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

Zequn Chen, Master.; zequn.chen.th@dartmouth.edu; (603) 729-6285; 15 Thayer Dr, Hanover, NH 03755 [1]
Wesley J. Marrero, Ph.D.; wesley.marrero@dartmouth.edu; (603) 646-3457; 15 Thayer Dr, Hanover, NH 03755

Word count: 4005 words

[1] Financial support for this study was provided entirely by Thayer School of Engineering from Dartmouth College. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The following authors are employed by the sponsor: Zequn Chen and Wesley J. Marrero.

Background. Unintended biases introduced by optimization and machine learning (ML) models are of great interest to medical professionals. Bias in healthcare decisions can cause patients from vulnerable populations (e.g., racially minoritized, low-income) to have lower access to resources, exacerbating societal unfairness. **Purpose.** This review aims to identify, describe, and categorize literature regarding bias types, fairness metrics, and bias mitigation methods in healthcare decision making. **Data Sources.** Google Scholar database was searched to identify published studies. **Study Selection.** Eligible studies were required to present 1) types of bias 2) fairness metrics and 3) bias mitigation methods within decision-making in healthcare. **Data Extraction.** Studies were classified according to the three themes mentioned in the "Study Selection". Information was extracted concerning the definitions, examples, applications, and limitations of bias types, fairness metrics, and bias mitigation methods. **Data Synthesis.** In bias type section, we included studies (n=15) concerning different biases. In the fairness metric section, we included studies (n=6) regarding common fairness metrics. In bias mitigation method section, themes included pre-processing methods (n=5), in-processing methods (n=16), and post-processing methods (n=4). **Limitations.** Most examples in our survey are from the United States since the majority of studies included in this survey were conducted in the United States. In the meanwhile, we limited the search language to English, so we may not capture some meaningful articles in other languages. **Conclusions**. Several types of bias, fairness metrics, and bias mitigation methods (especially optimization and machine learning-based methods) were identified in this review, with common themes based on analytical approaches. We also found topics such as explainability, fairness metric selection, and integration of prediction and optimization are promising directions for future studies.

Highlights:

- This review aims to articulate common bias types and fairness metrics and delves into applications of bias mitigation methods within the context of medical decision making.
- We explored optimization-based and machine learning-based methodologies for medical decision-making applications in a detailed manner.
- The relationship and restrictions of various fairness metrics were analyzed, which can help people understand and select appropriate fairness metrics based on the concrete scenario.
- We investigated multiple bias mitigation technologies that have not been applied in healthcare but can be easily extended to healthcare settings.

Keywords: fairness, decision-making, healthcare, optimization, machine learning

An increasing number of healthcare researchers and practitioners are leveraging quantitative methodologies to improve decision-making. Optimization and machine learning (ML) models have proven to be extremely helpful in medical decision-making and health policy[1]. However,

such advanced decision-making methods can lead to inequitable outcomes as they do not give sufficient attention to underrepresented groups[2], including those who are racially minoritized, low-income, or living in rural areas. Our survey presents a review of fair decision-making in healthcare, revealing how we can leverage methods such as fair optimization and ML to ensure people from disadvantaged groups also have sufficient access to healthcare.

The biases introduced by optimization and ML algorithms have become of special interest to practitioners and researchers[3,4]. Medical policies generated by reinforcement learning often afford black veterans fewer opportunities to receive cardiovascular screenings compared to white veterans[5]. Convolutional neural networks routinely underdiagnose Hispanic patients at a higher rate because Hispanic patients have limited access to healthcare resources[6]. Such bias may cause decision-makers to distribute fewer medical resources to racially minoritized subgroups. The COMPAS algorithm by Equivant is widely used by the American police system to evaluate the risk of recidivists and thus it influences sentencing decisions in court. However, racial biases in the outputs of the algorithm have been documented, specifically the underprediction of recidivism risks for white defendants and the overprediction of risks for black defendants.

Current literature reviews focus on the biases brought or perpetuated by ML prediction settings. For example, Ahmad and collaborators show that ML algorithms may show fewer satisfying outcomes in underrepresented groups since they have fewer data points from these populations[8]. Mishler and coauthors reveal that predictors are sensitive to distributions, and incorporating fairness definitions in predictors can help avoid fairness issues[9].

The main difference between our work and the previous reviews is instead of prediction, we focus on fairness within the context of decision-making. It is worth noting our review has some overlaps with Smith et al.'s survey[5]. The major differences between our surveys are: 1) their work focuses on fairness in reinforcement learning exclusively, while our paper explores other in-process techniques like optimization and classification; 2) we review different pre-processing techniques, such as fair data transformers and natural language processing; 3) we review post-processing methods such as Laplacian smoothing and multi-accuracy approaches. The fairness-enhanced reinforcement learning in healthcare articles cited in Smith et al. are included in our review for completion purposes[10-13]. However, we refer the interested reader to their review for an in-depth description of these works.

Methods

Search Strategy

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 one of the terms in "bias", "fairness", or "equity", and one of the terms in "optimization", "machine learning", "deep learning", "reinforcement learning",

"game", or "network". We include all records mentioning these keywords in the publication title, abstract, or full text. Chosen records focus on one of three domains: 1) bias types in decision-making; 2) fairness metrics; or 3) bias mitigation methods for healthcare decision-making. The publication language is restricted to English. The search's last update was on February 21, 2024.

Study Eligibility

The articles' titles and abstracts were screened by [], then the selected manuscripts were double-checked by []. Records were excluded if the publications: 1) did not address healthcare topics and cannot be extended to healthcare easily; 2) did not focus on decision making (e.g., papers focusing on predictions); 3) only included introductory text or conference abstract; or 4) did not focus on methodology. Of the remaining publications on methodology or review of fair decision-making in healthcare, the full text was screened by [] before the final discussion with []. The articles and surveys were categorized into three sections: bias in healthcare, fairness metrics, and bias mitigation methods. Furthermore, section of bias mitigation methods only includes articles mentioning an application of decision-making in healthcare with a methodological focus (i.e., we chose papers that use fair decision-making methodologies/algorithms to address healthcare issues). Discordance between reviewers was settled by discussion until the consensus had been achieved.

Fair decision-making concepts (i.e., types of bias, fairness metrics, or bias mitigation approaches) were extracted from the full text of the selected works by []. To ensure the accuracy and completeness of the extracted aspects, the selected concepts were double-checked by []. All chosen concepts were grouped into broad and non-overlapping sections to reflect their relations.

Results

Review Process

The systematic review led to 642 records, 158 of them were unrelated to the healthcare domain, and 176 of them did not focus on decision-making. Furthermore, 68 records were excluded because they were tutorial or conference abstracts. Of the remaining articles, 153 papers discuss fair decision making in healthcare without a focus on methodology. In the 87 papers left, we evaluate the full text. Twenty additional articles were excluded since they did not meet the inclusion criteria mentioned above. Of the remaining 67 papers, 18 were review papers. The flow diagram for the literature review is shown in Figure 1.

Study Characteristics

Of the 67 included papers, all concepts related to fair decision-making in healthcare were extracted. We create a structured and non-overlapping coverage. *Bias in Healthcare Decision Making* section categorizes biases into three categories: 1) algorithmic bias; 2) data bias; and 3) publication bias. *Fairness Metrics* section classifies fairness metrics as one of the following four types: 1) fairness through unawareness; 2) demographic parity; 3) equal opportunity; and 4) equal odds. *Bias Mitigation* section divides bias mitigation methods into three classes: pre-processing methods, in-processing methods, and post-processing methods.

Bias in Healthcare Decision Making

Biases can exist in data and algorithms, 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. *Bias Mitigation* section of our survey will summarize methods addressing these biases.

Algorithmic Bias

Algorithmic bias stems from computational procedures failing to consider fairness in their execution[7]. For example, some vaccine allocation algorithms set overall social welfare as their objective function. This objective may exacerbate demographic disparities since underrepresented populations may have less access to vaccines[14]. Many optimization-based ambulance allocation models set the overall survival rate as the sole objective[15]. These models fail to take fairness into account because they do not consider ambulance availability across different populations, such as people with lower socioeconomic status. Algorithmic bias also exists in ML models. For instance, Samorani and coauthors have found that machine-learning-based scheduling models have a higher likelihood of assigning black people to overbooked slots because they have a higher historical no-show rate[16].

Data Bias

Data bias refers to the unfairness generated by prejudiced data sources[17]. Socioeconomic and racial disparities in medical resource availability may lead to skewed datasets[18]. Two common data biases are aggregation biases and representation biases. Aggregation bias refers to the effect of aggregating data without considering disparities among subgroups[3,19]. Representation bias occurs when the data is not comprehensive and cannot represent the actual situation[20]. For instance, certain providers may have fewer electronic health records (EHR) about people from lower socioeconomic status as they may have limited access to healthcare. If decision-making models are built with data underrepresenting this population, the developed models may be biased against people with lower socioeconomic status[21,22].

Another source of data bias is response bias, which occurs when data are labeled inconsistently or collected by unreliable methods. Response bias frequently happens in self-reported data or surveys because of some participants' inaccurate answers[23]. Since policymakers may harness data to make public health decisions, response bias can skew decision-making of some subpopulations[24].

Publication Bias

Publication bias happens when the researchers' decision to publish a paper or not depends on the study results. Compared to studies without positive results, it is easier to publish medical studies with statistically significant and positive results[25]. This behavior may lead to overestimation of certain clinical treatments since only studies with satisfying outcomes are published. Medical practitioners may then make treatment decisions based on biased outcomes, giving rise to degraded patient outcomes for certain subgroups[26,27].

Fairness Metrics

In this section, we present the evaluation of four types of biases: fairness through unawareness, demographic parity, equal opportunity, and equal odds.

Fairness Through Unawareness

Fairness through unawareness is the base fairness metric[6]. It does not consider any sensitive attribute during the decision-making process. However, simply ignoring sensitive attributes may not remove inequity, because other variables can be highly correlated with the sensitive traits. This method has been proven to be invalid in many cases[28].

Demographic Parity

The demographic parity aims to ensure the actions generated from a decision-making model are independent of a sensitive attribute in the whole population[2]. Independence of sensitive attributes indicates the outcomes (i.e., expected cumulative rewards) must be equivalent in privileged and unprivileged groups[29]. The problem with demographic parity is it does not consider the population's ground-truth qualification.

Equal Opportunity

Similar to demographic parity, equal opportunity verifies whether the algorithmic recommendations for privileged and unprivileged groups are the same[7]. However, demographic parity applies to the entire population while equal opportunity applies solely to a truly qualified population. However, equal opportunity fails to investigate fairness among truly unqualified people.

Equal Odds

The equal odds ensures a decision-making model performs equally well across all groups[30]. It is more rigorous than demographic parity and equal opportunity since it not only requires decisions to be independent of a sensitive feature but also has equal outcomes for people with

different ground-truth outcomes. This metric states that algorithmic outcomes for majority and minority groups should be close among truly qualified and truly unqualified groups.

In summary, fairness through unawareness is the most straightforward metric but is invalid in many settings. Demographic parity evaluates if decision-making is independent of sensitive attributes. Equal opportunity measures whether decision-making is independent of sensitive attributes among a qualified group, which applies to smaller populations compared to demographic parity. Equal odds is the most rigorous metric since it ensures that algorithmic outcomes are independent of sensitive attributes among qualified and unqualified groups separately.

Bias Mitigation

In this section, we summarize different bias mitigation approaches used across healthcare domain. 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 fed for constructing decision-making models. In-process methodologies refer to building unbiased algorithms directly. Post-processing methods calibrate the algorithmic outcomes to achieve fairness[31–33]. The descriptions of pre-processing, in-processing, and post-processing bias mitigation methods are summarized in Supplemental Table 1 in the Appendix.

Pre-Processing

Datasets may be biased due to the sources discussed in *Bias in Healthcare Decision Making* section, which can cause skewed decisions. For instance, when a dataset is imbalanced, decisions may be skewed toward subpopulations with greater sizes[34]. Pre-processing methods can help circumvent possible biases in this setting. There are four popular approaches for pre-processing: reweighting the underrepresented populations, resampling, natural language processing, and fair data transformers. Reweighting the underrepresented populations assigns greater weights to data of underrepresented subgroups[3,36,39,40], so the model can give sufficient attention to vulnerable populations. Resampling ensures the data is balanced (i.e., has an equal number of instances from each subgroup)[34,37,41], hence the model can learn adequate information from all subgroups. Natural language processing gives computers the capability to understand and manipulate text like human beings[35,38,42]. Natural language processing can remove sensitive attributes from text data before feeding the data to decision-making models. A fair data transformer is a processor that can extract feature vectors (i.e., numeric representations of an object of interest) from the input data in a fair way[32,44]. Though fair data transformers have not been applied to healthcare so far, we can easily extend this approach to clinical areas. These four methods along with their respective reference, areas of application, and targeted fairness metrics are summarized in Table 2.

In-Processing

In-processing methodologies refer to building solution techniques to lessen the effect of biases in input data. Generally, there are two main branches under the umbrella of in-processing bias mitigation methods: optimization-based and ML-based methods. In-processing bias mitigation techniques are attracting increased attention within domains such as medical resource allocation, scheduling, and clinical treatment[16,45]. Table 3 demonstrates the methods and the corresponding references and applications.

1. Optimization-based Techniques

There are mainly two ways to achieve fair decision-making through optimization algorithms: adding fair constraints or constructing fair objective function. Emergency department overcrowding has become a nationwide crisis in the last decade[46]. To resolve the overcrowding issue, researchers have applied mixed integer programming to build fair medical resource distribution constraints. Mixed integer programming is a type of constrained optimization problem that allows for integer and continuous variables in its objective and constraints[55]. An example of fair constraints is that Acuna and coauthors added equity constraints to ensure the minimal quality of care for every emergency is greater or equal to a threshold in an ambulance allocation situation[46].

Researchers have also leveraged stochastic optimization techniques to generate in-processing bias mitigation techniques. stochastic optimization optimizes an objective function while representing uncertainty through probability distributions[56]. To optimize patient's waiting time, it is possible to add fair constraints in stochastic optimization models to limit the expected difference between the maximum waiting time and minimum waiting time.[47]

Another ubiquitous way to fulfill fairness requirements in healthcare decision-making is to modify the objective function of an optimization model. When medical resources are scarce, people from vulnerable groups such as low socioeconomic status 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.[14] Such fair objective ensures each subgroup receives necessary medical support.

2. Machine Learning-based Techniques

We also witness the application of ML algorithms for fair healthcare decision-making. 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[57]. Deep learning uses multiple neural network layers and activation functions to extract new features from the input data[52], and deep learning is capable of recapitulating and modeling complex patterns in data. Deep reinforcement learning is a combination of deep learning and reinforcement learning. Deep reinforcement learning can be applied to the Markov decision process (MDP)

model. 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[58]. In practice, an MDP may encompass a massive number of system configurations (i.e., states), becoming computationally intractable by traditional reinforcement learning methods. However, deep reinforcement learning can take advantage of deep learning to represent a policy (i.e., sequence of procedures for decision-making at each state) as a neural network and learn to find a policy that optimizes model outcomes (i.e., rewards)[51]. Yang et. al redefines the rewards of deep reinforcement learning to achieve fairness[12]. In their approach, the rewards of a certain subgroup are adjusted by the group size, and subgroups with smaller sizes are assigned with greater rewards. This definition of rewards helps the model give more attention to minority groups.

Fair survival models provide an additional tool for decision-making in healthcare settings[54]. Traditional survival analysis estimates the time until an event of interest. Fair survival models incorporate event probabilities and fairness metric violations. Specifically, the objective function of fair survival model incorporates both the log-likelihood of a Cox proportional-hazards model, which measures the probability of getting a disease during a certain period, and the fairness metric, which is the difference between the largest and lowest disease probabilities within a group[60]. Thereafter, they feed the input data to train the model. The fair model's outcome is used to generate a waitlist of patients, which decides the sequence of resource allocation.

While it has not been applied to healthcare settings yet, 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. Ge and coauthors have applied the Pareto frontier to identify the policy that optimizes weights of objectives regarding utility and fairness elements[53]. Specifically, they apply reinforcement learning to learn the optimal weight for each objective. Their result shows there exists a trade-off between utility and fairness performance, and we can choose the final recommendation based on user preferences[53]. Though multi-objective MDP has not been applied to fair decision-making in healthcare yet, 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 solutions that consider both vaccine utility and distribution fairness.

The Constrained Markov Decision Process (CMDP) is another prospective direction. Compared to traditional MDP, CMDP can accommodate fair constraints in deep learning framework. In CMDP, we can formulate the cost function regarding fairness, and we can only choose policies leading to fairness cost less or equal to the threshold[13]. This method has not been utilized in fair healthcare decision-making yet. 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 cost.

Post-Processing

Post-processing methods calibrate algorithmic outcomes to achieve fairness[31]. The Laplacian smoothing method is a technique to reduce the noise of the data while preserving the important characteristics of the solution technique[62]. We can take advantage of Laplacian smoothing to guarantee comparable results of similar individuals while preserving satisfying cost of loss function[31]. Laplacian smoothing can be extended to healthcare. For example, after a reinforcement learning algorithm produces treatment plans, a Laplacian smoothing method can guarantee comparable treatment plans are assigned to similar individuals.

We can also apply multi-accuracy approach to combine several weak learners to achieve high accuracy rates among all subpopulation groups[63]. After obtaining results from several pre-trained weak learners, multi-accuracy can assign larger weights to samples that are identified incorrectly in weak learners. Subsequently, the following weak learners pay extra attention to mistaken samples from underrepresented subpopulations and adjust their results accordingly[64]. Multi-accuracy can produce accurate and fair classification, therefore, physicians can deploy algorithm-based diagnoses to achieve fair medical outcomes.

Lastly, the clinical expertise of medical practitioners may help increase fairness in ML algorithms[10,11,61]. For example, reinforcement learning techniques can suggest several near-equivalent actions, then we can rely on clinicians' opinions to decide what actions can lead to the fairest outcome. This approach enables improved decisions while adhering to clinical standards, and it leverages practitioners' experience. Therefore, by embedding clinical suggestions, we can reduce bias in an explainable manner. The post-processing bias mitigation methods included are shown in Table 4.

Discussion

This review summarized the state-of-the-art fair decision-making approaches in healthcare settings. We found that even though a plethora of fairness methods have been proposed, most of them focused on prediction rather than decision-making.

Major Contributions

The main takeaway of the survey is that we explore two kinds of in-processing methodologies in a detailed manner: optimization-based methodologies and machine learning-based methodologies. Optimization-based techniques fulfill fairness by modifying objective functions or adding fair constraints. Mixed-integer programming and stochastic optimization are the mainstream methodologies to implement optimization-based models. Machine learning-based

approaches incorporate fairness in loss functions, and then learn parameters or decision rules to minimize the loss function. The survey investigates three machine learning-based techniques to achieve fairness: reinforcement learning, deep learning, and fair survival analysis.

Another contribution of the survey is that we explored multiple bias mitigation technologies that have not been applied in healthcare and illustrated how they may be extended to healthcare settings. Such bias mitigation techniques covered in the survey are the fair data transformer in pre-processing, several reinforcement learning techniques in in-processing, and Laplacian smoothing and multi-accuracy in post-processing.

Last but not least, we discuss the relationship and restrictions of fairness metrics mentioned in the survey. Based on metrics' relationship and properties, medical practitioners and researchers can decide what metric they want to employ based on the specific context.

Given the growing importance of optimization and machine learning-based decision-making approaches in healthcare, fairness considerations and bias-mitigation approaches are becoming increasingly vital. Our survey may aid practitioners in 1) understanding potential sources of bias in decision-making; 2) choosing the appropriate fairness metric before making decisions; and 3) selecting the appropriate pre-processing, in-processing, and post-processing techniques to reduce bias. In conclusion, this survey has shed 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 and just healthcare outcomes for all individuals.

Future Research Directions

Based on the current literature regarding fair decision making in healthcare, we find several promising directions to explore in the future. The first promising field is algorithm explainability. Many decision-making algorithms in healthcare are considered as black boxes, which are hard to understand. This lack of explainability is an obstacle for practitioners to identify if the model is relying on biased features[52,65]. Another emerging field is the study of what fairness metric to apply under a certain context. Researchers have found that different fairness metrics can be incompatible[64, 66], so we cannot expect a model to satisfy all fairness metrics. It is also worthwhile to cross the gap between prediction and fair decision-making. Current research usually follows a "prediction then optimization" pipeline, but we can explore innovative models to incorporate in the loss function the decision error induced by prediction[67]. Therefore, the model can achieve fair prediction and fair optimization concurrently.

Study Limitations

The survey contains limitations. Since the most relevant research projects to this review were conducted in the United States, most examples in our paper were cases in this country. Thus, the review may not sufficiently reflect the reality in other parts of the world. In addition, we limit our search to articles in English, so we are unable to capture insightful publications in other

languages. Finally, we might have missed keywords during our literature review and did not capture some works that used excluded terms.

Conclusion

This study provides how to fulfill fair medical decision-making, especially via optimization and machine learning. It is worth noting that most studies of fair methods were conducted within prediction rather than decision-making. We first presented different categories of biases for data and models, including algorithmic bias, data bias, and publication bias. Then, we described multiple fairness metrics that have been used to evaluate the model's fairness, including fairness through unawareness, demographic parity, equal opportunity, and equal odds. The survey categorizes the literature on fair decision-making methods in healthcare into pre-processing, in-processing, and post-processing bias mitigation methods. The pre-processing section articulates how to adjust or transform data to avoid unfairness, which includes methods of reweighting, resampling, natural language processing, and fair data transformer. In-processing section summarizes modeling methodologies to lessen the effect of biased input data. We focus on optimization-based methodologies and machine learning-based methodologies for fair decision-making, and clinical applications of these methodologies are also covered. The post-processing section elucidates methods to calibrate algorithmic outcomes to accomplish fairness.

Fairness in decision-making is an emerging field, poised to substantially reduce social inequities and improve the overall well-being of underrepresented subgroups. Our review can increase the awareness of fairness in healthcare decision making, as well as facilitate the selection of appropriate fairness metrics and decision-making approaches under varying scenarios.

Acknowledgments

We would like to express our sincere gratitude to [Name of Person or Organization] for their invaluable assistance and support in [specific contribution]. We also thank [Name of Person or Organization] for providing [specific resource or support]. Our appreciation extends to [Name of Person or Organization] for [specific contribution].

References

1. Mansell J, Lee Rhea C, Murray GR. Predicting the issuance of COVID-19 stay-at-home orders in Africa: Using machine learning to develop insight for health policy research. *Int J Disaster Risk Reduct*. 2023;88:103598. doi:10.1016/j.ijdrr.2023.103598

2. Caton S, Haas C. Fairness in Machine Learning: A Survey. Published online October 4, 2020. Accessed May 4, 2023. http://arxiv.org/abs/2010.04053

3. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv*. 2022;54(6):1-35. doi:10.1145/3457607

4. Swain S, Bhushan B, Dhiman G, Viriyasitavat W. Appositeness of Optimized and Reliable Machine Learning for Healthcare: A Survey. *Arch Comput Methods Eng*. 2022;29(6):3981-4003. doi:10.1007/s11831-022-09733-8

5. Smith B, Khojandi A, Vasudevan R. Bias in Reinforcement Learning: A Review in Healthcare Applications. *ACM Comput Surv*. Published online July 18, 2023:3609502. doi:10.1145/3609502

6. Chen RJ, Chen TY, Lipkova J, et al. Algorithm Fairness in AI for Medicine and Healthcare. Published online March 23, 2022. Accessed June 24, 2023. http://arxiv.org/abs/2110.00603

7. Grote T, Keeling G. On Algorithmic Fairness in Medical Practice. *Camb Q Healthc Ethics*. 2022;31(1):83-94. doi:10.1017/S0963180121000839

8. Ahmad MA, Patel A, Eckert C, Kumar V, Teredesai A. Fairness in Machine Learning for Healthcare. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM; 2020:3529-3530. doi:10.1145/3394486.3406461

9. Mishler A, Dalmasso N. Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings. Published online February 10, 2022. Accessed July 6, 2023. http://arxiv.org/abs/2202.05049

10. Lu M, Shahn Z, Sow D, Doshi-Velez F, Lehman L wei H. Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Hemodynamic Management in Sepsis Patients.

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

12. Yang J, Soltan AAS, Eyre DW, Clifton DA. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat Mach Intell*. 2023;5(8):884-894. doi:10.1038/s42256-023-00697-3

13. Ge Y, Liu S, Gao R, et al. Towards Long-term Fairness in Recommendation. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM; 2021:445-453. doi:10.1145/3437963.3441824

14. Munguía-López ADC, Ponce-Ortega JM. Fair Allocation of Potential COVID-19 Vaccines Using an Optimization-Based Strategy. *Process Integr Optim Sustain*. 2021;5(1):3-12. doi:10.1007/s41660-020-00141-8

15. Knight VA, Harper PR, Smith L. Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*. 2012;40(6):918-926. doi:10.1016/j.omega.2012.02.003

16. Samorani M, Harris SL, Blount LG, Lu H, Santoro MA. Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. *Manuf Serv Oper Manag*. 2022;24(6):2825-2842. doi:10.1287/msom.2021.0999

17. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques.

18. Aleem S, Huda NU, Amin R, Khalid S, Alshamrani SS, Alshehri A. Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions. *Electronics*. 2022;11(7):1111. doi:10.3390/electronics11071111

19. James LR. Aggregation Bias in Estimates of Perceptual Agreement.

20. Li Y, Vasconcelos N. REPAIR: Removing Representation Bias by Dataset Resampling. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2019:9564-9573. doi:10.1109/CVPR.2019.00980

21. Ng JH, Ye F, Ward LM, Haffer SC "Chris", Scholle SH. Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. *Health Aff (Millwood)*. 2017;36(3):548-552. doi:10.1377/hlthaff.2016.1044

22. Waite S, Scott J, Colombo D. Narrowing the Gap: Imaging Disparities in Radiology. *Radiology*. 2021;299(1):27-35. doi:10.1148/radiol.2021203742

23. IsHak W, Nikravesh R, Lederer S, Perry R, Ogunyemi D, Bernstein C. Burnout in medical students: a systematic review. *Clin Teach*. 2013;10(4):242-245. doi:10.1111/tct.12014

24. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J*. 2021;8(1):40-48. doi:10.7861/fhj.2020-0233

25. Scherer RW. Full Publication of Results Initially Presented in Abstracts: A Meta-analysis. *JAMA*. 1994;272(2):158. doi:10.1001/jama.1994.03520020084025

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

27. Yang Z, Zhang Y, Lazic Mosler E, et al. Topical benzoyl peroxide for acne. Cochrane Skin Group, ed. *Cochrane Database Syst Rev*. 2020;2020(3). doi:10.1002/14651858.CD011154.pub2

28. Datta A, Tschantz MC, Datta A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. Published online March 16, 2015. Accessed August 28, 2023. http://arxiv.org/abs/1408.6491

29. Wen M, Bastani O, Topcu U. Algorithms for fairness in sequential decision making. In: PMLR; 2021:1144-1152.

30. Chen RJ. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7.

31. Petersen F, Sun Y, Mukherjee D, Yurochkin M. Post-processing for Individual Fairness.

32. Biswas S, Rajan H. 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; 2021:981-993. doi:10.1145/3468264.3468536

33. Wan M, Zha D, Liu N, Zou N. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans Knowl Discov Data*. 2023;17(3):1-27. doi:10.1145/3551390

34. Kamiran F, Calders T, Kamiran F, Nl T, Calders T, Nl T. Classification with No Discrimination by Preferential Sampling.

35. Chowdhury GG. Natural Language Processing.

36. Kumar N, Shrestha R, Li Z, Wang L. Distributionally Robust Optimization and Invariant Representation Learning for Addressing Subgroup Underrepresentation: Mechanisms and Limitations. Published online August 11, 2023. Accessed September 27, 2023. http://arxiv.org/abs/2308.06434

37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953

38. Minot JR, Cheney N, Maier M, Elbers DC, Danforth CM, Dodds PS. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining

classification performance. Published online March 9, 2021. Accessed September 29, 2023. http://arxiv.org/abs/2103.05841

39. Nilsson A, Bonander C, Strömberg U, Canivet C, Östergren PO, Björk J. Reweighting a Swedish health questionnaire survey using extensive population register and self-reported data for assessing and improving the validity of longitudinal associations. Behrens T, ed. *PLOS ONE*. 2021;16(7):e0253969. doi:10.1371/journal.pone.0253969

40. Borland D, Zhang J, Kaul S, Gotz D. Selection-Bias-Corrected Visualization via Dynamic Reweighting. *IEEE Trans Vis Comput Graph*. 2021;27(2):1481-1491. doi:10.1109/TVCG.2020.3030455

41. Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front Artif Intell*. 2021;3:561802. doi:10.3389/frai.2020.561802

42. Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. *IEEE Rev Biomed Eng*. Published online 2022:1-17. doi:10.1109/RBME.2022.3210270

43. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and Mitigating Unintended Bias in Text Classification. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM; 2018:67-73. doi:10.1145/3278721.3278729

44. Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proc AAAI Conf Artif Intell*. 2018;32(1). doi:10.1609/aaai.v32i1.11296

45. Uhde A, Schlicker N, Wallach DP, Hassenzahl M. Fairness and Decision-making in Collaborative Shift Scheduling Systems. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM; 2020:1-13. doi:10.1145/3313831.3376656

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

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

48. Lodi A, Olivier P, Pesant G, Sankaranarayanan S. Fairness over time in dynamic resource allocation with an application in healthcare. *Math Program*. Published online November 7, 2022. doi:10.1007/s10107-022-01904-6

49. Radovanović S, Delibašić B, Marković A, Suknović M. Achieving MAX-MIN Fair Cross-efficiency scores in Data Envelopment Analysis. In: ; 2022. doi:10.24251/HICSS.2022.189

50. Ponce-Ortega JM. a Universidad Michoacana de San Nicolás de Hidalgo, Chemical Engineering Department, Building V1, Ciudad Universitaria, Santiago Tapia S/N, Morelia, Michoacán, México, 58060.

51. Budhiraja I, Kumar N, Tyagi S. Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication. *IEEE Internet Things J*. 2021;8(5):3143-3156. doi:10.1109/JIOT.2020.3014926

52. Chakraborty S, Tomsett R, Raghavendra R, et al. 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)*. IEEE; 2017:1-6. doi:10.1109/UIC-ATC.2017.8397411

53. Ge Y, Zhao X, Yu L, et al. 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*. ACM; 2022:316-324. doi:10.1145/3488560.3498487

54. Keya KN, Islam R, Pan S, Stockwell I, Foulds JR. Equitable Allocation of Healthcare Resources with Fair Cox Models. Published online October 14, 2020. Accessed October 28, 2023. http://arxiv.org/abs/2010.06820

55. Achterberg T, Wunderling R. Mixed Integer Programming: Analyzing 12 Years of Progress. In: Jünger M, Reinelt G, eds. *Facets of Combinatorial Optimization*. Springer Berlin Heidelberg; 2013:449-481. doi:10.1007/978-3-642-38189-8_18

56. Fouskakis D, Draper D. Stochastic Optimization: a Review. *Int Stat Rev*. 2002;70(3):315-349. doi:10.1111/j.1751-5823.2002.tb00174.x

57. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT Press; 1998.

58. Bellman R. A Markovian Decision Process. *Indiana Univ Math J*. 1957;6(4):679-684. doi:10.1512/iumj.1957.6.56038

59. Ohno-Machado L. Modeling Medical Prognosis: Survival Analysis Techniques. *J Biomed Inform*. 2001;34(6):428-439. doi:10.1006/jbin.2002.1038

60. Demeniconi C, Davidson I, eds. *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics; 2021. doi:10.1137/1.9781611976700

61. Hooker S. Moving beyond "algorithmic bias is a data problem." *Patterns*. 2021;2(4):100241. doi:10.1016/j.patter.2021.100241

62. Field DA. Laplacian smoothing and Delaunay triangulations. *Commun Appl Numer Methods*. 1988;4(6):709-712. doi:10.1002/cnm.1630040603

63. Kim MP, Ghorbani A, Zou J. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM; 2019:247-254. doi:10.1145/3306618.3314287

64. Jo N, Tang B, Dullerud K, Aghaei S, Rice E, Vayanos P. Fairness in Contextual Resource Allocation Systems: Metrics and Incompatibility Results. *Proc AAAI Conf Artif Intell*. 2023;37(10):11837-11846. doi:10.1609/aaai.v37i10.26397

65. Garcia, G.G.P., Steimle, L.N., Marrero, W.J. and Sussman, J.B., 2024. Interpretable policies and the price of interpretability in hypertension treatment planning. Manufacturing & Service Operations Management, 26(1), pp.80-94.

66. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *eBioMedicine*. 2022;84:104250. doi:10.1016/j.ebiom.2022.104250

67. Elmachtoub AN, Grigas P. Smart "Predict, then Optimize." *Manag Sci*. 2022;68(1):9-26. doi:10.1287/mnsc.2020.3922

Tables

Table 1: Definition of fairness metrics

Metric type	Metric definition
Fairness through unawareness	Measures whether a model contains sensitive variables
Demographic parity	Measures whether the decision-making is independent of sensitive attributes in the whole population
Equal opportunity	Measures whether decision-making is independent of sensitive attributes among a qualified group
Equal odds	Measures whether decision-making is independent of

	sensitive attributes for both qualified and unqualified groups

Table 2: Pre-processing bias mitigation methods.

Method(s)	**Reference(s)**	**Area of application**	**Fairness metric**
Reweighting	Nilsson et al.[39]	Medical diagnosis	Demographic parity
	Kumar et al.[36]	Medical diagnosis	Demographic parity
Resampling	Chawla et al.[37]	Treatment design	Demographic parity
Natural language processing	Minot et al.[38]	Medical diagnosis	Equal opportunity
Fair data transformer	Biswas et al.[32]	N/A	Demographic parity

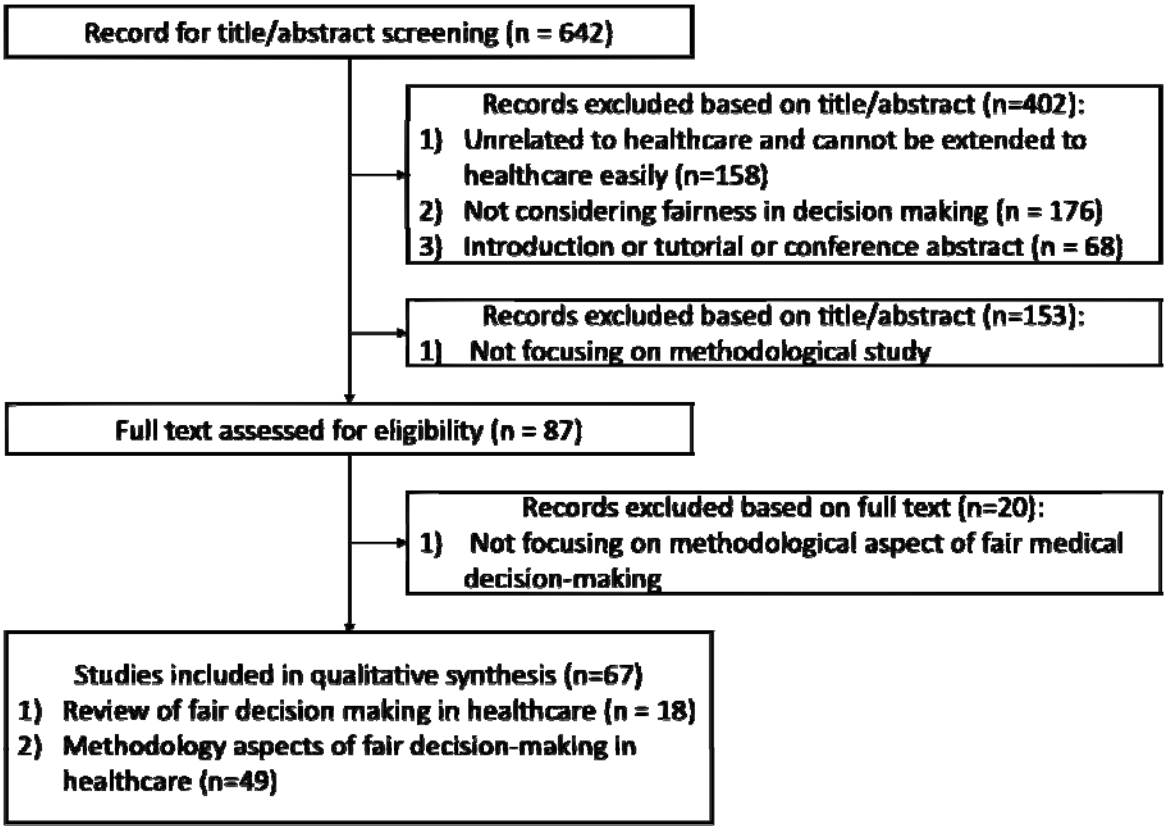
Table 3: In-processing bias mitigation methods

Optimization/Machine learning	**Method(s)**	**Reference(s)**	**Area of application**	**Fairness metric**
Optimization	Mixed-integer Programming/Stochastic Optimization	Acuna et al.[46]	Medical resource allocation	Demographic parity
		Ala et al.[47]	Appointment scheduling	Demographic parity
		Munguía-López et al.[14]	Medical resource allocation	Equal odds
		Lodi et al. [48]	Medical resource allocation	Equal opportunity
		Radovanović et al. [49]	Medical resource allocation	Demographic parity
		Ponce-Ortega [50]	Appointment scheduling	Equal opportunity
Machine learning	Reinforcement learning	Budhiraja et al.[51]	Medical scheduling	Equal opportunity
		Chakraborty et al.[52]	Clinical treatment	Equal opportunity
		Lu[10]	Clinical	Equal odds

Optimization/Machine learning	Method(s)	Reference(s)	Area of application	Fairness metric
			treatment	
		Yang et al. [12]	Clinical treatment	Demographic parity
		Tang et al.[11]	Clinical treatment	Equal opportunity
		Ge et al.[53]	N/A	Equal opportunity
		Ge et al.[13]	N/A	Demographic parity
	Deep learning	Budhiraja et al.[51]	Medical scheduling	Equal opportunity
		Lu M[10]	Clinical treatment	Demographic parity
	Fair survival analysis	Keya et al.[54]	Medical resource allocation	Demographic parity

Figures

Figure 1: Flow diagram for the systematic literature review of published fair decision making in healthcare.



Appendix

Supplemental Table 1: Bias-mitigation methods and descriptions

Stage	Method	Description
Pre-processing	Reweighting	Assign greater weights to underrepresented instances.
	Resampling	Randomly under-sample the majority groups (or over-sample the minority groups).
	Natural language processing	Remove sensitive attributes from text data before feeding data to decision making algorithms.
	Fair data transformer	T techniques used to transform input data and reduce biases.
In-processing	Mixed-integer	A type of constrained

	programming	optimization problem that allows for integer and continuous variables.
	Reinforcement learning	Algorithms that learn to make decisions by performing actions and observing the results in a given environment.
	Deep reinforcement learning	Uses multiple neural network layers and activation functions to extract new features from the input data.
	Fair survival analysis	Survival analysis adds a fairness penalty in the loss function to ensure equity.
Post-processing	Laplacian smoothing	A technique to reduce the noise of the data while preserving the important characteristics of the solution approach.
	Multi-accuracy	A technique to convert several weak learners into a strong one to achieve high accuracy rates among all subpopulation groups.