgroups.

1.1 General Trends in Fair Decision Making

Recent research in ML and optimization-based decision making in healthcare has increasingly centered on three complementary methodological paradigms: pre-processing, in-processing, and post-processing. *Pre-processing* techniques focus on transforming and rebalancing raw data before model training to mitigate historical biases and improve representation for underrepresented groups. For instance, methods such as fair data transformers and advanced natural language processing pipelines are being developed to ensure that the input data itself is less biased. *In-processing* methods embed fairness directly into the learning or optimization process by incorporating fairness penalties, constraints, or fairness considerations in the algorithm or model design. These methods include the adaptation of fair reward design in reinforcement learning, as well as formulation of optimization problems (e.g., mixed-integer or stochastic programming models) that inherently account for equitable decision criteria during model training. Finally, *post-processing* approaches adjust model outputs after training to correct any residual biases. Techniques like Laplacian smoothing and multi-accuracy adjustments are employed to refine decision policies, ensuring fair resource allocation and treatment recommendations across diverse patient populations. Collectively, these research avenues are advancing the state-of-the-art in achieving both high-performance and equitable decision making in healthcare.

1.2 Existing Systematic Reviews

Past literature reviews focus on the biases brought or perpetuated by ML in prediction settings. For example, Ahmad and collaborators show that ML-based predictions often yield fewer satisfying outcomes in underrepresented groups due to their insufficient data [18]. Mishler and coauthors reveal that predictors are sensitive to proportions of different populations, and incorporating fairness definitions can help avoid such issues [19]. In

contrast to previous reviews centering on prediction, we focus on fairness within the context of decision making. It is worth noting that our review has some overlaps with Smith et al.'s survey [14]. The main differences between our surveys are: 1) their work focuses on fairness in reinforcement learning exclusively, while our paper explores other in-processing techniques such as mixed-integer programming (MIP) and stochastic programming; 2) we review different pre-processing techniques, such as fair data transformers and natural language processing; and 3) we consider post-processing methods such as Laplacian smoothing and multi-accuracy approaches. The fair reinforcement learning methods cited in Smith et al. are included in our review for completion purposes [20–23]. However, we refer the interested reader to their review for an in-depth description of these works.

1.3 Contributions

Our review provides a comprehensive overview of fair decision-making approaches in healthcare. Its main contributions are: (1) We systematically identify and categorize biases affecting both data and models, summarizing multiple fairness metrics (and their limitations) through concrete examples; (2) We analyze the advantages and disadvantages of existing bias-mitigation strategies for decision making in pre-processing, in-processing, and post-processing techniques. Even though several reviews have surveyed fairness methods within the context of prediction [24–26], our work aims to summarize the understanding of fairness for decision-making settings; (3) We examine outstanding challenges in fair decision making and proposing promising directions for future research; (4) We explore emerging bias-mitigation technologies that, while not yet applied in healthcare, offer significant potential for future integration; and (5) We clarify both high-level concepts and detailed examples of each methodology, thereby serving as a practical resource for both practitioners and researchers.

1.4 Organization of the Systematic Review

We begin by describing our literature search strategy. Then, we portray bias categories commonly encountered in decision making, followed by a summary of fairness metrics. Next, we present bias mitigation techniques to extend the traditional decision-making framework while achieving fair choices. Finally, we outline our survey's contributions and limitations, how our survey can assist researchers and practitioners, as well as promising future directions.

2. Search Strategies

For our systematic review, we searched the Google Scholar database for records related to fair decision making in healthcare. The electronic search strategy used the terms "decision making" and "healthcare", combined with a term in "bias", "fairness", or "equity", and one of the terms in "optimization", "machine learning", "deep learning", "reinforcement learning", "game", or "network." The keyword combinations we used are demonstrated in Figure 1. We included all records mentioning these keywords in the publication title, abstract, or full text. The publication language was restricted to English. The search’s last update was in April of 2024.

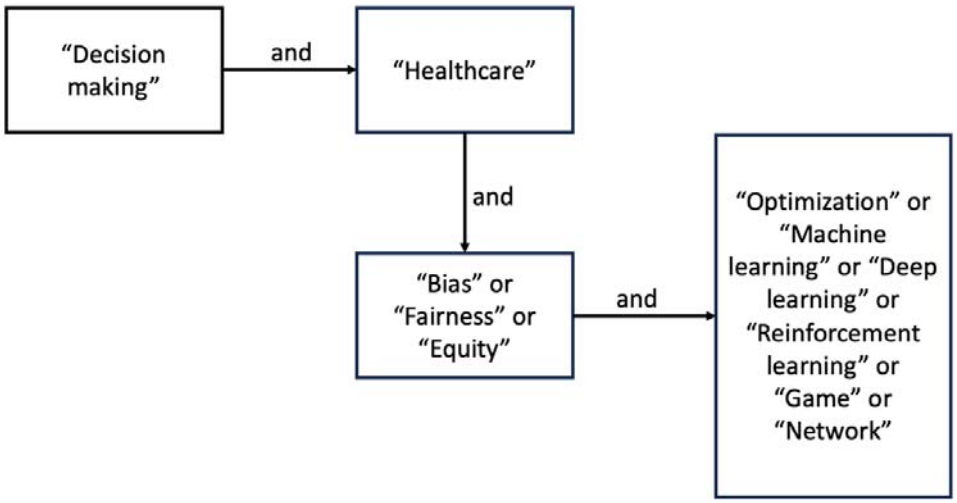

Figure 1: Key words combination for literature review

The titles and abstracts of the articles were screened by one author referred to as "investigator," then the selected manuscripts were double-checked by another author referred to as "researcher." Records were excluded if the publications: 1) did not address healthcare topics and could not be extended to healthcare easily; 2) did not focus on decision making (e.g., papers focusing on predictions); or 3) only included introductory text or conference abstract. Of the remaining publications on methodology or review of bias, fairness metrics, and bias mitigation methods, the full text was screened by investigator before the final discussion with researcher. The articles and surveys are categorized into three sections: bias in healthcare (section 3.1), fairness metrics (section 3.2), and fair decision-making models and algorithms (section 3.3). Discordance between authors was settled through discussion until the consensus had been achieved.

Fair decision-making concepts (i.e., types of biases, fairness metrics, or bias mitigation approaches) were extracted from the full text of the selected works by investigator. To ensure the accuracy and completeness of the extracted aspects, the selected concepts were double-

checked by researcher. All chosen concepts were grouped into section 3.1, section 3.2 or section 3.3 to reflect their relations.

3. Literature Review Results

The systematic review led to 801 records; 211 of them were unrelated and could not be easily transferred to the healthcare domain, and 212 of them did not focus on decision making. Furthermore, 72 records were excluded because they were introductory text or conference abstracts. Of the remaining articles, 155 papers discussed fair decision making in healthcare without a focus on methodology or specific examples of application. In the 151 papers left, we excluded 37 additional articles since they did not meet the inclusion criteria described in section 2. Of the remaining 114 papers, 25 were review papers and 89 were original papers. The flow diagram for literature review is shown in Figure 2.

Of the included papers, all concepts related to fair decision-making in healthcare were extracted. We created a structured coverage of the key findings in the following sections. Section 3.1 categorizes biases into three groups: 1) algorithmic bias, 2) data bias, and 3) publication bias. Section 3.2 classifies fairness metrics as one of the following three types: 1) fairness through unawareness, 2) demographic parity, and 3) equal opportunity. Section 3.3 categorizes bias mitigation techniques into three distinct classes: pre-processing, in-processing, and post-processing methodologies.

For the in-processing section, we describe the cutting-edge optimization and ML methods used to alleviate bias. At a high level, there are two mainstream approaches to achieving fairness in these methods: incorporating fairness in objectives or adding fairness-enhancing constraints.

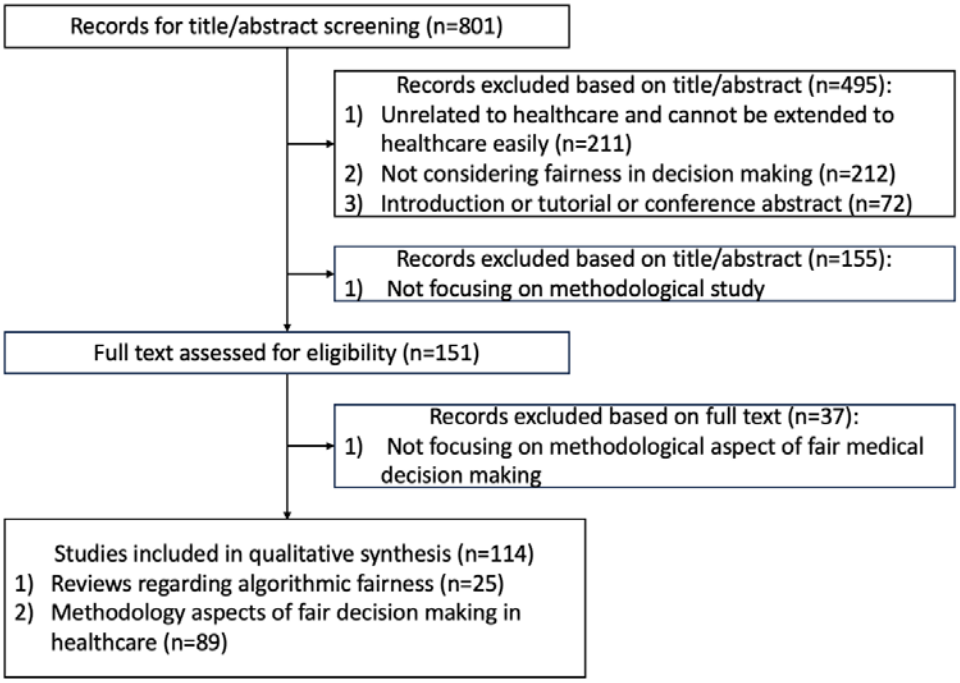

Figure 2: Flow diagram for the systematic literature review of fair decision making in healthcare

3.1 Bias in Healthcare Decision Making

Biases can exist in data and algorithmic design, which may impede decision-making systems from generating equitable outcomes among subgroups. In this section, we summarize some sources of bias impacting decision making across healthcare domains. These biases can be categorized into one of the following classes: algorithmic bias, data bias, and publication bias. Section 3.3 of our survey will summarize methods addressing these biases.

3.1.1 Algorithmic Bias

Algorithmic bias stems from computational procedures failing to consider fairness in their execution. This type of bias is a result of improper algorithmic design, which consequently

influences fairness of medical decisions [27]. For example, optimization-based vaccine allocation algorithms aiming to maximize overall social welfare may exacerbate demographic disparities since underrepresented populations may have less access to vaccines under this objective [28]. Similarly, ambulance allocation models may fail to consider fairness of unit availability across different populations by solely maximizing the overall survival rate. While the survival rate may be high among the entire population, it can be low among patients from vulnerable populations, such as people with lower socioeconomic status [29]. Algorithmic bias also exists in ML models. For instance, Samorani and coauthors have found that ML-based scheduling models have a higher likelihood of assigning Black patients to overbooked slots [30], resulting in worse service experience and longer waiting time.

3.1.2 Data Bias

Data bias refers to the unfairness generated by prejudiced data sources. Unbiased datasets are often necessary for high-quality decision-making models [31]. However, socioeconomic and racial disparities in resource availability may lead to skewed datasets [32]. Two common data biases in healthcare are aggregation biases and representation biases. Aggregation bias refers to the effect of aggregating data without considering disparities among subgroups [33]. For example, hemoglobin A1c level, a widely accepted indicator of diabetes, vary across sex and ethnicity. If we ignore the subgroup differences in the data and draw conclusions for subpopulations based on the entire population, we may introduce aggregation biases [10, 11, 34, 35]. Representation bias occurs when the data cannot represent the characteristics of all subgroups [36–40]. For instance, providers may have fewer electronic health records for people from lower socioeconomic status as they have limited access to electronic healthcare systems. Hence, we can expect fewer data among people from lower socioeconomic status, which indicates their data cannot demonstrate their overall characteristics. If decision-making

models are built with data underrepresenting this population, the developed models can be biased against those with lower socioeconomic status [41, 42].

Another source of data bias is response bias. Response bias occurs when data are labeled inconsistently or collected by unreliable methods. Response bias frequently happens in self-reported data or surveys due to participants' inaccurate answers. For instance, when medical students rate their mental health conditions in surveys, they tend to provide socially acceptable answers. However, such answers often do not align with their true mental conditions as students [43]. Since policymakers may harness data to make public health decisions, response bias can skew decision making [44], and models built from data with response bias may underestimate the seriousness of medical students' mental health problems [32].

3.1.3 Publication Bias

Publication bias happens when the researchers' decision to publish a paper depends on the study results. Compared with studies without positive results, publishing medical studies with positive results is generally easier [45]. This phenomenon may lead to the overestimation of certain clinical treatments as evidence against a treatment is not made available. Medical practitioners may then make treatment decisions based on biased outcomes, giving rise to degraded treatment effects.

An example of publication bias occurred during the COVID-19 pandemic. The academic papers concerning COVID-19 treatment were published rapidly within this period. Most publications showed promising outcomes of COVID-19 treatments, while less satisfying research results only had a slim chance of being published [46]. When related treatments

were applied to the general population, many of these treatments produced inferior outcomes compared to publication results [46]. Another example comes from the Cochrane Review on topical benzoyl peroxide, a widely used acne treatment. In 2019, Yang and coauthors found that most benzoyl peroxide studies before 2015 were still unpublished because of their negative results [47]. This finding suggests published literature could not reliably reflect the overall effect of benzoyl peroxide, and medical practitioners applying benzoyl peroxide may not have achieved the desired treatment effects.

3.2 Fairness Metrics

In this section, we present the evaluation of three types of metrics: fairness through unawareness, demographic parity and equal opportunity. To portray the ideas of different fairness metrics, we use a vaccine distribution example. For illustration purposes, we restrict our attention to only one binary sensitive attribute, denoted by A. An example of this type of attribute may be a dichotomized version of race, which includes White and people of color as its categories. The metrics covered by the section are summarized in Table 1.

Table 1: Definition of fairness metrics

	Fairness through unawareness	**Demographic parity**	**Equal opportunity**
Sensitive attribute	No	Yes	Yes
Individual qualification	No	No	Yes

Definition	$X \cap g = \emptyset$	After feeding X to decision-making models, $E(r_{g_1}) = E(r_{g_2})$	After feeding X to decision-making models, within a qualified subgroup, $E(r_{g_1}) = E(r_{g_2})$

In this review, X denotes entire characteristics of the population, g denotes the sensitive attribute (in our case sensitive attribute is race), r_{g_1} denotes the cumulative rewards of White patients while r_{g_2} denotes the cumulative rewards of people of color. A qualified subgroup represents a subset of the general population that may be of special interest to decision makers. For example, certain patients, such as senior citizens and those residing in areas with inadequate healthcare resources, could be considered qualified patients as they may be more susceptible to the disease.

3.2.1 Fairness Through Unawareness

Fairness through unawareness is the base fairness metric [15]. It does not consider any sensitive attribute during the decision-making process. Within the context of our vaccination distribution example, fairness through unawareness means the model is considered fair if it does not consider race while deciding the vaccination distribution. The benefit of fairness through unawareness is it is simple and convenient to implement. The advantage of fairness through unawareness lies in its simplicity and ease of implementation. However, simply ignoring sensitive attributes may not remove inequity, because other variables can be highly correlated with the sensitive traits. For instance, socioeconomic status or geographic location could be strongly correlated with race. Consequently, the model might still make decisions

that indirectly disadvantage certain people of color, perpetuating inequity. This method has been proven to be invalid in many cases since the method neglects the structural inequalities embedded in proxy variables [48].

3.2.2 Demographic Parity

The demographic parity fairness metric aims to ensure the expected cumulative reward of a decision-making model is independent of sensitive attributes [13]. Independence of sensitive attributes indicates the outcomes (i.e., expected cumulative rewards) must be equivalent in privileged and unprivileged groups [49]. In our vaccine distribution example, demographic parity ensures that, when other information is the same (e.g., age, socioeconomic status), White patients and patients of color have the same expected cumulative rewards (e.g., total symptom improvements). The problem with demographic parity is that it does not consider the population's ground-truth qualifications. For instance, consider a case where patients of color are statistically more vulnerable to a particular disease due to underlying health disparities or systemic inequities. These patients might have a higher ground-truth qualification for receiving the vaccine. Applying demographic parity in such a scenario would require the vaccines be distributed equally across racial groups, regardless of the groups' actual vulnerability to the disease. This distribution would neglect the real-world needs of patients of color and could lead to suboptimal outcomes, such as under-protection of the most vulnerable populations.

However, demographic parity may be appropriate in a scenario where there is no significant difference in vulnerability across racial groups, and fairness requires strict equality of treatment. For example, if the disease impacts White patients and patients of color equally and their ground-truth qualifications are the same (i.e., both groups are equally at risk and

stand to benefit equally from the vaccine), demographic parity ensures that neither group is disadvantaged in the vaccine distribution process. In such a situation, applying demographic parity would help prevent systemic biases from skewing vaccine access while maintaining fairness in outcomes.

3.2.3 Equal Opportunity

Similar to demographic parity, equal opportunity verifies whether the expected cumulative rewards for privileged and unprivileged groups are the same [27]. However, demographic parity applies to the entire population, while equal opportunity applies solely to a qualified population [49]. Within our example, this metric requires that within a qualified population, the cumulated expected reward of receiving vaccine among the unprivileged group ($A = people\ of\ color$) is the same as the cumulated expected reward of receiving vaccine among the privileged group ($A = White$).

Equal opportunity requires a well-defined notion of "true eligibility," meaning that individuals who meet the criteria for a treatment should be treated equally across groups. Within our vaccine distribution example, if White and patients of color both have a set of equally high-risk individuals, equal opportunity can guarantee that no group is unfairly advantaged in the eligible, high risk, population.

Nevertheless, equal opportunity becomes inappropriate when unqualified individuals are unfairly impacted when the qualification criteria themselves are biased or imperfectly measured. For instance, if systemic issues lead to more misclassifications of people of color as unqualified, focusing exclusively on the qualified population may overlook these disparities and exacerbate existing inequities.

3.2.4 Summary of Fairness Metrics

Fairness through unawareness is the simplest fairness metric, requiring that sensitive attributes, such as race or gender, are excluded from decision making. However, it is often invalid because other variables correlated with sensitive attributes can serve as proxies, perpetuating biases even when the sensitive attributes are omitted. Demographic parity, on the other hand, evaluates whether decision making is independent of sensitive attributes across the entire population, ensuring equality of outcomes. While this can help mitigate systemic biases, it ignores differences in ground-truth qualifications between groups, potentially leading to unfair decisions. Equal opportunity refines demographic parity by focusing on whether decision making is independent of sensitive attributes within a qualified subgroup, such as high-risk individuals in a vaccine distribution example. This approach ensures fairness among those most in need but does not address inequities among the unqualified population, such as those misclassified as ineligible or indirectly disadvantaged. These metrics vary in applicability, with their effectiveness depending on the context and the specific fairness goals.

3.3 Bias Mitigation

In this section, we summarize different bias mitigation approaches used across healthcare domains. The methods can be categorized into pre-processing, in-processing, and post-processing. Pre-processing mechanisms clean and manipulate the input data before it is used in decision-making models or algorithms [50]. In-processing methodologies refer to building unbiased models or algorithms directly [51]. Post-processing methods calibrate algorithmic outcomes to achieve fairness [52].

3.3.1 Pre-Processing

Datasets may be biased, which can cause skewed decisions. For instance, when a dataset is imbalanced, decisions may be biased toward subpopulations with smaller sizes [53]. Pre-processing methods can help circumvent possible biases in this setting. There are five commonly used approaches for pre-processing: reweighting the underrepresented populations, resampling, fair data transformers, natural language processing, and post-survey analysis. The identified pre-processing methods, along with their respective reference, areas of application, fairness metrics, and targeted bias categories, are summarized in Table 2. Please see section 1.1 in the Supplementary Information for an in-depth discussion of the strengths and limitations of the pre-processing methods presented in this review.

Table 2: Pre-processing bias mitigation methods

Method(s)	**Reference(s)**	**Area of application**	**Fairness metric**	**Targeted bias**
Reweighting	Nilsson et al. [54]	Medical diagnosis	Demographic parity	Representation bias
	Kumar et al. [55]	Medical diagnosis		
	Peacock et al. [56]	Resource allocation		
Resampling	Chawla et al. [57]	Clinical treatment		
	Borland et al. [58]	N/A[a]		
	Mohamed et al. [59]	Medical diagnosis		
Fair data transformers	Biswas et al. [50]	N/A[a]		
	Grgić-Hlača et al.	N/A[a]		

	[60]			
	Briscik et al. [61]	Medical diagnosis		
Natural language processing	Hajjar et al. [62]	N/A[a]		
	Zhou et al. [63]	Clinical treatment		
	Minot et al. [64]	Medical diagnosis	Equal opportunity	
Post-survey analysis	Serra et al. [65]	Medical diagnosis	Demographic parity	Response bias
	Fulton et al. [66]	N/A[a]		

[a] This method may be easily extended to healthcare settings.

Reweighting. Reweighting assigns greater weights to underrepresented instances [54]. Biases may be introduced to decision-making tasks if we fail to process underrepresented population's data. Skewed data can also lead to threatening consequences for underrepresented populations. For example, African Americans and Asians have fewer instances in genome studies, which gives rise to higher misclassification rates for the two subgroups in clinical research [11]. Classification methods can investigate the features of each piece of data and use them to decide which category or label the data belongs to. Since physicians may rely on classification methods for diagnosis and treatment design, varying misclassification rates for different subgroups can trigger bias in decision making. A remedy for similar issues is to assign greater weights to underrepresented instances, hence data from all populations play an equal role in the modeling process [54]. Kumar and coauthors have shown that distributing more weight to underrepresented groups can improve fairness in medical image classification by 8% [55].

One crucial factor influencing reweighting efficiency is the choice of weighting scheme. Complex or highly granular schemes (e.g., accounting for multiple intersecting protected attributes) can increase computational overhead. Another consideration is the distribution of protected subgroups: extremely small or imbalanced groups may lead to unstable or excessively large weights, thereby affecting convergence speed and model stability.

Resampling. Resampling is a technique to ensure the data is balanced (i.e., has an near-equal number of instances from each subgroup) by repeatedly drawing samples from the same data [53]. Using imbalanced data directly may favor the majority groups while disregarding the minority ones. To avoid this concern, we can resample from the minority groups, so the majority and minority groups have approximately the same size. The relative proportions of protected groups heavily influence efficiency of resampling [53]. When there is a severe imbalance between groups, oversampling the minority group or undersampling the majority group can lead to overfitting or information loss, respectively.

Fair data transformers. Fair data transformers extract features from the input data in a fair manner. Such transformers modify the input data to achieve fairness [50]. Data transformers, such as principal component analysis, are popular techniques for pre-processing data. In practice, multiple transformers are typically evaluated for a specific fairness metric (e.g., demographic parity). The transformers achieving the fairest output are used to produce inputs for ML and optimization algorithms. Biswas and collaborators have shown that a proper data transformer can significantly improve the fairness of outcomes [50]. Though fair data transformers have not been deployed in healthcare to the best of our knowledge, it is easy to extend a fair data transformer to healthcare data preprocessing. For example, in the context of vaccination distribution, we can collect data such as age, sex, population density, and

incidence rate in the areas where individuals live. Then, we apply several fair data transformers and feed the transformed data into the same decision-making algorithm. After the algorithm outputs vaccination distribution decisions, we can evaluate the fairness of outcomes and choose the data transformer that produces the fairest output.

The efficacy of a fair data transformer is sensitive to several factors, including the quality and representativeness of patient data, the complexity of fairness definitions, and the regulatory environment in which it operates. For instance, imbalanced datasets or limited data on certain patient subgroups can reduce the transformer's ability to produce equitable outcomes. Additionally, ensuring that the transformed features remain clinically interpretable is vital for practitioner trust, and any deployed system must be monitored over time to adapt to changing patient populations and medical standards.

Natural language processing. Natural language processing removes biased information from text data before feeding data to decision-making algorithms [63]. Due to the complexity of healthcare data, natural language processing is becoming increasingly popular in fair pre-processing settings.

Post-survey analysis. Post-survey analysis for data bias mitigation refers to the process of analyzing survey data after its collection to identify, assess, and correct various types of biases that may have been introduced during the data collection phase [31]. For example, some respondents might misremember their health history; hence the treatment effect for them is likely to be inferior compared to respondents remembering correctly. Researchers can mitigate this issue by cross-referencing survey responses with medical records to mitigate such bias [65].

3.3.2 In-Processing

In-processing methodologies incorporate one or more fairness metrics directly in the design of algorithms or models to lessen biases. These bias mitigation techniques are attracting increased attention within domains such as resource allocation, scheduling, and clinical treatment [30, 67]. Overall, we find six typically used in-processing methods in the literature: mixed-integer programming, stochastic programming, deep reinforcement learning, fair survival analysis, multi-objective Markov Decision Process, and constrained Markov Decision Process. Table 3 demonstrates the methods and the corresponding references and applications. Please refer to section 1.2 of the Supplementary Information for a detailed explanation of the strengths and limitations of in-processing methods.

Table 3: In-processing bias mitigation methods

Method(s)	**Reference(s)**	**Area of application**	**Fairness metric**	**Targeted bias**
Mixed-integer programming	Acuna et al. [29]	Resource allocation	Demographic parity	Algorithmic bias
	Radovanović et al. [68]	Resource allocation		
	Ala et al. [69]	Scheduling		
	Zhong et al. [70]	Scheduling		
	Neophytou et al. [71]	Resource allocation		

	Wolbeck et al. [72]	Scheduling		
	Gunnarsson et al. [73]	Resource allocation		
	Sepulveda et al. [74]	Resource allocation		
	Klyve et al. [75]	Scheduling		
	Gross et al. [76]	Scheduling		
	Akshat et al. [77]	Resource allocation		
	Azizi et al. [78]	Scheduling		
	Proano et al. [79]	Scheduling		
	Vahdani et al. [80]	Resource allocation		
	Lodi et al. [80]	Resource allocation	Equal opportunity	
	Argyris et al. [81]	Resource allocation		
	Rastegar et al. [82]	Resource allocation		
	López et al. [28]	Resource allocation		
Stochastic programming	Ala et al. [83]	Scheduling	Demographic parity	
	Yin et al. [84]	Resource allocation		

Deep reinforcement learning	Yang et al. [21]	Clinical treatment		
	Yu et al. [85]	Clinical treatment		
	Li et al. [86]	Resource allocation		
	Atwood et al. [87]	Resource allocation	Equal opportunity	
	Budhiraja et al. [88]	Scheduling		
Fair survival analysis	Keya et al. [89]	Resource allocation	Demographic parity	
Multi-objective Markov Decision Process	Ge et al. [90]	N/A[a]	Equal opportunity	
Constrained Markov Decision Process	Ge et al. [20]	N/A[a]	Demographic parity	

[a] This method may be easily extended to healthcare settings.

Mixed-integer programming (MIP). This methodology has been widely used to ensure fairness in decision making [29, 68–82]. Emergency department overcrowding has become a nationwide crisis over the last decade [29]. To resolve the overcrowding issue, researchers have applied MIP to build fair medical resource distribution models. An MIP is a type of constrained optimization problem that allows for both integer and continuous variables in its objective and constraints [91]. For example, Acuna and coauthors added equity constraints to

ensure that the minimal quality of care for every emergency is greater than or equal to a predefined threshold in an ambulance allocation situation [29]. We discuss this example in subsection 1.2.1 of the Supplementary Information.

Another ubiquitous way to fulfill fairness requirements in decision making is to modify the objective function of an optimization approach. When medical resources are scarce, people from vulnerable groups may have lower access to them. To ensure fairness towards vulnerable populations, the objective function of an algorithm can be set to maximize the smallest number of allocated resources across all population subgroups [28]. This objective ensures that each subgroup receives their required medical support.

Stochastic programming. Researchers have also leveraged stochastic programming techniques to generate in-processing bias mitigation techniques [83, 84]. These techniques optimize an objective function while representing uncertainty through probability distributions [92]. For example, we can add fair constraints in stochastic programming models to limit the expected difference between the maximum waiting time and minimum waiting time [83]. Please refer to subsection 1.2.2 of the Supplementary Information for a more detailed discussion of this example.

Deep reinforcement learning. Reinforcement learning and deep learning play a pivotal role in in-processing methods. Reinforcement learning is a type of ML where the algorithms learn to make decisions by performing actions and observing the results in an environment of interest [93]. Deep learning applies multiple neural network layers and activation functions to extract new features from the input data [94], being capable of recapitulating and modeling complex patterns in data. Deep reinforcement learning is a combination of deep learning and

reinforcement learning that can be used to solve Markov Decision Process (MDP) models. An MDP is a mathematical framework used for modeling decision making in situations where outcomes are partly random and partly under the control of a decision-maker [95]. In an MDP, the reward quantifies how desirable the outcome is after taking a specific action in a given situation. Higher rewards indicate more favorable outcomes. For instance, Yang et. al redefine the rewards of deep reinforcement learning by ensuring the absolute value of rewards of subgroups are smaller if they have a large size [21]. We discuss this example in more detail in subsection 1.2.3 of the Supplementary Information.

Fair survival analysis. Fair survival models provide an additional tool for decision making in healthcare settings [89]. Traditional survival analysis estimates the time until an event of interest [96]. Fair survival models incorporate event probabilities and fairness violations when providing the estimate. For example, the allocation sequence of a resource may be decided based on the risk outcomes of a survival model that penalizes the difference between the highest and lowest probabilities of disease incidence within a cohort [89]. Their numerical experiment shows the fair survival model can substantially reduce the group disease risk range. Please refer to subsection 1.2.4 of the Supplementary Information for a more detailed discussion of this example. By integrating these survival-based risk predictions into the waitlist formation, decision makers can systematically balance disease risk considerations with fairness goals when allocating limited resources.

Multi-objective Markov Decision Process. While it has not been applied to healthcare settings to the best of our knowledge, the multi-objective MDP is a promising approach to alleviate the potential effect of bias. This model is an extension of the traditional MDP with the difference that the reward function depends on a utility objective and a fairness objective

[97]. Ge and coauthors have applied the Pareto frontier to identify the policy that optimizes both utility and fairness elements [90]. Their results show that there exists a trade-off between utility and fairness performance, and we can choose the final policy—i.e., the action chosen in each situation—based on user preferences. We discuss this example in more detail in subsection 1.2.5 of the Supplementary Information. Though multi-objective MDP has not been applied to fair decision making in healthcare to the best of our knowledge, it is possible to deploy these methods to generate fair clinical decisions. For example, if we need to guarantee similar vaccination rates between males and females, we can add this fairness objective into our model. The Pareto frontier can return optimal policies that consider both vaccine utility and distribution fairness.

Constrained Markov Decision Process. The Constrained Markov Decision Process (CMDP) is another prospective direction. Compared to traditional MDP, CMDP can accommodate fair constraints in decision making [98]. In CMDP, we can formulate the cost function regarding fairness, and then choose policies leading to fairness cost less or equal to the threshold [20]. This model has not been utilized in fair healthcare decision making to the best of our knowledge. However, we can model fairness metrics as constraints and choose a set of policies that satisfy these constraints. Afterward, we can investigate which policy in this set gives the optimal discounted accumulated reward. Please refer to section 1.2.6 of the Supplementary Information for a more detailed explanation of this example.

<u>Summary of In-Processing Methods</u>

Each bias mitigation strategy suits different healthcare scenarios based on data availability, decision objectives, and computational constraints. For instance, mixed-integer programming MIP is effective for discrete resource allocation (e.g., scheduling or capacity planning) and

allows explicit fairness constraints, but it may struggle with high-dimensional patient data. Stochastic programming handles uncertainty in clinical pathways and can incorporate probabilistic fairness bounds, yet it requires robust probability estimates and may be computationally demanding. Deep reinforcement learning adapts well to complex state spaces and large-scale data, but it can be opaque, raising explainability concerns and requiring careful regularization to avoid overfitting. Fair survival-based decision-making is crucial for time-to-event outcomes, ensuring that interventions and resources are allocated in a way that balances long-term risks and benefits across demographic groups. However, enforcing fairness at a specific time horizon can introduce temporal distortions in subsequent decisions, as ensuring equity at one point in time may inadvertently shift survival probabilities—and thus decision outcomes—at later points. Multi-objective MDP models allow clinicians to weigh multiple criteria (e.g., treatment effectiveness vs. equity), though specifying trade-offs can be challenging. Constrained MDP models impose hard fairness requirements that are useful for high-stakes decisions, but they may reduce flexibility in optimizing overall performance. Ultimately, the choice depends on the clinical setting, data availability, ethical priorities, and the extent to which one can tolerate reduced performance for stronger fairness guarantees.

3.3.3 Post-Processing

Post-processing methods calibrate algorithmic outcomes to achieve fairness. We identify the following post-processing methods relevant to healthcare applications: Laplacian smoothing, multi-accuracy approaches, and expert systems. The post-processing bias mitigation methods included in this review are shown in Table 4. Please refer to section 1.3 of the Supplementary Information for a more detailed discussion of the strengths and limitations of post-processing methods.

Table 4: Post-processing bias mitigation methods

Method(s)	Reference(s)	Area of application	Fairness metrics	Targeted bias
Laplacian smoothing	Petersen F et al. [52]	N/A[a]	Demographic parity	Algorithmic bias
Multi-accuracy	Kim et al. [99]			
Expert systems	Tang et at. [22]	Clinical treatment		
	Lu et at. [100]			
	Tang et at. [22]		Equal opportunity	
	Yu et al. [101]			

[a] This method may be easily extended to healthcare settings.

Laplacian smoothing. The Laplacian smoothing method is a technique to reduce the noise of the data while preserving the important characteristics of the solution technique [102]. For instance, we can take advantage of this method to guarantee comparable treatment plans among similar individuals while preserving the performance of the algorithms [52]. Selecting an appropriate smoothing parameter is critical: too much smoothing risks diluting meaningful subgroup distinctions, while too little smoothing can leave small groups vulnerable to statistical noise. Balancing these factors is key to leveraging Laplacian smoothing as an efficient and robust tool for improving fairness in ML applications.

Multi-accuracy. We can apply multi-accuracy approaches to combine several weak learners to achieve high accuracy rates among all subpopulation groups [99]. These methods play vital role in classification techniques by assigning larger weights to samples identified incorrectly in weak learners. Subsequently, the following weak learners pay extra attention to

misidentified samples and adjust their results accordingly [103]. With accurate and fair classification for target populations, physicians can deploy algorithm-based treatment design to achieve desirable medical outcomes.

In addition, the efficiency of multi-accuracy methods hinges on several factors several factors: the complexity and granularity of subgroup identification, the number of iterations required for convergence, and the computational overhead associated with retraining or correcting the model. Efficient implementation demands careful selection of error thresholds to detect performance gaps without incurring excessive retraining cycles.

Expert systems. Lastly, the clinical expertise of medical practitioners may help increase the fairness of algorithms [101, 104, 105]. For example, reinforcement learning techniques can suggest several near-equivalent actions, then we can rely on clinicians to decide what actions can lead to the fairest outcome [22]. This approach may enable improved decisions to overcome potential biases while leveraging practitioners' experience. In practice, one of the most critical factors in expert systems is the complexity of the knowledge base, including the number of rules, the granularity of each rule, and the amount of domain-specific information encoded; as these grow, so does the computational cost of rule matching and inference. Additionally, update frequency—how often new medical guidelines or clinical insights are incorporated—can influence both performance and maintenance overhead.

A summary of the strengths and limitations of the bias mitigation methods presented in this survey is shown in Table 5. We discuss these advantages and disadvantages in more detail in section 1 of the Supplementary Information.

Table 5: Strengths and limitations of bias mitigation methods

Field	Method(s)	Strengths	Limitations
Pre-Processing	Reweighting	Preserves original data without generating synthetic samples	Requires careful tuning to avoid overemphasizing minority groups
	Resampling	Addresses class or group imbalance directly	Oversampling can lead to overfitting while undersampling risks losing valuable information
	Fair data transformers	Adjusts or masks sensitive features to reduce bias	May lose important information if transformations are too aggressive
	Natural language processing	Useful for extracting rich features from clinical notes	Computationally intensive for large text corpora
	Post-survey analysis	Provides direct feedback on perceived fairness	Time-consuming to conduct and interpret
In-Processing	Mixed-integer programming	Guarantees an optimal solution under well-defined constraints	Computationally expensive for large-scale problems
	Stochastic programming	Handles uncertainty in parameters or data	Modeling complexity increases with more uncertainty sources
	Deep reinforcement learning	Can adapt to changing patient populations over time	Requires large amounts of training data

	Fair survival analysis	Accounts for time-to-event data and censoring	Limited standard tools for enforcing fairness constraints
	Multi-objective Markov Decision Process	Simultaneously optimizes multiple criteria	Finding a balanced solution among competing objectives can be complex
	Constrained Markov Decision Process	Directly enforces fairness or risk constraints within the MDP framework	Potentially high computational overhead for large state/action spaces
Post-Processing	Laplacian smoothing	Stabilizes parameter estimates, especially for underrepresented groups	Can over-smooth differences among subgroups
	Multi-accuracy	Ensures balanced performance across subgroups by iterative audits	Repeated audits can be time-consuming and computationally expensive
	Expert systems	Transparent, rule-based reasoning aligned with clinical guidelines	Updating and maintaining rules is labor-intensive, especially for evolving medical knowledge

The distribution of bias mitigation literature is presented in Supplementary Figure 1 of the section 2 of the Supplementary Information. We include the trend of bias mitigation approaches over the last decade as Figure 3 and the distribution of papers across areas of application in Figure 4.

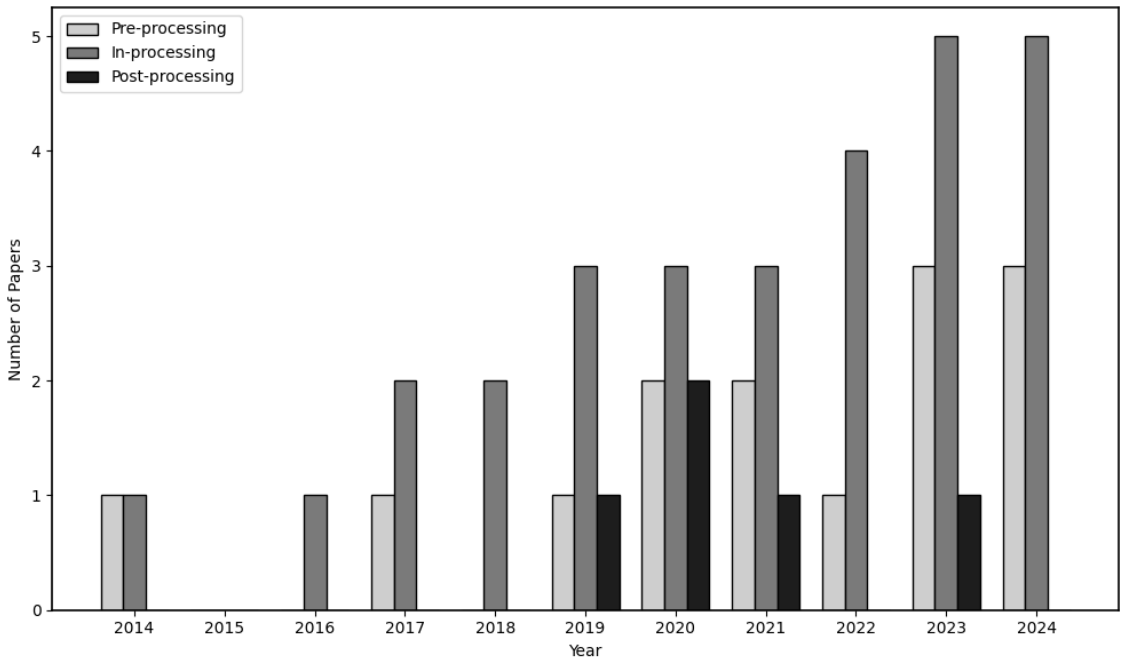

Figure 3: Trend of bias mitigation approaches over the last decade

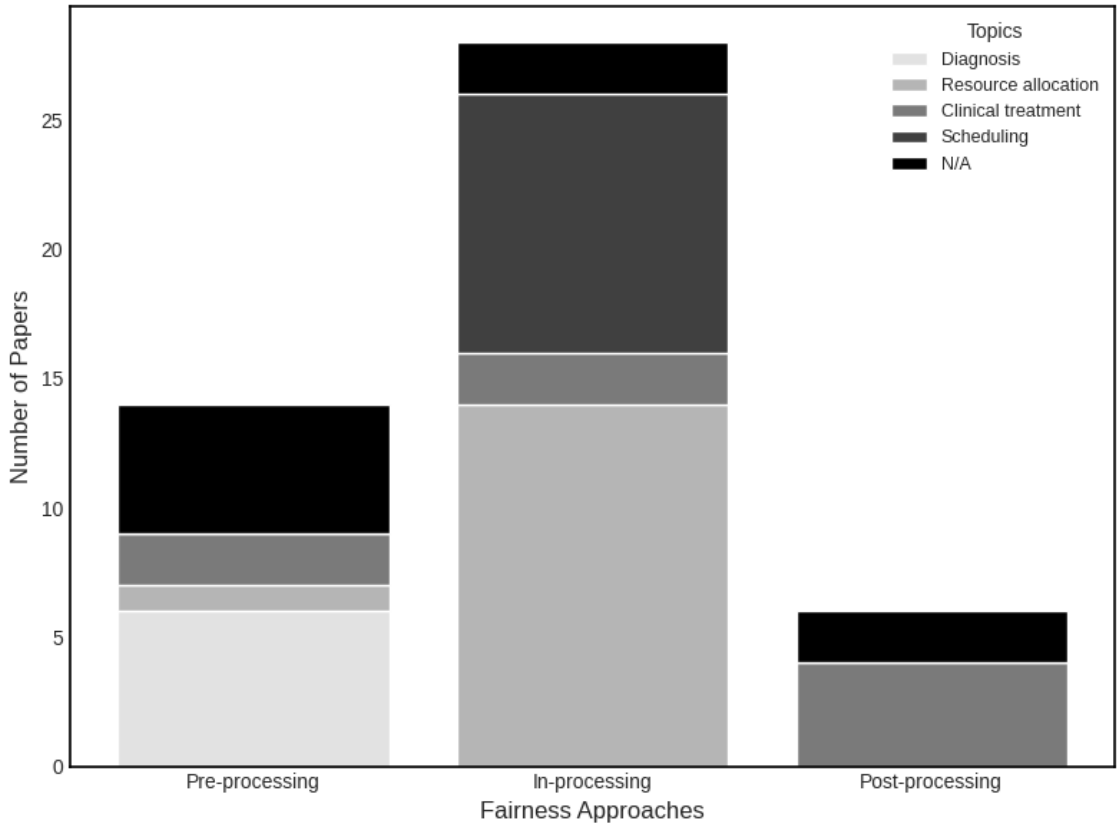

Figure 4: Distribution of papers across topics

3.3.4 Fair Decision Making Nationally and Internationally

In the U.S., fair decision-making in healthcare is often shaped by regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA), alongside a fragmented healthcare system that influences data-sharing practices [106]. Research in the U.S. tends to focus on algorithmic bias mitigation within electronic health record data, reflecting the prominence of private healthcare providers and diverse data ownership models. Conversely, some non-U.S. settings—especially in countries with nationalized healthcare systems—can leverage centralized patient databases, which may offer more uniform coverage but also raise unique privacy and ethical concerns. Regulatory regimes like the European Union's General Data Protection Regulation (GDPR) further shape data handling and bias detection by imposing stricter consent and transparency requirements [107]. Moreover, cultural differences in attitudes toward data collection, resource allocation, and equity lead to variations in how fairness metrics are defined and prioritized. Despite these contextual disparities, both U.S. and international studies share common challenges—such as representation bias and response bias—pointing to the need for continued global collaboration to develop robust, context-aware fairness solutions.

4. Conclusion and Future Research Directions

Bias in decision making can stem from various factors, such as algorithm design, data sources, publication practices, limited resources, patient diversity, and cultural differences. Commonly discussed fairness metrics include fairness through unawareness, demographic parity, and equal opportunity. Multiple metrics have been developed to conceptualize and quantify fairness across various stages of the decision-making pipeline, including the dataset, algorithmic design, and resulting outcomes. To address biases stemming from these distinct sources, researchers have proposed three primary categories of bias mitigation strategies: (1)

Pre-processing methods, which focus on identifying and rectifying biases within datasets prior to their use in modeling and algorithms, thereby ensuring more equitable representations; (2) *In-processing methods*, which aim to incorporate fairness constraints or objectives directly into the algorithm and model design, reducing disparities across subgroups during the modeling process; and (3) *Post-processing methods*, which adjust model outcomes to achieve fairer results without altering the underlying algorithm and model. These approaches collectively represent a comprehensive framework for addressing fairness by optimization and ML.

4.1 Extensions and Future Work

Continuing research on fair decision making in healthcare is vital because existing methods do not yet fully capture the complexity and diversity of real-world clinical settings. Many current approaches rely on assumptions or metrics that overlook individual patient nuances, such as evolving clinical states or social determinants of health. Moreover, high-stakes healthcare decisions demand both transparency and ethical accountability, though existing "black box" models can obscure sources of bias. As a result, there is a pressing need for new, context-aware frameworks that integrate domain expertise, robust optimization, and interpretability to ensure equitable outcomes for all patients.

Moving forward, several questions are worth exploration: The first interesting open question is explainability for fair algorithms. Many fair algorithms in healthcare are considered black boxes that are hard to understand. The lack of explainability is an obstacle for practitioners to identify if the model is relying on biased features. Another promising field is algorithmic interpretability. If a fair algorithm doesn't align with practitioners' intuition and domain knowledge, practitioners may be skeptical to adopt the algorithmic output. In addition, it is

worthwhile to explore how to keep the balance between interpretability and fairness. While increased interpretability can earn more trust among practitioners, it may hurt the model's fairness [64]. Furthermore, context-aware fairness metrics is a significant topic to explore. Researchers have found that different fairness metrics can be incompatible. Thus, we cannot expect a model to satisfy all fairness metrics [103]. In such contexts, it is critical to understand which type of metric we should consider in a specific circumstance. Identifying the best fairness metric for a specific problem will likely require cooperation between modelers and domain experts [26]. Another promising methodological innovation is to bridge the divide between fair prediction and fair decision making. Typically, current approaches follow a "fair prediction then optimization" pipeline, operating under the assumption that fair predictions will naturally result in fair decisions. However, this causal link is not assured. To address this divide, innovative methodologies are required that explicitly incorporate the cost of fair decision making—stemming from fair predictions—into the optimization objective. Such approaches could enable the simultaneous achievement of both fair prediction and fair optimization [108].

Furthermore, we need to consider fairness beyond expected values, and extensions such as fairness of variances across subgroups can fulfill the goal by evaluating the spread of outcomes, not just their expectations [21]. In addition, ideas from unsupervised and supervised learning, such as representation learning, can reveal hidden subgroups for personalized care before policy optimization, supporting fair decision making in healthcare [55]. Finally, another big open question is to establish individual-level fairness in large populations. Individual-level fairness is the notion that two individuals who are sufficiently similar in clinically relevant ways should receive similar outcomes. Establishing this notion within large and diverse healthcare populations remains an open challenge. This task involves

defining appropriate measures of similarity in complex, high-dimensional data and ensuring that these measures do not conflict with broader group-level fairness criteria or clinical objectives. Successfully addressing individual-level fairness would enable more personalized and equitable decision making, ultimately improving patient care [49].

We acknowledge that factors such as information asymmetry, resource constraints, patient diversity, and cultural differences can intersect with or exacerbate the biases discussed in this review. For instance, limited or skewed data can intensify representation bias, and uneven access to resources may deepen representation bias. Although a detailed exploration of these broader challenges is beyond our current scope, we note their significance and encourage future research to address them in conjunction with the biases we identify. Furthermore, the bias mitigation techniques mentioned in subsection 3.3 can be applied to various healthcare challenges, including information asymmetry, limited resources, patient diversity, and cultural differences. For example, reweighting or resampling can address data imbalances from vulnerable groups, while fairness constraints may help ensure each group receives a minimum standard of care. Furthermore, methods like constrained MDP models can dynamically adapt recommendations to diverse patient characteristics, promoting more equitable outcomes.

Because our search was restricted to English-language sources, this review may not fully capture the perspectives and developments present in non-English-speaking regions. Additionally, we might have missed keywords during our literature review, leading to the omission of works that used excluded terms.

4.2 Practical Considerations

The main sources of agreement and disagreement in fair decision-making in healthcare revolve around fairness metrics, algorithmic approaches, and their practical deployment. There is broad agreement on the importance of addressing biases in healthcare data and decision-making processes, with researchers emphasizing the necessity of fairness-aware methodologies such as preprocessing, in-processing, and post-processing techniques. However, disagreements arise in the selection and application of fairness metrics, as different metrics often produce conflicting results [103]. Moreover, there is disagreement over the assumption that fair predictions automatically lead to fair decision making, highlighting the need for more integrated approaches [105]. These points of divergence underscore the complexity of developing fair healthcare methodologies and the necessity for context-sensitive, interdisciplinary solutions [94].

To effectively deploy methods above, practitioners need accessible tools and frameworks that bridge the gap between theoretical advancements and real-world applications. First, the development of user-friendly software libraries and platforms incorporating fairness-aware algorithms and explainability features can empower practitioners to implement these methods without requiring extensive technical expertise. Open-source fair ML libraries, such as Fairlearn in Python and fairml in R, provide a valuable foundation for implementing bias mitigation strategies in predictive modeling, yet ongoing commitments are necessary to address the full complexity of implementing fair decision-making methods. Second, integrating domain-specific knowledge into these tools is critical, as it enables models to be tailored to the unique challenges and priorities of healthcare. Third, practitioners should engage in interdisciplinary collaboration with data scientists, ethicists, and domain experts to ensure that fairness and interpretability align with practical needs and ethical standards. Finally, regular audits and monitoring systems should be established to continuously evaluate

the performance and fairness of deployed models, providing feedback loops to refine methodologies and address emerging biases. By adopting these approaches, practitioners can effectively leverage innovative fairness methods to create decision-making systems that are robust, equitable, and transparent.

For researchers, our work offers a comprehensive synthesis of state-of-the-art fairness approaches, highlighting current trends, identifying research gaps, and suggesting promising future directions. This review not only facilitates methodological comparisons and benchmarking but also serves as a roadmap for developing novel fairness metrics, mitigation strategies, and evaluation frameworks. By integrating insights from both theory and practice, our survey aims to help researchers advance the field and stimulate further research into robust, equitable, and transparent decision-making systems in healthcare.

4.3 Conclusions

Given the growing importance of decision-making approaches in healthcare, fairness considerations and bias-mitigation approaches are increasingly vital. Our survey may aid practitioners and researchers in 1) understanding potential sources of biases in decision making; 2) choosing the appropriate fairness metric to evaluate decision-making models; and 3) selecting the appropriate pre-processing, in-processing, and post-processing techniques to reduce bias. In conclusion, this survey sheds light on the current state and challenges of fair decision-making in healthcare, highlighting the crucial need for continuous improvement in policies and practices to ensure equitable healthcare outcomes for all individuals.

References

1. Awaysheh A, Wilcke J, Elvinger F, et al (2019) Review of Medical Decision Support and Machine-Learning Methods. Vet Pathol 56:512–525. https://doi.org/10.1177/0300985819829524

2. Brnabic A, Hess LM (2021) Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak 21:54. https://doi.org/10.1186/s12911-021-01403-2

3. Zerouaoui H, Idri A (2021) Reviewing Machine Learning and Image Processing Based Decision-Making Systems for Breast Cancer Imaging. J Med Syst 45:8. https://doi.org/10.1007/s10916-020-01689-1

4. Peiffer-Smadja N, Rawson TM, Ahmad R, et al (2020) Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 26:584–595. https://doi.org/10.1016/j.cmi.2019.09.009

5. Abdalkareem ZA, Amir A, Al-Betar MA, et al (2021) Healthcare scheduling in optimization context: a review. Health Technol 11:445–469. https://doi.org/10.1007/s12553-021-00547-5

6. Wang L, Demeulemeester E (2023) Simulation optimization in healthcare resource planning: A literature review. IISE Trans 55:985–1007. https://doi.org/10.1080/24725854.2022.2147606

7. Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. Eur J Oper Res 258:3–34. https://doi.org/10.1016/j.ejor.2016.06.064

8. Wang L, Zhang W, He X, Zha H (2018) Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp 2447–2456

9. Mizan T, Taghipour S (2022) Medical resource allocation planning by integrating machine learning and optimization models. Artif Intell Med 134:102430. https://doi.org/10.1016/j.artmed.2022.102430

10. Balayn A, Lofi C, Houben G-J (2021) Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. VLDB J 30:739–768. https://doi.org/10.1007/s00778-021-00671-8

11. Mehrabi N, Morstatter F, Saxena N, et al (2022) A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv 54:1–35. https://doi.org/10.1145/3457607

12. Swain S, Bhushan B, Dhiman G, Viriyasitavat W (2022) Appositeness of Optimized and Reliable Machine Learning for Healthcare: A Survey. Arch Comput Methods Eng 29:3981–4003. https://doi.org/10.1007/s11831-022-09733-8

13. Caton S, Haas C (2024) Fairness in Machine Learning: A Survey. ACM Comput Surv 56:166:1-166:38. https://doi.org/10.1145/3616865

14. Smith B, Khojandi A, Vasudevan R (2024) Bias in Reinforcement Learning: A Review in Healthcare Applications. ACM Comput Surv 56:1–17. https://doi.org/10.1145/3609502

15. Chen RJ, Wang JJ, Williamson DFK, et al (2023) Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 7:719–742. https://doi.org/10.1038/s41551-023-01056-8

16. Sharabiani A, Bress A, Douzali E, Darabi H (2015) Revisiting Warfarin Dosing Using Machine Learning Techniques. Comput Math Methods Med 2015:e560108. https://doi.org/10.1155/2015/560108

17. Syn NL, Wong AL-A, Lee S-C, et al (2018) Genotype-guided versus traditional clinical dosing of warfarin in patients of Asian ancestry: a randomized controlled trial. BMC Med 16:104. https://doi.org/10.1186/s12916-018-1093-8

18. Ahmad MA, Patel A, Eckert C, et al (2020) Fairness in Machine Learning for Healthcare. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp 3529–3530

19. Mishler A, Dalmasso N (2022) Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. ArXiv Prepr ArXiv220205049

20. Ge Y, Liu S, Gao R, et al (2021) Towards Long-term Fairness in Recommendation. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 445–453

21. Yang J, Soltan AAS, Eyre DW, Clifton DA (2023) Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nat Mach Intell 5:884–894. https://doi.org/10.1038/s42256-023-00697-3

22. Tang S, Modi A, Sjoding M, Wiens J (2020) Clinician-in-the-Loop Decision Making: Reinforcement Learning with Near-Optimal Set-Valued Policies. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, pp 9387–9396

23. Mingyu Lu (2020) Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Hemodynamic Management in Sepsis Patients - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075511/. Accessed 26 Apr 2024

24. Correa R, Shaan M, Trivedi H, et al (2022) A Systematic Review of 'Fair' AI Model Development for Image Classification and Prediction. J Med Biol Eng 42:816–827. https://doi.org/10.1007/s40846-022-00754-z

25. Paulus JK, Kent DM (2020) Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. Npj Digit Med 3:99. https://doi.org/10.1038/s41746-020-0304-9

26. Xu J, Xiao Y, Wang WH, et al (2022) Algorithmic fairness in computational medicine. eBioMedicine 84:. https://doi.org/10.1016/j.ebiom.2022.104250

27. Grote T, Keeling G (2022) On Algorithmic Fairness in Medical Practice. Camb Q Healthc Ethics 31:83–94. https://doi.org/10.1017/S0963180121000839

28. Munguía-López ADC, Ponce-Ortega JM (2021) Fair Allocation of Potential COVID-19 Vaccines Using an Optimization-Based Strategy. Process Integr Optim Sustain 5:3–12. https://doi.org/10.1007/s41660-020-00141-8

29. Acuna JA, Zayas-Castro JL, Charkhgard H (2020) Ambulance allocation optimization model for the overcrowding problem in US emergency departments: A case study in Florida. Socioecon Plann Sci 71:100747. https://doi.org/10.1016/j.seps.2019.100747

30. Samorani M, Harris SL, Blount LG, et al (2022) Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. Manuf Serv Oper Manag 24:2825–2842. https://doi.org/10.1287/msom.2021.0999

31. Seker E, Talburt JR, Greer ML (2022) Preprocessing to Address Bias in Healthcare Data. In: Challenges of Trustable AI and Added-Value on Health. IOS Press, pp 327–331

32. Aleem S, Huda N ul, Amin R, et al (2022) Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. Electronics 11:1111. https://doi.org/10.3390/electronics11071111

33. James LR (1982) Aggregation bias in estimates of perceptual agreement. J Appl Psychol 67:219–229. https://doi.org/10.1037/0021-9010.67.2.219

34. Dhabliya D, Dari SS, Dhablia A, et al (2024) Addressing Bias in Machine Learning Algorithms: Promoting Fairness and Ethical Design. E3S Web Conf 491:02040. https://doi.org/10.1051/e3sconf/202449102040

35. Nazer LH, Zatarah R, Waldrip S, et al (2023) Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2:e0000278. https://doi.org/10.1371/journal.pdig.0000278

36. Chang C, Deng Y, Jiang X, Long Q (2020) Multiple imputation for analysis of incomplete data in distributed health data networks. Nat Commun 11:5467. https://doi.org/10.1038/s41467-020-19270-2

37. Furukawa MF, Raghu TS, Shao BBM (2010) Electronic Medical Records, Nurse Staffing, and Nurse-Sensitive Patient Outcomes: Evidence from California Hospitals, 1998–2007. Health Serv Res 45:941–962. https://doi.org/10.1111/j.1475-6773.2010.01110.x

38. Hu Z, Melton GB, Arsoniadis EG, et al (2017) Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J Biomed Inform 68:112–120. https://doi.org/10.1016/j.jbi.2017.03.009

39. Jakobsen JC, Gluud C, Wetterslev J, Winkel P (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 17:162. https://doi.org/10.1186/s12874-017-0442-1

40. Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice - FM Hajjaj, MS Salek, MKA Basra, AY Finlay, 2010. https://journals.sagepub.com/doi/full/10.1258/jrsm.2010.100104. Accessed 22 May 2024

41. Ng JH, Ye F, Ward LM, et al (2017) Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. Health Aff (Millwood) 36:548–552. https://doi.org/10.1377/hlthaff.2016.1044

42. Waite S, Scott J, Colombo D (2021) Narrowing the Gap: Imaging Disparities in Radiology. Radiology 299:27–35. https://doi.org/10.1148/radiol.2021203742

43. IsHak W, Nikravesh R, Lederer S, et al (2013) Burnout in medical students: a systematic review. Clin Teach 10:242–245. https://doi.org/10.1111/tct.12014

44. Gopal DP, Chetty U, O'Donnell P, et al (2021) Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J 8:40–48. https://doi.org/10.7861/fhj.2020-0233

45. Scherer RW, Dickersin K, Langenberg P (1994) Full Publication of Results Initially Presented in Abstracts: A Meta-analysis. JAMA 272:158–162. https://doi.org/10.1001/jama.1994.03520020084025

46. Raynaud M, Zhang H, Louis K, et al (2021) COVID-19-related medical research: a meta-research and critical appraisal. BMC Med Res Methodol 21:1. https://doi.org/10.1186/s12874-020-01190-w

47. Yang Z, Zhang Y, Mosler EL, et al (2020) Topical benzoyl peroxide for acne. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD011154.pub2

48. Datta A, Tschantz MC, Datta A (2014) Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. ArXiv Prepr ArXiv14086491

49. Wen M, Bastani O, Topcu U (2021) Algorithms for Fairness in Sequential Decision Making. In: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. PMLR, pp 1144–1152

50. Biswas S, Rajan H (2021) Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, Athens Greece, pp 981–993

51. Wan M, Zha D, Liu N, Zou N (2023) In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. ACM Trans Knowl Discov Data 17:1–27. https://doi.org/10.1145/3551390

52. Petersen F, Mukherjee D, Sun Y, Yurochkin M (2021) Post-processing for Individual Fairness. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp 25944–25955

53. Kamiran F, Calders T (2010) Classification with no discrimination by preferential sampling. Citeseer

54. Nilsson A, Bonander C, Strömberg U, et al (2021) 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. https://doi.org/10.1371/journal.pone.0253969

55. Kumar N, Shrestha R, Li Z, Wang L (2023) Distributionally Robust Optimization and Invariant Representation Learning for Addressing Subgroup Underrepresentation: Mechanisms and Limitations. In: Wesarg S, Puyol Antón E, Baxter JSH, et al (eds) Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging. Springer Nature Switzerland, Cham, pp 183–193

56. Peacock WF, Slatkin N, Gagnon-Sanschagrin P, et al (2022) Opioid-Induced Constipation: Cost Impact of Approved Medications in the Emergency Department. Adv Ther 39:2178–2191. https://doi.org/10.1007/s12325-022-02090-9

57. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953

58. Borland D, Zhang J, Kaul S, Gotz D (2021) Selection-Bias-Corrected Visualization via Dynamic Reweighting. IEEE Trans Vis Comput Graph 27:1481–1491. https://doi.org/10.1109/TVCG.2020.3030455

59. Mohamed R, Azizan NH, Perumal T, et al (2023) Discovering and Recognizing of Imbalance Human Activity in Healthcare Monitoring using Data Resampling Technique and Decision Tree Model. J Adv Res Appl Sci Eng Technol 33:340–350. https://doi.org/10.37934/araset.33.2.340350

60. Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A (2018) Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. Proc AAAI Conf Artif Intell 32:. https://doi.org/10.1609/aaai.v32i1.11296

61. Briscik M, Dillies M-A, Déjean S (2023) Improvement of variables interpretability in kernel PCA. BMC Bioinformatics 24:282. https://doi.org/10.1186/s12859-023-05404-y

62. Hajjar J, Ma W, Vosoughi S (2022) DartmouthCS at SemEval-2022 task 8: Predicting multilingual news article similarity with meta-information and translation. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). pp 1157–1162

63. Zhou B, Yang G, Shi Z, Ma S (2022) Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng 1–17. https://doi.org/10.1109/RBME.2022.3210270

64. Minot JR, Cheney N, Maier M, et al (2022) Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. ACM Trans Comput Healthc 3:1–41

65. Moreno-Serra R, Anaya-Montes M, León-Giraldo S, Bernal O (2022) Addressing recall bias in (post-)conflict data collection and analysis: lessons from a large-scale health survey in Colombia. Confl Health 16:14. https://doi.org/10.1186/s13031-022-00446-0

66. Fulton BR, Bilgen I, Pineau V, et al (2022) Evaluating Nonresponse Bias for a Hypernetwork Sample Generated from a Probability-Based Household Panel. J Quant Methods 6:1–25. https://doi.org/10.29145/jqm.0602.03

67. Uhde A, Schlicker N, Wallach DP, Hassenzahl M (2020) Fairness and Decision-making in Collaborative Shift Scheduling Systems. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, pp 1–13

68. Radovanović S, Delibašić B, Marković A, Suknović M (2022) Achieving MAX-MIN Fair Cross-efficiency scores in Data Envelopment Analysis. Proc 55th Hawaii Int Conf Syst Sci

69. Ala A, Alsaadi FE, Ahmadi M, Mirjalili S (2021) Optimization of an appointment scheduling problem for healthcare systems based on the quality of fairness service using whale optimization algorithm and NSGA-II. Sci Rep 11:19816. https://doi.org/10.1038/s41598-021-98851-7

70. Zhong X, Zhang J, Zhang X (2017) A two-stage heuristic algorithm for the nurse scheduling problem with fairness objective on weekend workload under different shift designs. IISE Trans Healthc Syst Eng 7:224–235. https://doi.org/10.1080/24725579.2017.1356891

71. Neophytou N, Taik A, Farnadi G (2024) Promoting Fair Vaccination Strategies through Influence Maximization: A Case Study on COVID-19 Spread. Proc AAAI Conf Artif Intell 38:22285–22293. https://doi.org/10.1609/aaai.v38i20.30234

72. Wolbeck L, Kliewer N, Marques I (2020) Fair shift change penalization scheme for nurse rescheduling problems. Eur J Oper Res 284:1121–1135. https://doi.org/10.1016/j.ejor.2020.01.042

73. Gunnarsson B, Björnsdóttir KM, Dúason S, Ingólfsson Á (2023) Locating helicopter ambulance bases in Iceland: efficient and fair solutions. Scand J Trauma Resusc Emerg Med 31:70. https://doi.org/10.1186/s13049-023-01114-9

74. Sepúlveda IA, Aguayo MM, De la Fuente R, et al (2024) Scheduling mobile dental clinics: A heuristic approach considering fairness among school districts. Health Care Manag Sci 27:46–71. https://doi.org/10.1007/s10729-022-09612-5

75. Klyve KK, Senthooran I, Wallace M (2023) Nurse rostering with fatigue modelling. Health Care Manag Sci 26:21–45. https://doi.org/10.1007/s10729-022-09613-4

76. Gross CN, Brunner JO, Blobner M (2019) Hospital physicians can't get no long-term satisfaction – an indicator for fairness in preference fulfillment on duty schedules. Health Care Manag Sci 22:691–708. https://doi.org/10.1007/s10729-018-9452-8

77. Akshat S, Gentry SE, Raghavan S (2024) Heterogeneous donor circles for fair liver transplant allocation. Health Care Manag Sci 27:20–45. https://doi.org/10.1007/s10729-022-09602-7

78. Azizi S, Aygül Ö, Faber B, et al (2023) Select, route and schedule: optimizing community paramedicine service delivery with mandatory visits and patient prioritization. Health Care Manag Sci 26:719–746. https://doi.org/10.1007/s10729-023-09646-3

79. Proano RA, Agarwal A (2018) Scheduling internal medicine resident rotations to ensure fairness and facilitate continuity of care. Health Care Manag Sci 21:461–474. https://doi.org/10.1007/s10729-017-9403-9

80. Lodi A, Olivier P, Pesant G, Sankaranarayanan S (2024) Fairness over time in dynamic resource allocation with an application in healthcare. Math Program 203:285–318. https://doi.org/10.1007/s10107-022-01904-6

81. Argyris N, Karsu Ö, Yavuz M (2022) Fair resource allocation: Using welfare-based dominance constraints. Eur J Oper Res 297:560–578. https://doi.org/10.1016/j.ejor.2021.05.003

82. Rastegar M, Tavana M, Meraj A, Mina H (2021) An inventory-location optimization model for equitable influenza vaccine distribution in developing countries during the COVID-19 pandemic. Vaccine 39:495–504. https://doi.org/10.1016/j.vaccine.2020.12.022

83. Ala A, Simic V, Pamucar D, Tirkolaee EB (2022) Appointment Scheduling Problem under Fairness Policy in Healthcare Services: Fuzzy Ant Lion Optimizer. Expert Syst Appl 207:117949. https://doi.org/10.1016/j.eswa.2022.117949

84. Yin X, Büyüktahtakın İE (2021) A multi-stage stochastic programming approach to epidemic resource allocation with equity considerations. Health Care Manag Sci 24:597–622. https://doi.org/10.1007/s10729-021-09559-z

85. Yu G, Siddique U, Weng P (2023) Fair Deep Reinforcement Learning with Preferential Treatment. In: Gal K, Nowé A, Nalepa GJ, et al (eds) Frontiers in Artificial Intelligence and Applications. IOS Press

86. Li Y, Mao C, Huang K, et al (2023) Deep Reinforcement Learning for Efficient and Fair Allocation of Health Care Resources. ArXiv Prepr ArXiv230908560

87. Atwood J, Srinivasan H, Halpern Y, Sculley D (2019) Fair treatment allocations in social networks. ArXiv Prepr ArXiv191105489

88. Budhiraja I, Kumar N, Tyagi S (2021) Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication. IEEE Internet Things J 8:3143–3156. https://doi.org/10.1109/JIOT.2020.3014926

89. Keya KN, Islam R, Pan S, et al (2020) Equitable allocation of healthcare resources with fair cox models. ArXiv Prepr ArXiv201006820

90. Ge Y, Zhao X, Yu L, et al (2022) Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 316–324

91. Achterberg T, Wunderling R (2013) Mixed Integer Programming: Analyzing 12 Years of Progress. In: Jünger M, Reinelt G (eds) Facets of Combinatorial Optimization: Festschrift for Martin Grötschel. Springer, Berlin, Heidelberg, pp 449–481

92. Fouskakis D, Draper D (2002) Stochastic Optimization: a Review. Int Stat Rev 70:315–349. https://doi.org/10.1111/j.1751-5823.2002.tb00174.x

93. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge, Mass

94. Chakraborty S, Tomsett R, Raghavendra R, et al (2017) Interpretability of deep learning models: A survey of results. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). pp 1–6

95. Puterman ML (2014) Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons

96. Kleinbaum DG, Klein M (1996) Survival analysis a self-learning text. Springer

97. Roijers DM, Vamplew P, Whiteson S, Dazeley R (2013) A survey of multi-objective sequential decision-making. J Artif Intell Res 48:67–113

98. Altman E (2021) Constrained Markov Decision Processes. Routledge, New York

99. Kim MP, Ghorbani A, Zou J (2019) Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, pp 247–254

100. Lu M, Shahn Z, Sow D, et al (2020) Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients. American Medical Informatics Association, p 773

101. Yu C, Liu J, Nemati S, Yin G (2023) Reinforcement Learning in Healthcare: A Survey. ACM Comput Surv 55:1–36. https://doi.org/10.1145/3477600

102. Field DA (1988) Laplacian smoothing and Delaunay triangulations. Commun Appl Numer Methods 4:709–712. https://doi.org/10.1002/cnm.1630040603

103. Jo N, Tang B, Dullerud K, et al (2023) Fairness in Contextual Resource Allocation Systems: Metrics and Incompatibility Results. Proc AAAI Conf Artif Intell 37:11837–11846. https://doi.org/10.1609/aaai.v37i10.26397

104. Tang S, Modi A, Sjoding MW, Wiens J Clinician-in-the-Loop Decision Making: Reinforcement Learning with Near-Optimal Set-Valued Policies

105. Lu M, Shahn Z, Sow D, et al Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Hemodynamic Management in Sepsis Patients

106. Cohen IG, Mello MM (2018) HIPAA and Protecting Health Information in the 21st Century. JAMA 320:231–232. https://doi.org/10.1001/jama.2018.5630

107. REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation)

108. Elmachtoub AN, Grigas P (2022) Smart "Predict, then Optimize." Manag Sci 68:9–26. https://doi.org/10.1287/mnsc.2020.3922