Algorithmic Fairness Beyond Legally Protected Groups and When Group Labels Are Unknown

Abdoul Jalil D. Mahamadou[1], Judy W. Gichoya[2], Artem A. Trotsyuk[1]
[1]Stanford University
[2]Emory University
abdjiber@stanford.edu, judywawira@emory.edu, atrotsyuk@stanford.edu

Abstract

The algorithmic fairness literature has focused on defining fairness group labels based on legally protected groups. This assumes that populations at risk of unfairness are known and that equity for these groups translates to broader fairness. However, these risks miss emerging or context-specific at-risk populations. We illustrate this through a review of 73 fairness in healthcare AI studies published between 2020 and 2024, as well as three case studies conducted at Stanford Health Care. The review reveals disproportionate use of protected characteristics (90%), socioeconomic factors (19%), clinical factors (14%), and system and institutional factors (5%) as group labels. Through the case studies, we show how stakeholder engagement in ethical AI assessment, primarily designed to surface value conflicts, helps identify case-specific vulnerable populations that can inform fairness interventions. This study shows the need to expand fairness group label definitions to include a broader range of context-informed attributes. Doing so can help ensure that bias mitigation strategies are better grounded in real-world social contexts, leading to more context-aware definitions of harm and equity.

Introduction

Historically, the field of algorithmic fairness has focused on protected demographic categories, such as race, ethnicity, and sex, to prevent explicit forms of algorithmic discrimination. However, relying solely on these attributes assumes that populations at risk of discrimination are known a priori and that equity for these groups translates to broader fairness, a premise increasingly challenged in real-world healthcare deployments of artificial intelligence (AI) systems. For instance, in a study examining the challenges faced by AI developers in developing more equitable AI systems, many practitioners reported difficulties in identifying which subpopulations to consider for specific machine learning applications (Holstein et al. 2019). Developers emphasized that fairness concerns often extend beyond standard demographic categories and must be defined based on the context and domain of the system. As one technical director working on general-purpose ML tools explained:

> *Most of the time, people start thinking about attributes like [ethnicity and gender...]. But the biggest problem I found is that these cohorts should be defined based on the domain and problem. For example, for [automated writing evaluation] maybe it should be defined based on [...whether the person is] a native speaker.* Holstein et al. 2019 (p. 7)

Despite ongoing efforts, teams frequently discover fairness issues only after deployment. A software engineer working on image classification reflected this uncertainty, asking:

> *How do you know the unknowns that you're being unfair towards? [...] You just have to put your model out there, and then you'll know if there's fairness issues if someone raises hell online.* Holstein et al. 2019 (p. 7)

Moreover, even when protected characteristics are used as group labels in fairness interventions, researchers have found inconsistencies in how they are constructed. For example, one study found that racial categories are often defined based on data availability (e.g., Black/White, or White/Non-white) or constrained by technical requirements, such as binary labeling (Abdu, Pasquetto, and Jacobs 2023). The authors caution that "the algorithmic fairness community is an emerging race-making institution that merits further attention" (p. 2), pointing to the need for more critical reflection on how social categories are constructed and operationalized in technical systems.

To address these issues, a growing number of proposals advocate for sociotechnical approaches that move beyond purely technical fixes to bias and fairness in AI systems. One promising direction involves building closer collaboration

between AI developers and a broader set of stakeholders, including domain experts such as ethicists, healthcare workers, and social scientists (Chen et al. 2023). These collaborations can help ensure that fairness considerations are grounded in real-world social contexts, leading to more context-aware definitions of harm and equity.

This study aims to investigate how AI practitioners define population subgroups in the context of algorithmic fairness. We reviewed 73 healthcare AI studies published between 2020 and 2024, sampled from the PubMed Central (PMC) database. The results revealed four variable categories used to define group labels: demographic variables (90%), socioeconomic variables (19%), clinical variables (14%), and system and institutional variables (5%).

We complement the literature search with three case studies at Stanford Health Care (SHC), where diverse stakeholders, including patients, clinicians, and AI developers, were engaged in the ethical evaluation of AI tools through the Fair, Useful, and Reliable AI Models (FURMs) framework (Callahan et al. 2024). The framework is designed to ensure that AI tools considered for use at SHC are fair, useful, and reliable through ethical assessments, utility simulations, cost evaluations, and monitoring of the tools. While the ethics component is designed to surface stakeholder value conflicts, we found that it can help define case-specific vulnerable populations, which can inform fairness interventions. In these case studies, we identified eleven such subgroups, primarily within the socioeconomic and clinical label categories from the literature review.

These findings indicate that although protected characteristics remain important, exclusive focus on them can overlook other relevant sources of disparity. By engaging domain experts and affected communities, developers can identify subgroups more closely aligned with real-world risks and needs. This study provides AI practitioners with evidence-based guidance to broaden their approach when defining subgroups, encouraging more detailed and context-aware fairness assessments that can improve the equity and effectiveness of AI systems in practice. Moreover, we discuss directions for more sociotechnical approaches to fairness evaluations.

In the remaining manuscript, we first outline related work on context-specific fairness and standard methods for defining subgroups in algorithmic fairness, including the creation of racial and ethnic categories, intersectional groups, proxy attributes, and country-specific categories. We then describe the methods used for our literature review and case studies, followed by a presentation and discussion of the results. Finally, we explore future directions that focus on human oversight, context-specific subgroup analysis, and outcome-based evaluations to achieve more effective AI fairness. We focus intentionally on healthcare because clinical workflow, disease phenotypes, and institutional infrastructure are common, consequential axes of disparity that differ from other domains (e.g., finance, criminal justice). Our goal is depth on healthcare-specific labels and practices rather than breadth across sectors.

Preliminaries

We identified four areas central to defining and operationalizing group labels in healthcare AI: (i) construction of racial/ethnic categories; (ii) intersectionality; (iii) fairness without explicit demographics; and (iv) country-specific categories. These domains reveal the complexity of subgroup definition and motivate our analysis, especially when labels are missing, unstable, or contested.

Construction of Racial and Ethnic Categories

Racial/ethnic classifications in fairness research are inconsistent and often oversimplified. Abdu et al. (2023) identified 14 distinct schemes in 60 FAccT papers, none of which aligned with U.S. Census standards. Datasets frequently collapse categories (Simson, Fabris, and Kern 2024), erasing smaller groups and obscuring disparities. Moreover, treating race as a fixed attribute ignores its social and political context (Benthall and Haynes 2019); yet ignoring it can mask inequality. Context-sensitive, transparent documentation of classification decisions (Mickel 2024) is essential, as small definitional shifts can double disparity measures (Jaime and Kern 2024).

Intersectional Categories

Intersectionality (Crenshaw 1989) addresses the compounded harms that result from overlapping identities. AI systems may appear fair for broad groups but still fail at intersections (e.g., gender × skin tone) (Buolamwini and Gebru 2018). Intersectional auditing (Gohar and Cheng 2023) is critical to reveal such hidden disparities.

Proxy Attributes

When protected labels are absent, incomplete, or restricted, proxy-based methods can uncover at-risk groups. Approaches include Max-Min fairness (Barsotti and Koçer 2022), latent group discovery (Chakrabarti 2023), and statistical estimation using census data (Kallus, Mao, and Zhou 2022). Adversarial techniques (Lahoti et al. 2020) can detect correlations between hidden group membership and outcomes. Complementing these exemplars, we now summarize survey-level approaches to fairness without demographics, noting where healthcare adoption remains limited. We add this to clarify how our stakeholder-led subgrouping is a complementary practice-first approach when group labels are unknown or unavailable (Ashurst and Weller 2023).

Country-specific Group Labels

Protected classes vary across legal, cultural, and social contexts. Attributes such as caste in South Asia (Franz 2023) or language and rural-urban divide in Africa (Asiedu et al. 2023) may be critical locally but absent from U.S. frameworks. Fairness perceptions also differ cross-nationally; Sasaki et al. (2024) found significant metric preference variation among China, France, Japan, and the U.S.

Related Work on Context-specific Fairness

While extensive literature addresses fairness metrics and mitigation strategies, there is a limited amount of work that explicitly examines how practitioners define vulnerable populations in real-world deployments. Holstein et al. (2018) documented practitioners' struggles with identifying relevant subpopulations but did not systematically analyze which groups are actually considered. Recent work on fairness without demographics (Dwork et al. 2023) provides technical solutions but does not address the upstream question of which groups matter in specific contexts. Our work fills this gap by empirically documenting current practices and demonstrating how stakeholder engagement can identify context-specific vulnerabilities that technical approaches alone might miss.

Methods

This section describes the literature search method, summarized in **Figure 1**, and the case studies conducted at SHC.

Literature Search

We searched research papers published between 2020 and 2024, indexed by PubMed Central (PMC), to investigate how AI practitioners define population subgroups for identifying and mitigating algorithmic bias. PMC is a free digital archive of full-text biomedical and life sciences journal articles maintained by the US National Library of Medicine at the National Institutes of Health. We chose PMC due to its comprehensive coverage of peer-reviewed biomedical literature and its open-access policy, making it a suitable source for identifying trends in how population subgroups are defined in AI fairness research. Additionally, PMC provides full-text access, which enables a more in-depth examination of how subgroup variables are discussed and operationalized across studies. While this study is limited to a single database and may not capture subgroup definitions used in other domains or publication venues, it offers a focused overview of how population subgroups are currently identified within healthcare AI research, highlighting emerging patterns and limitations in subgroup conceptualization.

We used a search query that combines terms related to AI, bias mitigation, and healthcare as follows:

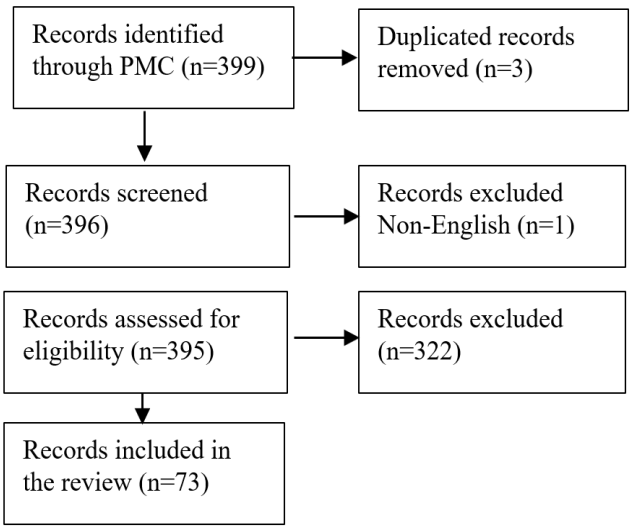

Figure 1: Flow diagram of the PMC search.

("artificial intelligence"[MeSH] OR "machine learning"[MeSH] OR "deep learning"[All Fields] OR "AI"[All Fields] OR "ML"[All Fields] OR "predictive analytics"[All Fields]) AND "bias mitigation"[All Fields] AND ("medicine"[MeSH] OR "medical"[All Fields] OR "clinical"[All Fields] OR "health informatics"[All Fields] OR "healthcare algorithms"[All Fields] OR "diagnostic"[All Fields] OR "electronic health records"[All Fields] OR "patient-centered"[All Fields]) AND ("2020/01/01"[PubDate] : "2024/12/31"[PubDate])

The search performed on March 14th, 2025, yielded 399 records, three of which were duplicates, and one record was in a language other than English, which was excluded from the analysis. The records were first screened for relevance from titles and abstracts (396 records). We excluded records that do not assess group fairness (322 records). This includes, for instance, review papers, papers focusing on individual fairness, fairness perceptions, and other forms of bias (e.g., AI adoption bias). For each paper in the final sample (73 records), we assessed how subgroups were labeled. If a paper reported on multiple labels (e.g., race and sex), each attribute was counted separately in our analysis.

The FURM Ethical Assessment Framework

The Fair, Useful, and Reliable AI Models (FURMs) is an AI evaluation framework introduced by (Callahan et al. 2024) to assess the ethics, usefulness, and reliability of AI systems considered for use at SHC. The framework involves ethical assessments, utility simulations, cost evaluations, and monitoring of the tools. The ethics component is primarily designed to identify potential value conflicts among stakeholders with key roles in developing the tools, including internal and external developers, as well as those who use or could

be impacted by the tools, such as patients, clinicians, and other healthcare personnel. The process assesses stakeholders' perceptions of AI risks across various ethical dimensions, including privacy and confidentiality, bias and fairness, model training and performance, and informed consent. Through semi-structured interviews, stakeholders are prompted with questions, such as "Are there any groups of patients or clinicians who might be more likely to be harmed than others?" and "Do you have any concerns about how this model was trained?" to explicitly surface concerns related to bias and fairness, and unintended harm of the tools.

In this study, we present three predictive AI use cases of the ethical assessments. The first tool, referred to here as **LabAlert**, is an AI-based tool designed to reduce low-value laboratory tests. This was motivated by unnecessary laboratory tests ordered by doctors, which could increase the cost of testing and lead to a shortage of supplies, such as blood draw tubes. Given a patient, the tool predicts the likelihood that the following laboratory test will be stable based on previous laboratory results, vital signs, medications, and demographics. When the tool identifies patients for whom a test is likely to be of low value, it triggers an alert in the patient's electronic health record for the doctors. The second tool referred to here, **HeartRead**, aims to improve the diagnosis of hypertrophic cardiomyopathy (HCM). HCM is a rare heart condition characterized by the thickening of the heart muscle, which makes it more difficult for the heart to pump blood. It affects about 1 in 500 people (Maron et al. 2022). HeartRead uses electrocardiogram (ECG) data to identify patients who may benefit from an echocardiogram, determining whether a patient has HCM. The model was trained on ECG data from four academic medical centers in the US and Asia. The third tool, **SendOff**, aims to reduce hospital readmissions by predicting patients at high risk of readmission within 30 days of discharge. The model was trained on patient electronic health record data, including demographic and clinical information. Sendoff helps prioritize patients who are most likely to benefit from SHC's Aging Adult Services Transition of Care program, designed to ensure a smooth transition from hospital to home or another healthcare facility, particularly for patients who are at risk of experiencing complications after discharge, such as those with complex medical conditions.

We conducted 28 semi-structured stakeholder interviews between August 2024 and January 2025, comprising both online and in-person sessions, with 20 individual interviews and two patient focus groups. Patients (coded as P) were recruited through SHC's Patient and Family Partnership Program. In contrast, clinicians (coded as C) and developers (coded as D) were recruited as prospective users of the tools and as contributors to their development. Specifically, for LabAlert, we interviewed three patients, two prospective physician users of the tool, and three physician-informaticists with key roles in developing the tool. For HeartRead, we interviewed four prospective clinical users of the tool, three members of the design team, and a focus group comprising four patients. For SendOff, we interviewed two members of the SHC team proposing the use case, one representative of the vendor, two physician prospective clinical users, and a focus group comprising four patients. Each interview lasted between 30 minutes and an hour, while the focus groups were conducted for one hour each. We selected the FURM framework for several reasons: (1) it explicitly incorporates stakeholder perspectives through structured ethical assessments, enabling discovery of context-specific vulnerabilities; (2) it has been validated in real-world healthcare settings; and (3) its semi-structured interview protocol specifically probes for differential impacts across patient groups. Alternative frameworks, such as traditional Health Technology Assessment (HTA) (WHO, n.d.-a), typically focus on aggregate outcomes rather than subgroup disparities, whereas pure technical auditing frameworks lack mechanisms for stakeholder input.

The three AI tools were selected to represent diverse clinical use cases: LabAlert addresses resource utilization (laboratory testing), HeartRead targets the detection of rare diseases, and SendOff focuses on care transitions (readmission prevention). This diversity ensures that our findings are not specific to a single clinical context, although we acknowledge that all cases come from a single institution, which may limit their generalizability.

Results

From the PMC literature search, we identified four variable categories summarized in **Table 1** to define group labels: demographic, socioeconomic, and system and institutional variables. Due to the citation format and space limitations, we report the same table in the **Appendix**[2] with the references for each variable. From the AI use cases at SHC, we identified 11 groups at risk of unfairness: six from **LabAlert**, one from **HeartRead**, and four from **SendOff**.

Group Label Definitions from The Literature Search

Demographic Variables

AI practitioners defining group labels from demographic variables was the most commonly used approach in the final 73 samples. It accounts for 66 (90%) of the sample. Sex, race, and ethnicity were the most frequently used variables, followed by age, accounting for 49 (67%), 47 (64%), and 26

[2] https://github.com/abdjiber/AIES2025-APPENDIX-FAIRNESS

(36%) cases, respectively. Region and Nationality, marital status, and handedness were less common, representing 7 (10%),

Group Label Categories	**Total**
Demographic Variables	
Age	26/73 (36%)
Sex	49/73 (67%)
Race/Ethnicity	47/73 (64%)
Region/Nationality	7/73 (10%)
Marital Status	2/73 (3%)
Handedness	1/73 (1%)
Grand Total	**66/73 (90%)**
Socioeconomic Variables	
Income	5/73 (7%)
Education	5/73 (7%)
Employment Status	1/73 (1%)
Working Position	1/73 (1%)
Level of Work	1/73 (1%)
Working Conditions (e.g., 35h/week)	1/73 (1%)
Insurance	6/73 (8%)
Self-reported Health Status	1/73 (1%)
Smoking	1/73 (1%)
Grand Total	**14/73 (19%)**
Clinical Variables	
Disease Status (Parkinson's Disease, Human Papilloma-virus, Unilateral Vocal Fold Paralysis)	3/73 (4%)
Comorbidities (Cardiovascular Disease, Diabetes)	1/73 (1%)
Cancer Types (Ovarian, Bowel, Lung, Breast)	1/73 (1%)
Parkinson's Disease	1/73 (1%)
Bowel Obstruction	1/73 (1%)
Tumor Size Status (Head and Neck tumors)	1/73 (1%)
Lymph Node Size (Head and Neck tumors)	1/73 (1%)
Chemotherapy (Head and Neck tumors)	1/73 (1%)
BMI	1/73 (1%)
Radiological grade	1/73 (1%)
Oxford Knee Scores	1/73 (1%)
Disease Region and Intensity (e.g., brain spatial location)	1/73 (1%)
Grand Total	**10/73 (14%)**
System/Institutional Variables	
Scanner Type (e.g., T1-weighted, Harmonized T1-weighted, log-Jacobians)	1/73 (1%)
Scanner Vendor (e.g., General Electric, Philips, Siemens)	1/73 (1%)
Scanner Field (e.g., 1.5T, 3T)	1/73 (1%)
Hospital Location (e.g., high-income and low-income countries)	2/73 (3%)
Grand Total	**4/73 (5%)**

Table 1*:* Group label definitions from the PMC literature search. Counts are per paper; percentages are of 73 papers; totals within a category need not sum to 100%.

2 (3%), and 1 (1%) of the sample. Of these variables, only handedness is not protected by law, making 65 (89%) of the group labels defined by protected classes.

Socioeconomic Variables

Socioeconomic variables were the second largest category of group labels, accounting for 14 (19%) of the sample. Specifically, insurance types such as Medicare, Medicaid, commercial, government, and self-pay (Bing et al. 2022) were the most commonly used socioeconomic variables, accounting for 6 (8%) of the sample, followed by education and income, each at 5 (7%). Besides these variables, Buda et al. used employment status, as well as the level and position of work of participants, to assess fairness (Buda et al. 2022). In contrast, Yaseliani et al. used working conditions, such as 35 hours per week (Yaseliani, Noor-E-Alam, and Hasan 2024), and Libin et al. used self-reported health status ranging from excellent to poor (Libin et al. 2024).

Clinical Variables

Ten out of the 73 studies (14%) used clinical variables to define the group labels, among which the most common were disease status 3 (4%), such as Parkinson's Disease, Human Papillomavirus (Salahuddin et al. 2023), and Unilateral Vocal Fold Paralysis (Zhang et al. 2023). Other studies used comorbidities such as cardiovascular disease and diabetes (Dang et al. 2024), cancer types, including Ovarian, Bowel, Lung, Breast (Watson et al. 2024), and Parkinson's Disease (Patil and Ford 2024), bowel obstruction (Juwara, El-Hussuna, and El Emam 2024), head and neck tumor size, lymph node size, and chemotherapy to define the attributes (Salahuddin et al. 2023). In addition, Heidari et al. used BMI, radiological grade, and the Oxford Knee Scores to assess subpopulation performance (Heidari et al. 2024).

System and Institutional Variables

System and institutional variables were also used to define group labels. These include scanner characteristics such as the type (e.g., T1-weighted), vendor (e.g., General Electric scan), magnetic field (e.g., 1.5T), and the hospital (e.g., in high-income and low-income countries). These variables represent 4 (5%) of the sample.

Group Labels Identified with The FURM Ethical Assessment

We identified 11 group labels, summarized in **Table 2**, from the LabAlert, HeartRead, and SendOff ethical evaluations. These include patients on anticoagulants, thrombocytopenic patients, patients undergoing chemotherapy, patients on dialysis, disease conditions (e.g., cancer, CVD, sepsis, mental health, diabetes, immunocompromised and frail patients), young Black-American athletes, patients with limited English proficiency, low socioeconomic status (e.g., unhoused, unstable housing, uninsured, financial difficulties), and hospital vs non-hospital patients.

Case Study: LabAlert

For LabAlert, a clinician expressed concerns about patients on anticoagulants, those with thrombocytopenia, and those undergoing chemotherapy in the following:

*"I think that, you know, patients where they're high risk for bleeding, you know, our GI [gastrointestinal] patients specifically, patients who have been on **anticoagulants** are always a big concern, or if they're, you know, if they're very **thrombocytopenic**, if they're actively **undergoing chemo** [chemotherapy], those are all kind of things we look out for, and that's kind of why that makes…it would make me a little bit nervous about this kind of tool because sometimes, you know, if a patient's hemoglobin, for example, has been stable but let's say, you know, they recently had a GI bleeding and we're reintroducing, you know, their anticoagulation back in, you know, we're still going to want several days more of monitoring to ensure that, you know, they're stable after, you know, restarting their anticoagulation. So it's things like that that AI, I don't know, can necessarily account for, at least in this early stage."* (C1)

Another clinician expressed concerns for patients on dialysis due to potential fluctuations in the laboratory tests that can induce errors in the tool, but noted no problems related to racial groups.

"*I guess it might be hard for people who are like on **dialysis** for this tool to be…it might get kind of like confused since there's going to be a lot more fluctuation in the labs, but I don't know that it…I'm trying to think of like how would it affect the patients if…because it would…probably wouldn't get the alert but maybe we'd still be drawing unnecessary blood from those patients? In terms of racial minorities or other groups, nothing comes to mind on how there would be like inequity in the tool. Yeah, can't think of anything.*" (C2)

A tool developer noted that patients with more extended hospital stays will likely benefit most from it.

*"I think there are…like are there certain patients who may be more subjected to unnecessary labs right now that this [tool] would identify? I think the patients who tend to have longer durations of stay, **lengths of stay**. Like I said, there's just that inertia and you kind forget what [laboratory] orders are active, and it just keeps going. So I would say like probably the longer you're in the hospital the less beneficial labs will be. Therefore the more likely this [tool] will help if clinicians are not, you know, being very systematic about revisiting those orders daily.*" (D1)

Similarly, one developer highlighted the importance of accounting for clinical subgroups based on underlying

health conditions, such as individuals undergoing cancer treatment, those with cardiovascular disease, and patients experiencing systemic infections.

*"There would certainly be a difference in terms of like clinical subgroup**, the onc [oncology] patient versus the cardiac patient versus the sepsis patient**, I can imagine those would be different scenarios, but it's just clinically different.*" (D2).

Case Study: HeartRead

A developer of HeartRead highlighted the need for greater diversity in the development of medical tools. For instance, specific ECG changes, such as T wave inversion, may be misinterpreted in young African-American athletes, potentially leading to misdiagnosis compared to individuals of Northern European ancestry (Davis et al. 2022).

"*I think this tool, like all the tools unfortunately to date, are limited by the populations in which they're developed, and those populations are definitely not diverse enough…You know, there are certain things that are just very clear across racial groups from our own experience, changes on an ECG that are true, but equally there are changes that everybody learns even for their Board exams, for example, about T wave inversion in African-American individuals, especially young athletes where you really need a new, you could see T wave inversion and QRS complexes that are very high and very large across the entire precordial leads and, you know, that could…you know, that was enough. Someone with Northern European ancestry you would think this has to be hypertrophic cardiomyopathy, but in let's say a **young African-American football player** could be completely normal.*" (D3)

Case Study: SendOff

A clinician and patient focus group identified patients with limited English proficiency at risk of SendOff. In the following, a patient noted:

"There are patients who have language barriers, which might be a problem… Right, and there may not be clarity for the patient to understand…Or they might have trouble, you know, figuring out what medications they have to pick up at the pharmacy or, you know, because it's not in their native language" (P1)

Moreover, all stakeholders noted that patients with low socioeconomic status could be vulnerable to SendOff. These include unhoused patients and those with unstable housing, uninsured patients, and those with financial difficulties. A patient expressed concerns that the tool might underperform for non-Stanford patients due to the underrepresentation of such patients in the training data used to develop the tool.

Fairness/Risk Factors	**Tools**
Patients on Anticoagulants	LabAlert
Thrombocytopenic Patients	LabAlert
Patients Undergoing Chemotherapy	LabAlert
Patient on Dialysis	LabAlert
Hospital Length of Stay	LabAlert
Disease Conditions (e.g., cancer, CVD, sepsis)	LabAlert
Young Black Athletes	HeartRead
Limited English Proficiency Patients	SendOff
Patients with Low Socioeconomic Status (e.g., unhoused, unstable housing, uninsured, financial difficulties)	SendOff
Stanford vs non-Stanford patients in the training data	SendOff
Disease Conditions (e.g., mental health, diabetes, immunocompromised and frail patients)	SendOff

Table 2: Group labels identified from the FURM ethical assessment.

These concerns are complemented with those of a developer who noted that previous versions of the tool underperformed for patients with mental health issues, or diabetes, immunocompromised, and frail. As the tool was trained on data before the COVID-19 pandemic, one patient expressed concerns about the skewness of the data distribution.

Discussion

This section examines the different influences on how group labels are defined in studies from the literature search. Legal standards have played a crucial role in shaping current fairness metrics; however, they have limitations when applied to healthcare settings. Furthermore, focusing solely on legally protected groups does not always guarantee fair treatment across all relevant populations. The following subsections review how legal definitions, data availability, technical factors, and stakeholder input have guided the choice of group labels. Of the stakeholder-surfaced subgroups (**Table 2**), LEP and low-SES are socioeconomic; dialysis, chemotherapy, anticoagulation, thrombocytopenia, frailty, mental health, and diabetes are clinical; 'Stanford vs non-Stanford patients' is system/institutional; and 'young Black athletes' combines demographic with clinical context (athlete ECG patterns). Notably, LEP did not appear as a subgroup variable in our 73-paper sample, highlighting a salient, practice-critical blind spot of protected-class-only audits.

Legal Discrimination Influence on Group Label Definitions

Legal definitions of discrimination have inspired the development of fairness metrics. For instance, disparate treatment refers to situations where individuals are treated differently, explicitly based on a protected attribute, such as race or gender, regardless of the intent behind the treatment. In contrast, disparate impact occurs when a seemingly neutral policy or model disproportionately harms a protected group, even without explicit intent to discriminate (Barocas and Selbst 2016). These legal concepts have influenced the conceptualization of fairness in AI, often leading researchers to frame fairness metrics in terms of group-based outcomes and the prevention of unequal treatment across protected categories (Verma and Rubin 2018; Gajane and Pechenizkiy 2017). Another line of work under the disparate impact prism is the 80% rule, a guideline established by the Equal Employment Opportunity Commission Code of Federal Regulations. According to this rule, a selection rate for any protected group less than 80% of the rate for the group with the highest selection rate is considered evidence of adverse impact. This threshold has been widely adopted as a baseline for detecting potential bias in algorithmic systems, for instance, (Foulds et al. 2020) used this rule to define an intersectional fairness metric. Although these legal frameworks provide foundations for fairness, their application to healthcare AI reveals limitations. These frameworks were established before the emergence of modern AI, failing to capture the complex, intersectional interactions among clinical, socioeconomic, and institutional factors that affect algorithmic performance.

Fairness to Protected Groups Does Not Imply Fairness to All Groups

A trivial assumption is that fairness to protected groups does not imply fairness to all groups. This assumption is validated in (Brown-Mulry et al. 2025). The authors evaluated subgroup performances of an FDA-cleared breast cancer screening AI system. While the model was fair across demographic groups (race, ethnicity, and age), it underperformed for imaging and pathological subgroups. If the subgroup analysis were limited to demographic groups, the model could have been considered fair.

Data Availability Influence on Group Label Definitions

While this study did not examine the justification of group label definitions, a similar study (Abdu, Pasquetto, and Jacobs 2023) found that the availability of data often justified the use of racial categories in fairness interventions. These findings could be extended to our research. Collecting new data, such as demographic information, for algorithmic fairness purposes can be time-consuming and expensive, even if permitted by laws and regulations. Thus, researchers can rely on available public or private data for fairness audits. Popular public datasets used in fairness evaluation, such as the Adult, German Credit Approval, and COMPAS datasets, primarily contain protected attributes that researchers utilize to evaluate subgroup performance. It has been demonstrated that even these datasets are noisy and contain outdated group labels, such as racial categories (93), prompting researchers to advocate for the development of newer and more up-to-date datasets (Mahamadou and Trotsyuk 2025), which may raise ethical concerns (Andrus and Villeneuve 2022).

Disproportionate Use of Protected Groups

The results show that researchers place disproportionate emphasis on certain protected classes, particularly race, ethnicity, and sex, while others receive little to no attention. Notably, none of the studies included disability as a group label despite its recognition as a protected class under multiple legal frameworks. This imbalance could also be attributed to the limited availability of data.

Socioeconomic Determinants of Health Disparities

Socioeconomic factors, including income, education, employment status, occupation, work level, working conditions, insurance coverage, health status, and smoking habits, are well-established determinants of health equity. Numerous studies have demonstrated that these factors profoundly influence individuals' access to resources, exposure to risks, and overall well-being, leading to significant disparities in health outcomes across populations (Chelak and Chakole 2023; WHO, n.d.-b; Zajacova and Lawrence 2018). For instance, income and education are closely linked to life expectancy and morbidity rates, with lower socioeconomic status consistently associated with poorer health outcomes and shorter lifespans (Zajacova and Lawrence 2018; Finkelstein et al. 2022). Employment conditions, occupational status, and job-related exposures also contribute to health disparities through access to material resources and psychosocial stressors (U.S. National Cancer Institute 2017). Smoking and health insurance coverage are also crucial, as they significantly impact disease risk and access to care. Incorporating these socioeconomic indicators enables researchers and policymakers to more effectively capture the complex social and economic contexts that drive health inequities, contexts that increasingly inform the definition of fairness group labels in algorithmic systems.

Infrastructure-Driven Sources of Unfairness and Dataset Shift

Technical factors, such as scanner type, vendor-specific hardware, and imaging protocols, can introduce biases in

medical AI systems (102, 103), resulting in unfair performance disparities across patient populations. For instance, changes in hardware from different vendors can result in covariate shift, i.e., changes in the AI input data, which can lead to poor model performance (Tejani et al. 2024; Schrouff et al. 2024; Barrainkua et al. 2023). Geographic bias in training data further exacerbates these disparities. A 2020 study analyzing U.S. clinical machine learning literature found that algorithms were trained on cohorts from three states: California, Massachusetts, and New York (Kaushal, Altman, and Langlotz 2020). This underrepresentation creates systemic fairness risks, as models may fail to generalize to populations with differing demographic, environmental, or socioeconomic profiles (e.g., rural areas or regions with distinct disease prevalence patterns).

Stakeholder-Informed Group Labels

The case studies demonstrate promising directions for AI participation in algorithmic fairness, particularly, how stakeholders can contribute to context-specific group label definitions. While the identified groups primarily fall under the categories of socioeconomic and clinical variables from the literature search, each tool is associated with specific vulnerable populations. Specifically, LabAlert may underperform for patients with short stays or those underrepresented in the training data, resulting in less accurate or suboptimal care. Additionally, if longer stays are disproportionately associated with specific demographic groups, the system may inadvertently perpetuate existing healthcare disparities in delivery. Group labels, such as Young Black American Athletes, highlight the importance of considering intersectional categories, where overlapping social and contextual identities, such as race, age, and occupation, can compound vulnerability to algorithmic bias. The use of SendOff could increase the risk of hospital readmission for patients with limited English proficiency. Many discharge instructions and medication labels are written in English, and patients who struggle with comprehension may unintentionally misuse medications or fail to follow care instructions correctly. As a result, some patients are readmitted not due to clinical deterioration, but because of language barriers, an unfairness that would remain invisible to the model without stakeholder input.

Insights from Case Studies

Our case studies reveal several patterns not apparent from the literature review alone. First, stakeholder-identified vulnerabilities often cut across traditional categories - for example, 'young Black athletes' represents an intersection of age, race, and occupation specific to ECG interpretation. Second, system-level factors, such as language barriers, emerged prominently in practice but rarely appear in published fairness evaluations. Third, clinical stakeholders consistently identified medication-specific vulnerabilities (anticoagulants, chemotherapy) that require domain expertise to recognize. These findings suggest that purely data-driven approaches to fairness, even with perfect demographic information, would miss essential sources of disparity. Further, findings from a single health system risk institutional bias. We mitigated this by engaging diverse roles (patients, clinicians, developers) across three distinct tools and by grounding subgroup proposals in concrete failure modes. Replication at additional sites is a priority future step.

Recommendations

This section presents strategies to strengthen fairness evaluations in healthcare AI. Current practices often rely on narrow definitions centered on legally protected demographics, overlooking disparities tied to clinical or institutional factors. We recommend expanding fairness definitions and adopting evaluation methods grounded in healthcare practice, including subgroup analysis beyond standard metrics, linking outputs to clinical relevance, promoting transparency, and supporting replicability through diverse datasets.

Adopting a Broader Group Fairness Definition

Limiting group fairness to protected attributes ignores other sources of disparity. Our findings indicate that unfairness can also arise from clinical, systemic, or institutional characteristics. We urge expanding group labels to capture these dimensions, enabling detection of bias in contexts not covered by traditional protected classes.

Towards Sociotechnical Fairness Evaluations

A sociotechnical approach to fairness acknowledges that neither single metrics nor universal definitions can fully capture the spectrum of model bias in clinical contexts. Because regulatory review cannot anticipate all failures, ongoing real-world surveillance and context-specific evaluation are essential for uncovering hidden risks. The strategies below emphasize methods for detecting, understanding, and addressing unfairness in populations that are most likely to be overlooked.

A. **Invest upstream in data collection and documentation**. Several context-specific subgroups (e.g., LEP, unstable housing) are not reliably captured in routine data. We recommend adding standardized fields and provenance (e.g., interpreter use, preferred language for discharge instructions, housing stability, payer granularity) to enable equity-relevant subgroup audits.

B. **Maintain the human in the loop in AI deployment.** Human oversight allows clinicians to detect atypical errors caused by shortcut learning tied to patient, hospital,

or disease features (Banerjee et al. 2023). Trained users can recognize when outputs may disproportionately harm certain groups and intervene before adverse impacts occur.

C. **Move from broad metrics to clinical subgroup evaluation.** Assessing performance within clinically defined subgroups, beyond just race, sex, or age, can reveal disparities masked in aggregate metrics. For example, addressing poor performance in dense-breast patients reduced disparities for Black patients in breast cancer prediction (Brown-Mulry et al. 2025; Woo et al. 2025).

D. **Link model performance to clinical outcomes.** Connecting predictions to actionable outcomes (e.g., pairing cardiomegaly detection with ejection fraction) ensures fairness assessments reflect meaningful differences in care. This approach highlights when disparities truly affect clinical decisions versus when they are statistically detectable but clinically irrelevant.

E. **Incentivize rather than penalize bias detection.** Positive incentives, such as bias "bounty" programs (Malladi and Subramanian 2020), encourage deeper post-deployment audits and transparency in reporting model failures, increasing the likelihood of discovering risks to marginalized populations.

F. **Use benchmarks cautiously.** Benchmark scores can hide subgroup harms. Supplementing them with failure audits, subgroup analysis, and outcome linkage prevents overreliance on one-time performance claims and reveals high-risk groups overlooked in headline metrics.

Addressing Data Collection Challenges

We acknowledge that implementing broader definitions of fairness presents practical challenges, particularly in terms of data availability. Even basic demographic data are often missing or unreliable, and collecting additional variables raises concerns about cost, privacy, and regulation. To address these challenges, we propose several strategies. First, prioritized data collection should leverage stakeholder input to identify the most critical additional variables for specific use cases, rather than attempting comprehensive data collection across all possible dimensions. Second, privacy-preserving approaches, such as federated learning and differential privacy, can enable fairness assessments without centralizing sensitive data, thereby addressing both regulatory compliance and patient privacy concerns. Third, standardization efforts are necessary to develop standardized data models for fairness-relevant variables beyond demographics, thereby facilitating consistent evaluation across institutions and studies. Finally, regulatory alignment with federal agencies is crucial for incorporating broader fairness considerations into approval processes, thereby creating incentives for developers to collect and evaluate relevant subgroup data from the earliest stages of development. These strategies can address the upstream data challenges that currently limit the implementation of context-aware fairness assessments in healthcare AI systems.

Conclusion

We investigate how subgroups are defined in the context of algorithmic fairness through a literature review. We show that current fairness assessments overwhelmingly rely on demographic categories, which may obscure other essential sources of bias rooted in clinical, social, and systemic factors. While the study is limited by its focus on the healthcare domain and a modest sample size, it offers a valuable overview of how group labels have been defined in the literature. We present three case studies from SHC, where stakeholder engagement revealed group labels that would likely be missed through conventional approaches. Our findings advocate for shifting from traditional fairness group labels, which are primarily centered on demographic categories, toward comprehensive, stakeholder-informed sociotechnical frameworks that can capture the complex determinants of health outcomes. Future research should prioritize the development of explicit methodologies and guidelines for engaging diverse stakeholder perspectives throughout AI design and evaluation processes.

Acknowledgements

This work is supported by the Stanford-GSK.ai program (AJDM and AAT). We are grateful to the reviewers for their valuable comments and feedback.

References

Abdu, Amina A., Irene V. Pasquetto, and Abigail Z. Jacobs. 2023. "An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '23, June 12, 2023, 1324–33. https://doi.org/10.1145/3593013.3594083.

Andrus, McKane, and Sarah Villeneuve. 2022. "Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '22, June 20, 2022, 1709–21. https://doi.org/10.1145/3531146.3533226.

Ashurst, Carolyn, and Adrian Weller. 2023. "Fairness Without Demographic Data: A Survey of Approaches."

Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (New York, NY, USA), EAAMO '23, October 30, 2023, 1–12. https://doi.org/10.1145/3617694.3623234.

Asiedu, Mercy Nyamewaa, Awa Dieng, Abigail Oppong, Maria Nagawa, Sanmi Koyejo, and Katherine Heller. 2023. "Globalizing Fairness Attributes in Machine Learning: A Case Study on Health in Africa." arXiv:2304.02190. Preprint, arXiv, April 4. https://doi.org/10.48550/arXiv.2304.02190.

Banerjee, Imon, Kamanasish Bhattacharjee, John L. Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N. Patel, Rakesh Shiradkar, and Judy Gichoya. 2023. "'Shortcuts' Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation." *Journal of the American College of Radiology : JACR* (United States) 20 (9): 842–51. https://doi.org/10.1016/j.jacr.2023.06.025.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." SSRN Scholarly Paper 2477899. Rochester, NY. https://doi.org/10.2139/ssrn.2477899.

Barrainkua, Ainhize, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. 2023. "Preserving the Fairness Guarantees of Classifiers in Changing Environments: A Survey." *ACM Computing Surveys*, ahead of print, December 15, 2023. https://doi.org/10.1145/3637438.

Barsotti, Flavia, and Rüya Gökhan Koçer. 2022. "MinMax Fairness: From Rawlsian Theory of Justice to Solution for Algorithmic Bias." *AI & SOCIETY*, ahead of print, November 30, 2022. https://doi.org/10.1007/s00146-022-01577-x.

Benthall, Sebastian, and Bruce D. Haynes. 2019. "Racial Categories in Machine Learning." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAT* '19, January 29, 2019, 289–98. https://doi.org/10.1145/3287560.3287575.

Bing, Simon, Andrea Dittadi, Stefan Bauer, and Patrick Schwab. 2022. "Conditional Generation of Medical Time Series for Extrapolation to Underrepresented Populations." *PLOS Digital Health* 1 (7): e0000074. https://doi.org/10.1371/journal.pdig.0000074.

Brown-Mulry, Beatrice, Rohan Satya Isaac, Sang Hyup Lee, Ambika Seth, KyungJee Min, Theo Dapamede, Frank Li, et al. 2025. "Subgroup Performance of a Commercial Digital Breast Tomosynthesis Model for Breast Cancer Detection." arXiv:2503.13581. Preprint, arXiv, March 17. https://doi.org/10.48550/arXiv.2503.13581.

Buda, Teodora Sandra, João Guerreiro, Jesus Omana Iglesias, Carlos Castillo, Oliver Smith, and Aleksandar Matic. 2022. "Foundations for Fairness in Digital Health Apps." *Frontiers in Digital Health* (Switzerland) 4: 943514. https://doi.org/10.3389/fdgth.2022.943514.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, January 21, 2018, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Callahan, Alison, Duncan McElfresh, Juan M. Banda, Gabrielle Bunney, Danton Char, Jonathan Chen, Conor K. Corbin, et al. 2024. "Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems." *NEJM Catalyst* 5 (10): CAT.24.0131. https://doi.org/10.1056/CAT.24.0131.

Chakrabarti, Deepayan. 2023. "SURE: Robust, Explainable, and Fair Classification without Sensitive Attributes." *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 6, 2023, 179–89. https://doi.org/10.1145/3580305.3599514.

Chelak, Khushbu, and Swarupa Chakole. 2023. "The Role of Social Determinants of Health in Promoting Health Equality: A Narrative Review." *Cureus* 15 (1): e33425. https://doi.org/10.7759/cureus.33425.

Chen, You, Ellen Wright Clayton, Laurie Lovett Novak, Shilo Anders, and Bradley Malin. 2023. "Human-Centered Design to Address Biases in Artificial Intelligence." *Journal of Medical Internet Research* 25 (March): e43251. https://doi.org/10.2196/43251.

Crenshaw, Kimberle. 1989. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum* 1989 (1). https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

Dang, Vien Ngoc, Anna Cascarano, Rosa H. Mulder, Charlotte Cecil, Maria A. Zuluaga, Jerónimo Hernández-González, and Karim Lekadir. 2024. "Fairness and Bias Correction in Machine Learning for Depression Prediction across Four Study Populations." *Scientific Reports* 14 (April): 7848. https://doi.org/10.1038/s41598-024-58427-7.

Davis, Angus J., Christopher Semsarian, John W. Orchard, Andre La Gerche, and Jessica J. Orchard. 2022. "The Impact of Ethnicity on Athlete ECG Interpretation: A Systematic Review." *Journal of Cardiovascular Development and Disease* 9 (6): 183. https://doi.org/10.3390/jcdd9060183.

Dwork, Cynthia, Daniel Lee, Huijia Lin, and Pranay Tankala. 2023. "From Pseudorandomness to Multi-Group Fairness and Back." *Proceedings of Thirty Sixth Conference on Learning Theory*, July 12, 2023, 3566–614. https://proceedings.mlr.press/v195/dwork23a.html.

Finkelstein, Daniel M., Jessica F. Harding, Diane Paulsell, Brittany English, Gina R. Hijjawi, and Jennifer Ng'andu. 2022. "Economic Well-Being And Health: The Role Of Income Support Programs In Promoting Health And Advancing Health Equity." *Health Affairs* 41 (12): 1700–1706. https://doi.org/10.1377/hlthaff.2022.00846.

Foulds, James R., Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. "An Intersectional Definition of Fairness." *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, April 2020, 1918–21. https://doi.org/10.1109/ICDE48307.2020.00203.

Franz, Jesse. 2023. "Councilmember Sawant and South Asian Community Leaders Introduce First-in-the-Nation Legislation To Ban Caste Discrimination." *Seattle City Council Blog*, January 25, 2023. https://council.seattle.gov/2023/01/24/councilmember-sawant-and-south-asian-community-leaders-introduce-first-in-the-nation-legislation-to-ban-caste-discrimination/.

Gajane, Pratik, and Mykola Pechenizkiy. 2017. "On Formalizing Fairness in Prediction with Machine Learning." *arXiv Preprint arXiv:1710.03184*, 2017.

Gohar, Usman, and Lu Cheng. 2023. "A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges." *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, August 2023, 6619–27. https://doi.org/10.24963/ijcai.2023/742.

Heidari, Nima, Stefano Olgiati, Davide Meloni, James Parkin, Brady Fish, Mark Slevin, and Leonard Azamfirei. 2024. "A Gender-Bias-Mitigated, Data-Driven Precision Medicine System to Assist in the Selection of Biological Treatments of Grade 3 and 4 Knee Osteoarthritis: Development and Preliminary Validation of precisionKNEE." *Cureus* 16 (3): e55832. https://doi.org/10.7759/cureus.55832.

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '19, May 2, 2019, 1–16. https://doi.org/10.1145/3290605.3300830.

Jaime, Sofia, and Christoph Kern. 2024. "Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '24, June 5, 2024, 237–53. https://doi.org/10.1145/3630106.3658902.

Juwara, Lamin, Alaa El-Hussuna, and Khaled El Emam. 2024. "An Evaluation of Synthetic Data Augmentation for Mitigating Covariate Bias in Health Data." *Patterns* 5 (4): 100946. https://doi.org/10.1016/j.patter.2024.100946.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2022. "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination." *Management Science* 68 (3): 1959–81. https://doi.org/10.1287/mnsc.2020.3850.

Kaushal, Amit, Russ Altman, and Curt Langlotz. 2020. "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms." *JAMA* 324 (12): 1212–13. https://doi.org/10.1001/jama.2020.12067.

Lahoti, Preethi, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. "Fairness without Demographics through Adversarially Reweighted Learning." arXiv:2006.13114. Preprint, arXiv, November 3. https://doi.org/10.48550/arXiv.2006.13114.

Libin, Alexander, Jonah T. Treitler, Tadas Vasaitis, and Yijun Shao. 2024. "Evaluating and Reducing Subgroup Disparity in AI Models: An Analysis of Pediatric COVID-19 Test Outcomes." *medRxiv*, September 19, 2024, 2024.09.18.24313889. https://doi.org/10.1101/2024.09.18.24313889.

Mahamadou, Abdoul Jalil Djiberou, and Artem A. Trotsyuk. 2025. "Revisiting Technical Bias Mitigation Strategies." *Annual Review of Biomedical Data Science*, ahead of print, April 8, 2025. https://doi.org/10.1146/annurev-biodatasci-103123-095737.

Malladi, Suresh S., and Hemang C. Subramanian. 2020. "Bug Bounty Programs for Cybersecurity: Practices, Issues, and Recommendations." *IEEE Software* 37 (1): 31–39. https://doi.org/10.1109/MS.2018.2880508.

Maron, Barry J., Milind Y. Desai, Rick A. Nishimura, Paolo Spirito, Harry Rakowski, Jeffrey A. Towbin, Ethan J. Rowin, Martin S. Maron, and Mark V. Sherrid. 2022. "Diagnosis and Evaluation of Hypertrophic Cardiomyopathy: JACC State-of-the-Art Review." *Journal of the American College of Cardiology* 79 (4): 372–89. https://doi.org/10.1016/j.jacc.2021.12.002.

Mickel, Jennifer. 2024. "Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What

They Represent." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '24, June 5, 2024, 2484–94. https://doi.org/10.1145/3630106.3659050.

Patil, Pranita, and W. Randolph Ford. 2024. "Parkinson's Disease Recognition Using Decorrelated Convolutional Neural Networks: Addressing Imbalance and Scanner Bias in Rs-fMRI Data." *Biosensors* 14 (5): 259. https://doi.org/10.3390/bios14050259.

Salahuddin, Zohaib, Yi Chen, Xian Zhong, Henry C. Woodruff, Nastaran Mohammadian Rad, Shruti Atul Mali, and Philippe Lambin. 2023. "From Head and Neck Tumour and Lymph Node Segmentation to Survival Prediction on PET/CT: An End-to-End Framework Featuring Uncertainty, Fairness, and Multi-Region Multi-Modal Radiomics." *Cancers* 15 (7): 1932. https://doi.org/10.3390/cancers15071932.

Schrouff, Jessica, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, et al. 2024. "Diagnosing Failures of Fairness Transfer across Distribution Shift in Real-World Medical Settings." *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS '22, April 3, 2024, 19304–18.

Simson, Jan, Alessandro Fabris, and Christoph Kern. 2024. "Lazy Data Practices Harm Fairness Research." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '24, June 5, 2024, 642–59. https://doi.org/10.1145/3630106.3658931.

Tejani, Ali S., Yee Seng Ng, Yin Xi, and Jesse C. Rayan. 2024. "Understanding and Mitigating Bias in Imaging Artificial Intelligence." *RadioGraphics*, ahead of print, April 18, 2024. world. https://doi.org/10.1148/rg.230067.

U.S. National Cancer Institute. 2017. *Occupation, the Work Environment, and Tobacco-Related Health Disparities*. A Socioecological Approach to Addressing Tobacco-Related Health Disparities. https://cancercontrol.cancer.gov/brp/tcrb/monographs/monograph-22.

Verma, Sahil, and Julia Rubin. 2018. "Fairness Definitions Explained." *Proceedings of the International Workshop on Software Fairness* (New York, NY, USA), FairWare '18, May 29, 2018, 1–7. https://doi.org/10.1145/3194770.3194776.

Watson, Matthew, Pinkie Chambers, Luke Steventon, James Harmsworth King, Angelo Ercia, Heather Shaw, and Noura Al Moubayed. 2024. "From Prediction to Practice: Mitigating Bias and Data Shift in Machine-Learning Models for Chemotherapy-Induced Organ Dysfunction across Unseen Cancers." *BMJ Oncology* 3 (1): e000430. https://doi.org/10.1136/bmjonc-2024-000430.

WHO. n.d.-a. "Health Technology Assessment." Accessed August 10, 2025. https://www.who.int/health-topics/health-technology-assessment.

WHO. n.d.-b. "Social Determinants of Health." Accessed April 18, 2025. https://www.who.int/health-topics/social-determinants-of-health.

Woo, MinJae, Linglin Zhang, Beatrice Brown-Mulry, In-Chan Hwang, Judy Wawira Gichoya, Aimilia Gastounioti, Imon Banerjee, Laleh Seyyed-Kalantari, and Hari Trivedi. 2025. "Subgroup Evaluation to Understand Performance Gaps in Deep Learning-Based Classification of Regions of Interest on Mammography." *PLOS Digital Health* 4 (4): e0000811. https://doi.org/10.1371/journal.pdig.0000811.

Yaseliani, Mohammad, Md Noor-E-Alam, and Md Mahmudul Hasan. 2024. "Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework." *JMIR AI* 3 (August): e55820. https://doi.org/10.2196/55820.

Zajacova, Anna, and Elizabeth M. Lawrence. 2018. "The Relationship between Education and Health: Reducing Disparities through a Contextual Approach." *Annual Review of Public Health* 39 (April): 273–89. https://doi.org/10.1146/annurev-publhealth-031816-044628.

Zhang, Jiaqing, Sabyasachi Bandyopadhyay, Faith Kimmet, Jack Wittmayer, Kia Khezeli, David J. Libon, Catherine C. Price, and Parisa Rashidi. 2023. "FaIRClocks: Fair and Interpretable Representation of the Clock Drawing Test for Mitigating Classifier Bias against Lower Educational Groups." *Research Square*, October 9, 2023, rs.3.rs-3398970. https://doi.org/10.21203/rs.3.rs-3398970/v1.