arXiv:2009.10576v1 [cs.CY] 22 Sep 2020

Ethical Machine Learning in Health Care

Irene Y. Chen,[1] Emma Pierson,[2] Sherri Rose,[3] Shalmali Joshi,[4] Kadija Ferryman,[5] and Marzyeh Ghassemi[4,6]

[1]Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA; email: iychen@mit.edu
[2]Microsoft Research, Cambridge, MA, 02143, USA
[3]Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA, 94305, USA
[4]Vector Institute, Toronto, ON, Canada
[5]Department of Technology, Culture, and Society, Tandon School of Engineering, New York University, Brooklyn, NY, 11201, USA
[6]Department of Computer Science, University of Toronto, Toronto, ON, Canada

Xxxx. Xxx. Xxx. Xxx. 2021. AA:1–24

https://doi.org/10.1146/((please add article doi))

Copyright © 2021 by Annual Reviews. All rights reserved

Keywords

machine learning, bias, ethics, health, health care, health disparities

Abstract

The use of machine learning (ML) in health care raises numerous ethical concerns, especially as models can amplify existing health inequities. Here, we outline ethical considerations for equitable ML in the advancement of health care. Specifically, we frame ethics of ML in health care through the lens of social justice. We describe ongoing efforts and outline challenges in a proposed pipeline of ethical ML in health, ranging from problem selection to post-deployment considerations. We close by summarizing recommendations to address these challenges.

Contents

1. INTRODUCTION 2
2. PROBLEM SELECTION 4
   2.1. Global Health Injustice 4
   2.2. Racial Injustice 5
   2.3. Gender Injustice 5
   2.4. Diversity of the Scientific Workforce 5
3. DATA COLLECTION 6
   3.1. Heterogeneous Data Losses 6
   3.2. Population-specific Data Losses 7
4. OUTCOME DEFINITION 8
   4.1. Clinical Diagnosis 9
   4.2. Health Care Costs 9
5. ALGORITHM DEVELOPMENT 10
   5.1. Understanding Confounding 10
   5.2. Feature Selection 11
   5.3. Tuning Parameters 12
   5.4. Performance Metrics 12
   5.5. Group Fairness Definition 12
6. POST-DEPLOYMENT CONSIDERATIONS 13
   6.1. Quantifying Impact 13
   6.2. Model Generalizability 14
   6.3. Model and Data Documentation 14
   6.4. Regulation 14
7. RECOMMENDATIONS 15

1. INTRODUCTION

Machine learning (ML): The study of computer algorithms that improve automatically through experience

ML model: An algorithm that has been trained on data for a specific use case

Algorithm: A finite sequence of well-defined instructions used to solve a class of problems

As machine learning (ML) models proliferate into many aspects of our lives, there is growing concern regarding their ability to inflict harm. In medicine, excitement about human-level performance (1) of ML for health is balanced against ethical concerns, such as the potential for these tools to exacerbate existing health disparities (2, 3, 4, 5). For instance, recent work has demonstrated that state-of-the-art clinical prediction models underperform on women, ethnic and racial minorities, and those with public insurance (6). Other research has shown that when popular contextual language models are trained on scientific articles, they complete clinical note templates to recommend "hospitals" for violent white patients and "prison" for violent Black patients (7). Even more worrisome, health care models designed to optimize referrals to long-term care-management programs for millions of patients were found to exclude Black patients with similar health conditions compared to white patients from care management programs (8).

A growing body of literature wrestles with the social implications of machine learning and technology. Some of this work, referred to as critical data studies, is from a social science perspective (9, 10), whereas other work leads with a technical and computer science perspective (11, 12, 13). While there is scholarship addressing social implications and algorithmic fairness in general, there has been less work at the intersection of health, ML, and fairness (14, 15, 16), despite the potential life-or-death impacts of those models (8, 17).

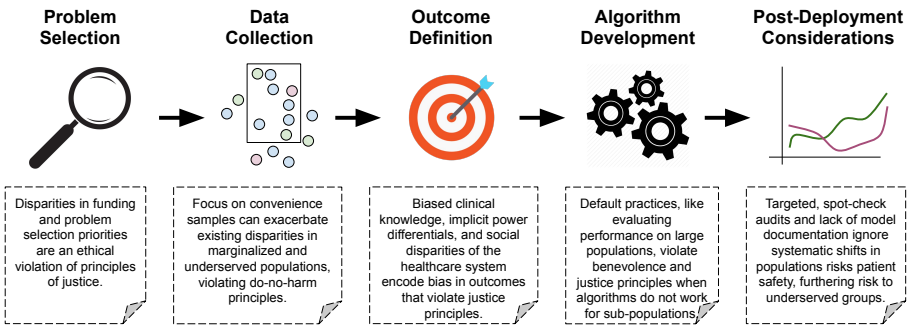


Figure 1

We motivate the five steps in the ethical pipeline for health care model development. Each stage contains considerations for ML where ignoring technical challenges violate the bioethical principle of justice, either by exacerbating existing social injustices or by creating the potential for new injustices between groups. Although this review's ethical focus is on social justice, the challenges that we highlight may also violate ethical principles such as justice and beneficence. We highlight a few in this illustration.

While researchers looking to develop ethical ML models for health can begin by drawing on bioethics principles (18, 19), these principles are designed to inform clinical care practices. How these principles could inform the ML model development pipeline remains understudied. We note that there has been significant work on other important ethical issues that relate to ML and health, including reviews of consent and privacy (20), which we do not address here. Instead, we focus on equity in ML models that operate on health data. We focus primarily on differences between groups induced by, or related to, the model development pipeline, drawing on both the bioethics principle of justice, and the established social justice centering of public health ethics (21). Unjust differences in quality and outcomes of health care between groups often reflect existing societal disparities for disadvantaged groups. We consider other bioethics principles such as benevolence and non-maleficence, but focus them primarily on groups of patients rather than on individuals.

We organize this review by describing the ethical considerations that arise at each step of the pipeline during model development for ML in health (**Figure 1**), from research funding to post-deployment. Here we motivate the ethical considerations in the pipeline with a case study of Black mothers in the United States, who die in childbirth at a rate three times higher than white women (22). This inequity is unjust because it connects to a history of reproductive injustices faced by Black women in the United States, from gynecological experimentation on enslaved women to forced sterilizations (23, 24).

1. This disparity occurs in part during **problem selection** because maternal mortality is an understudied problem (25).
2. Even after accounting for problem selection, **data collection** from hospitals may differ in quality and quantity. For example, 75% of Black women give birth at hospitals that serve predominantly Black patients (26) but Black-serving hospitals have higher rates of maternal complications than other hospitals (27).
3. Once data are collected, the choice of **outcome definition** can obscure underlying issues, e.g., differences in clinical practice. General model outcome definitions for

Ethical pipeline: The model development process and the corresponding ethical considerations

Bioethics: The study of ethical issues emerging from advances in biology and medicine

Justice: The principle that obligates equitably distributed benefits, risks, costs, and resources

Benevolence: The principle that requires that care be provided with the intent of doing good for the patient involved

Non-maleficence: The principle that forbids harm or injury to the patient, either through acts of commission or omission

Model outcome: The output of interest for predictive models

Deployment: The process through which a machine learning model is integrated into an existing production environment

Risk score: A calculated number denoting the likelihood of adverse event

maternal health complications might overlook conditions specific to Black mothers, e.g., fibroids (28).

4. During **algorithm development**, models may not be able to account for the confounding presence of societal bias. Black mothers in the wealthiest neighborhoods in Brooklyn, New York have worse outcomes than white, Hispanic, and Asian mothers in the poorest ones, demonstrating a gap despite factors that should improve Black mothers' outcomes — living in the same place, and having a higher income — likely due to societal bias that impacts Black women (29).
5. Finally, after a model is trained, **post-deployment considerations** may not fully consider the impact of deploying a biased prediction model into clinical settings that have large Black populations. Because Black women have a heightened risk of pregnancy-related death across income and education levels (30), a biased prediction model could potentially automate policies or risk scores that disadvantage Black mothers.

We organize the rest of this review sequentially expanding on each of the five steps in the pipeline described above and in **Figure 1**. First, we look at problem selection, and explain how funding for ML for health research can lead to injustice. We then examine how data collection processes in funded research can amplify inequity and unfairness. We follow this by exploring outcome definition and algorithm building, listing the multitude of factors that can impact model performance, and how these differences in performance relate to issues of justice. We close with audits that should be considered for more robust and just deployments of models in health, and recommendations to practitioners for ethical, fair, and just ML deployments.

2. PROBLEM SELECTION

There are many factors that influence the selection of a research problem, from interest to available funding. However, problem selection can also be a matter of justice if the research questions that are proposed, and ultimately funded, focus on the health needs of advantaged groups. Below we provide examples of how disparities in research teams and funding priorities exacerbate existing socioeconomic, racial, and gender injustices.

2.1. Global Health Injustice

The "10/90" gap refers to the fact that the vast majority of health research dollars are spent on problems that affect a small fraction of the global population (31, 32). Diseases that are most common in lower-income countries receive far less funding than diseases that are most common in high-income countries (33) (relative to the number of individuals they affect). As an example, 26 poverty-related diseases account for 14% of the global disease burden, but receive only 1.3% of global health-related research and development expenditure. Nearly 60% of the burden of poverty-related neglected diseases occurs in Western and Eastern sub-Saharan Africa as well as South Asia. Malaria, tuberculosis, and HIV/AIDS all have shares of global health-related research and development expenditure that are at least five times smaller than their share of global disease burden (33). This difference in rates of funding represents an injustice because it further exacerbates the disadvantages faced by Global South populations. While efforts like the "All Of Us" Project (34) and the 23andMe's Call for Collaboration (35) seek to collect more inclusive data, these efforts have come

under criticism for not reflecting global health concerns, particularly among Indigenous groups (36).

Global South: Countries on one side of the NorthSouth divide, the other side being the countries of the Global North

2.2. Racial Injustice

Racial bias affects which health problems are prioritized and funded. For example, sickle cell disease and cystic fibrosis are both genetic disorders of similar severity, but sickle cell disease is more common in Black patients, while cystic fibrosis is more common in white patients. In the United States (US), however, cystic fibrosis receives 3.4 times more funding per affected individual from the US National Institutes of Health (NIH), the largest funder of US clinical research, and hundreds of times more private funding (37). The disparities in funding persist despite the 1972 Sickle Cell Anemia Control Act, which recognizes that sickle cell has been neglected by the wider research community. Further, screening for sickle cell disease is viewed by some as unfair targeting (38), and Black patients with the disease who seek treatment are often maligned as drug abusers (39).

2.3. Gender Injustice

Women's health conditions like endometriosis are poorly understood; as a consequence, even basic statistics like the prevalence of endometriosis remain unknown, with estimates ranging from 1% to 10% of the population (40, 41). Similarly, the menstrual cycle is stigmatized and understudied (40, 42), producing a dearth of understanding that undermines the health of half the global population. Basic facts about the menstrual cycle — including which menstrual experiences are normal and which are predictive of pathology — remain unknown (40). This lack of focus on the menstrual cycle propagates into clinical practice and data collection despite evidence that it affects many aspects of health and disease (43, 44). Menstrual cycles are also not often recorded in clinical records and global health data (40). In fact, the NIH did not have an R01 grant, the NIH's original and historically oldest grant mechanism, relating to the influence of sex and gender on health and disease until 2019 (45). Notably, recent work has moved to target such understudied problems via ambulatory women's health-tracking mobile apps. These crowd-sourcing efforts stand to accelerate women's health research by collecting data from cohorts that are orders of magnitude larger than those used in previous studies (40).

2.4. Diversity of the Scientific Workforce

The diversity of the scientific workforce profoundly influences the problems studied, and contributes to the biases in problem selection (46). Research shows that scientists from underrepresented racial and gender groups tend to prioritize different research topics. They produce more novel research, but their innovations are taken up at lower rates (47). Female scientists tend to study different scientific subfields, even within the same larger field — for example, within sociology, they have been historically better-represented on papers about sociology of the family or early childhood (48) — and express different opinions about ethical dilemmas in computer science (49). Proposals from white researchers in the US are more likely to be funded by the NIH than proposals from Black researchers (50, 51), which in turns affects what topics are given preference. For example, a higher fraction of NIH proposals from Black scientists study community and population-level health (50). Overall, this evidence suggests that diversifying the scientific workforce will lead to problem selection

that more equitably represents the interests and needs of the population as a whole.

3. DATA COLLECTION

The role of health data is ever-expanding, with new data sources routinely being integrated into decision-making around health policy and design. This wealth of high-quality data, coupled with advancements in ML models, has played a significant role in accelerating the use of computationally informed policy and practice to strengthen health care and delivery platforms. Unfortunately, data can be biased in ways that have (or can lead to) disproportionate negative impacts on already marginalized groups. First, data on group membership can be completely absent. For instance, countries such as Canada and France do not record race and ethnicity in their nationalized health databases (52, 53), making it impossible to study race-based disparities and hypotheses around associations of social determinants of health. Second, data can be imbalanced. Recent work on acute kidney injury achieved state-of-the-art prediction performance in a large dataset of 703,782 adult patients using 620,000 features; however, they note that model performance was lower in female patients since female patients comprised 6.38% of patients in the training data (54). Other work has indicated that this issue can not be simply addressed by "pre-training" a model in a more balanced data setting prior to fine-tuning on an imbalanced dataset (55). This indicates that a model cannot be "initialized" with a balanced baseline representation which ameliorates issues of imbalance in downstream tasks, and suggests we must solve this problem at the root, be it with more balanced are comprehensive data, specialty learning algorithms, or combinations therein.Finally, while some sampling biases can be recognized and possibly corrected, others may be difficult to correct. For example, work in medical imaging has demonstrated that models may overlook unforeseen stratification of conditions, like rare manifestations of diseases, which can result in harm in clinical settings (56, 16).

In this section, we discuss common biases in data collection. We consider two types of processes that result in a loss of data. First, processes that affect what kind of information is collected, or heterogenous data loss, across varying input types. For example, clinical trials with aggressive inclusion criteria or social media data that reflects those with access to devices. Second, we examine processes that affect whether an individual's information is collected, or population-specific data losses, where individuals are impacted by their population type, often across data input categories. For example, undocumented immigrants may fear deportation if they participate in health care systems.

3.1. Heterogeneous Data Losses

Some data loss is specific to the data type, due to assumptions about noise that may have been present during the collection process. However, data noise and missingness can cause unjust inequities that impact populations in different ways. We cover four main data types: randomized controlled trials (RCTs), electronic health care records (EHR), administrative health data, and social media data.

Randomized Controlled Trials Randomized controlled trials are often run specifically to gather "unbiased" evidence of treatment effects. However, RCTs have notoriously aggressive exclusion (or inclusion) criteria (57), which create study cohorts that are not representative of general patient populations (58). In one study of RCTs used to define asthma treatment, an estimated 94% of the adult asthmatic population would not have

been eligible for the trials (59). There is a growing methodological literature designing methods to generalize RCT treatment effects to other populations (60). However, current empirical evidence indicates that such generalizations can be challenging given available data or may require strong assumptions in practice.

Electronic Health Records Much recent work in ML also leverages large electronic health records data. EHR data are a complex reflection of patient health, health care systems, and providers, where data missingness is a known, and meaningful, problem (61). As one salient example, a large study of laboratory tests to model three-year survival found that health care process features had a stronger predictive value than the patient's physiological feaures (62). Further, not all treatments investigated in RCTs can be easily approximated in EHR (63).

Biases in EHR data may arise due to differences in patient populations, access to care, or the availability of EHR systems (64). As an example, the widely-used MIMIC-III EHR dataset includes most patients who receive care at the intensive care units in Beth Israel Deaconess Medical Center (BIDMC), but this sample is obviously limited by which individuals have access to care at BIDMC, which has a largely white patient population (14). In the United States, uninsured Black and Hispanic or Latin(o/x) patients, as well as Hispanic or Latin(o/x) Medicaid patients, are less likely to have primary care providers with EHR systems, as compared to white patients with private insurance (65). Other work has shown that gender discrimination in health care access has not been systematically studied in India, primarily due to a lack of reliable data (66).

Administrative Health Records In addition to RCTs and EHR, health care billing claims data, clinical registries, and linked health survey data are also common data sources in population health and health policy research (67, 68), with known biases in which populations are followed, and who is able to participate. Translating such research into practice is a crucial part of maintaining health care quality, and limited participation of minority populations by sexual orientation and gender identity (69), race and ethnicity (70), and language (71) can lead to health interventions and policies that are not inclusive, and can create new injustices for these already marginalized groups.

Social Media Data Data from social media platforms and search-based research by nature consists only of individuals with internet access (72). Even small choices like limiting samples to those from desktop versus mobile platforms are a problematic distinction in non-North American contexts (73). Beyond concerns about access to resources or geographic limitations, data collection and scraping pipelines for most social media cohorts do not yield a random sample of individuals. Further, the common practice of limiting analysis to those satisfying a specified threshold of occurrence can lead to skewed data. As an example, when processing the large volume of Twitter data (7.6 billion tweets) researchers may first restrict to users who can be mapped to a US county (1.78 billion), then to those Tweets that contain only English (1.64 billion tweets), and finally remove users who made less than 30 posts (1.53 billion) (74).

3.2. Population-specific Data Losses

As with data types, the modern data deluge does not apply equally to all communities. Historically underserved groups are often underrepresented, misrepresented, or entirely missing from health data that inform consequential health policy decisions. When individuals from disadvantaged communities appear in observational datasets, they are less likely to be

Training data: Information that a ML model fits to and learns patterns from

Heterogeneous data loss: The process where data can be lost in collection due to data type

Population-specific data loss: The process where data can be lost in collection due to the features of the population

Data noise: Meaningless information added to data that obscures the underlying information of the data

Missingness: The manner in which data is absent from a sample of the population

Randomized controlled trial (RCT): A study in which subjects are allocated by chance to receive one of several interventions

Electronic health record (EHR): Digital version of a patient's clinical history that is maintained by the provider over time

Intervention: A treatment, procedure, or other action taken to prevent or treat disease, or improve health in other ways

accurately captured due to errors in data collection and systemic discrimination. Larger genomics datasets often target European populations, producing genetic risk scores which are more accurate in individuals of European ancestry than other ancestries (75). We note four specific examples of populations that are commonly impacted: low- and middle-income nationals, transgender and gender non-conforming individuals, undocumented migrants, and pregnant women.

Low- and Middle-Income Nationals Health data are infrequently collected due to resource constraints, and even basic disease statistic data such as prevalence of mortality rates can be challenging to find for low- and middle-income nations (73). When data are collected, it is not digitized, and often contains errors. In 2001, the World Health Organization found that only 9 out of the 46 member states in Sub-Saharan Africa could produce death statistics for a global assessment of the burden of disease, with data coverage often less than 60% in these countries (76).

Transgender and Gender Non-conforming Individuals The health care needs and experiences of transgender and gender non-conforming individuals are not well-documented in datasets (77) because documented sex, not gender identity, is what is usually available. However, documented sex is often discordant with gender identity for transgender and gender non-conforming individuals. Apart from health documentation concerns, transgender people are often concerned about their basic physical safety when reporting their identities. In the US, it was only in 2016, with the release of the US Transgender Survey that there was a meaningfully sized dataset — 28,000 respondents — to enable significant analysis and quantification of discrimination and violence that transgender people face (77).

Undocumented Immigrants Safety concerns are important in data collection for undocumented migrants, where socio-political environments can lead to individuals feeling unsafe during reporting opportunities. When immigration policies limit access to public services for immigrants and their families, these restrictions lead to spillover effects on clinical diagnoses. As one salient example, autism diagnoses for Hispanic children in California fell following aggressive federal anti-immigrant policies requiring citizenship verification at hospitals (78).

Pregnant Women Despite pregnancy being neither rare nor an illness, the US continues to experience rising maternal mortality and morbidity rates. In the US, the maternal mortality rate has more than doubled from 9.8 per 100,000 live births in 2000 to 21.5 in 2014 (79). Importantly, disclosure protocols recommend suppression of information in nationally available datasets when the number of cases or events in a data "cell" is low, to reduce the likelihood of a breach of confidentiality. For example, the US Centers for Disease Control suppresses numbers for counties with fewer than 10 deaths for a given disease (80). Although these data omissions occur because of patient privacy, such censoring on the *dependent* variable introduces particularly pernicious statistical bias and, as a result, much remains to be understood about what community, health facility, patient, and provider-level factors drive high mortality rates.

Censoring: The mechanism through which data values are removed from observation

4. OUTCOME DEFINITION

The next step in the model pipeline is to define the outcome of interest for a health care task. Even seemingly straightforward tasks like defining whether a patient *has* a disease can be skewed by how prevalent diseases are, or how they manifest in some patient populations. For example, a model predicting if a patient will develop heart failure will need labeled examples

both of patients who have heart failure, and patients without heart failure. Choosing these patients can rely on parts of the EHR that may be skewed due to lack of access to care, or abnormalities in clinical care: e.g., economic incentives may alter diagnosis code logging (81), clinical protocol affects the frequency and observation of abnormal tests (62), historical racial mistrust may delay care and affect patient outcomes (82), and naive data collection can yield inconsistent labels in chest X-rays (56). Such biased labels, and the resulting models, may cause clinical practitioners to allocate resources poorly.

We discuss social justice considerations in two examples of commonly modelled health care outcomes: clinical diagnosis and health care costs. In each example, it is essential that model developers choose a reliable proxy and account for noise in the outcome labels as these choices can have a large impact on performance and equity of the resulting model.

4.1. Clinical Diagnosis

Clinical diagnosis is a fundamental task for clinical prediction models, e.g., models for computer-aided diagnosis from medical imaging. In clinical settings, researchers often select patient disease occurrence as the prediction label for models. However, there are many options for the choice of a disease occurrence label. For example, the outcome label for developing cardiovascular disease could be defined through the occurrence of specific phrases in clinical notes. However, women can manifest symptoms of acute coronary syndrome differently (83) and receive delayed care as a result (84), which may then manifest in diagnosis labels derived from the clinical notes being gender-skewed. Because differences in label noise results in disparities in model impact, researchers have the responsibility to choose and improve disease labels, so that these inequalities do not further exacerbate disparities in health.

Label noise: Errors or otherwise obscuring information that affects the quality of the labels

Diagnosis code: A label in patient records of disease occurrence, which may be subject to misclassification, used primarily for billing purposes

Additionally, it is important to consider the health care system in which disease labels are logged. For example, health care providers leverage diagnosis codes for billing purposes, not clinical research. As a result, diagnosis codes can create ambiguities because of overlap and hierarchy in codes. Moreover, facilities have incentives to under-report (81) and over-report (85, 86) outcomes, yielding differences in model representations.

Recent advances in improving disease labels target statistical corrections based on estimates of the label noise. For instance, a positive label may be reliable, but the omission of a positive label could either indicate a negative label (i.e., no disease) or merely a missed positive label. Methods to address the positive-unlabeled setting use estimated noise rates (87) or hand-curated labels that are strongly correlated with positive labels, known also as "silver-standard" labels, from clinicians (88). Clinical analysis of sources of error in disease labels can also guide improvements (89) and identify affected groups (56).

4.2. Health Care Costs

Developers of clinical models may choose to predict health care costs, meaning the ML model seeks to predict which patients will cost the health care provider more in the future. Some model developers may use health care costs as a proxy for future health needs to guide accurate targeting of interventions (8), with the underlying assumption that addressing patients with future health need will limit future costs. Others may explicitly want to understand patients who will have high health care cost to reduce the total cost of health care (90). However, because socioeconomic factors affect both access to health care *and* access to financial resources, these models may yield predictions that exacerbate inequities.

For model developers seeking to optimize for health need, health care costs can deviate from health need on an individual level because of patient socioeconomic factors. For instance, in a model used to allocate care management program slots to high-risk patients, the choice of future health care costs as a predictive outcome led to racial disparities in patient allocation to the program (8). Health care costs can differ from health need on an institutional level due to underinsurance and undertreatment within the patient population (91). After defining health disparities as all differences except those due to clinical need and preferences, researchers have found racial disparities in mental health care. Specifically, white patients had higher rates of initiation of treatment for mental health compared to Black and Hispanic or Latin(o/x) patients. Because the analysis controls for health need, the disparities are solely a result of differences in health care access and systemic discrimination (92).

Addressing issues that arise from the use of health care costs depends on the setting of the ML model. In cases where health need is of highest importance, a natural solution is to choose another outcome definition besides health care costs, e.g., the number of chronic diseases as a measure of health need. If a model developer is most concerned with cost, it is possible to correct for health disparities in predicting health care costs by building fairness considerations directly into the predictive model objective function (93). Building these types of algorithmic procedures is further discussed in Section 5.

5. ALGORITHM DEVELOPMENT

Loss function: The relation that determines the error between algorithm output and given label, which the algorithm uses to optimize

Algorithm development considers the construction of the underlying computation for the ML model and presents a major vulnerability and opportunity for ethical ML in health care. Just as data are not neutral, algorithms are not neutral. A disproportionate amount of power lies with research teams who, after determining the research questions, make decisions about critical components of an algorithm such as the loss function (94, 46). In the case of loss functions, common choices like the L_1 absolute error loss and L_2 squared error loss do not target the same conditional functions of the outcome, instead minimizing the error in the median and mean respectively. Using a surrogate loss (e.g., hinge loss for the error rate) can provide computational efficiency, but it may not reflect the ethical criteria that we truly care about. Recent work has shown that models trained with a surrogate loss may exhibit "approximation errors" that disproportionately affect undersampled groups in the training data (95). Similarly, one might choose to optimize the worst-case error across groups as opposed to the average overall error. Such choices may seem purely technical, but reflect value statements about what should be optimized, potentially leading to differences in performance among marginalized groups (96).

In this section, we review several crucial factors in model development that potentially impact ethical deployment capacity: understanding (and accounting for) confounding, feature selection, tuning parameters, and defining "fairness" itself.

5.1. Understanding Confounding

Developing models that use sensitive attributes without a clear causal understanding of their relationship to outcomes of interest can significantly affect model performance and interpretation. This is relevant to algorithmic problems focused on prediction, not just causal inference. "Confounding" features — i.e., those features that influence both the

independent variables and the dependent variable — require careful attention. The vast majority of models learn patterns based on observed correlations between training data, even when such correlations do not occur in test data. For instance, recent work has demonstrated that classification models designed to detect hair color learn gender-biased decision boundaries when trained on confounded data, i.e., if women are primarily blond in training data, the model incorrectly associates gender with the hair label in test samples (97).

As ML methods are increasingly used for clinical decision support, it is critical to account for confounding features. In one canonical example, asthmatic patients presenting with pneumonia are given aggressive interventions that ultimately improve their chances of survival over non-asthmatic patients (98). When the hospital protocol assigned additional treatment to patients with asthma, those patients had improved outcomes. Thus the treatment policy was a confounding factor that altered the data in a seemingly straight-forward prediction task such that patients with asthma were erroneously predicted by models to have *lower* risk of dying from pneumonia.

Simply controlling for confounding features by including them as features in classification or regression models may be insufficient to learn reliable models because features can have a mediating or moderating effect (post-treatment effect on outcomes of interest) and have to be incorporated differently into model design (99).

Modern ML and causal discovery techniques can identify sources of confounding at scale (100), although validation of such methods can be challenging because of the lack of counterfactual data. ML methods have also been proposed to estimate causal effects from observational data (101, 102). In practice, when potential hidden confounding is suspected, either mediating features or proxies can be leveraged (103, 99) or sensitivity analysis methods can be used to determine potential sources of errors in effect estimates (104). Data-augmentation and sampling methods may also be used to mitigate effects of model confounding. For example, augmenting X-ray images with rotated and translated variants can help train a model that is not sensitive to orientation of an image (105).

Confounding: The condition in which a feature influences both the dependent variable and independent variable, causing a spurious association

Test data: Unseen information that a model predicts on and is evaluated against

5.2. Feature Selection

With large-scale digitization of EHR and other sources, sensitive attributes like race and ethnicity may be increasingly available (although prone to misclassification and missingness). However, blindly incorporating factors like race and gender in a predictive model may exacerbate inequities for a wide range of diagnostics and treatments (106). These resulting inequities can lead to unintended and permanent embedding of biases in algorithms used for clinical care. For example, vaginal birth after cesarean (VBAC) scores are used to predict success of "trial of labor" of pregnant women with a prior cesarean section; however, these scores explicitly include a race component as an input which reduces the chance of VBAC success for Black and Hispanic women. Although researchers found that previous observational studies showed correlation between racial identity and success of trial of labor (107), the underlying cause of this association is not well-understood. Such naïve inclusion of race information could exacerbate disparities in maternal mortality. This ambiguity calls race-based 'correction' in scores like VBAC into question (106).

Automation in feature selection does not eliminate the need for contextual understanding. For example, stepwise regression is commonly used and taught as a technique for feature selection despite known limitations (108). While specific methods have varying ini-

Sensitive attribute: A specified patient feature (e.g., race, gender) which is considered important for fairness considerations

Stepwise regression: A method of estimation where each feature is sequentially considered by addition or subtraction to the existing feature set

tialization (e.g., start with an empty set of features or full set of features) and processing steps (e.g., deletion vs. addition of features), most rely on p-values, R^2, or other global fit metrics to select features. Weaknesses of stepwise regressions include the misleading nature of p-values and the final set depending on if and when features were considered (109). In ML, penalized regressions like lasso regression are popular for automated feature selection, but the lasso trades potential increases in estimation bias for reductions in variance by shrinking some feature coefficients to zero. Features selected by lasso may be co-linear with other features not selected (110). Over-interpretation of the selected features in any automated procedures should therefore be avoided in practice given these pitfalls. Researchers should also consider the humans-in-the-loop framework where incorporation of automated procedures is blended with investigator knowledge (111).

Tuning parameters: Algorithm components used for prediction that are tuned toward solving an optimization problem

AUC: A measure of the sensitivity and specificity of a model for each decision threshold

AUPRC: A measure of precision and recall of a model for each decision threshold

Calibration: A measure of how well ML risk estimates reflect true risk

5.3. Tuning Parameters

There are many tuning parameters that may be set a priori or selected via cross-validation (110). These range from the learning rate in a neural network to the minimum size of the terminal leaves in a random forest. In the latter example, default settings in R for classification will allow trees to grow until there is just one observation in a terminal leaf. This can lead to overfitting the model to the training data and a loss of generalizability to the target population. Lack of generalizability is a central concern for ethical ML given the previously discussed issues in data collection and study inclusion. When data lack diversity and are not representative of the target population where the model would be deployed, overfitting algorithms to this data has the potential to disproportionately harm marginalized groups (112). Using cross-validation to select tuning parameters does not automatically solve these problems as cross-validation still operates with respect to an a priori chosen optimization target.

5.4. Performance Metrics

There are many commonly used performance metrics for model evaluation such as area under the receiver-operating characteristic (AUC), precision-recall curves (AUPRC), and calibration (113). However, the appropriate metrics to optimize depend on intended use case and relative value of true positives, false positives, true negatives, and false negatives. Not only can AUC be misleading when considering other global fit metrics (e.g., high AUC masking weak true positive rate), but it does not describe the impact of the model across selected groups. Further, even "objective" metrics and scores can be deeply flawed, and lead to over or under-treatment of minorities if blindly applied (114). Note that robust reporting of results should include an explicit statement of other non-optimized metrics, including the original intended use case, the training cohort and case, or level of model uncertainty.

5.5. Group Fairness Definition

The specific definition of fairness for a given application often impacts the choice of a loss function, and therefore the underlying algorithm. Individual fairness imposes classifier performance requirements that operate over pairs of individuals, e.g., similar individuals should be treated similarly (115). Group fairness operates over "protected groups" (based on some sensitive attribute) by requiring that a classifier performance metric be balanced across

those groups. (116, 117). For instance, a model may be partially assessed by calculating the true positive rate separately among rural and urban populations to ensure risk score similarity. Regressions subject to group fairness constraints or penalties optimizing toward joint global and group fit considerations have also been developed (118, 119, 93).

Recent work has focused on identifying and mitigating violations of fairness definitions in health care settings. While most of these algorithms have emerged outside the field of health care, researchers have designed penalized and constrained regressions to improve the performance of health insurance plan payment. This payment system impacts tens of millions of lives in the United States and is known to undercompensate insurers for individuals with certain health conditions, including mental health and substance use disorders, in part because billing codes do not accurately capture diagnoses (120). Undercompensation creates incentives for insurers exclude individuals with these health conditions from enrollment, limiting their access to care. Regressions subject to group fairness constraints or penalties were successful in removing nearly all undercompensation for a single group with negligible impacts on global fit (93). Subsequent work incorporating multiple groups into the loss function also saw improvements in undercompensation for the majority of groups not included (121).

Performance metric: Score or other quantitative representation of a model's quality and ability to achieve goals

Algorithmic fairness: The study of definitions and methods related to the justice of models

Group fairness: A principle where pre-defined patient groups should receive similar model performance

6. POST-DEPLOYMENT CONSIDERATIONS

Often the goal of model training is to ultimately deploy it in a clinical, epidemiological, or policy service. However, deployed models can have lasting ethical impact beyond the model performance measured in development. For example, in the inclusion of race in the clinical risk scores described earlier that may lead to chronic over- or under-treatment (106). Here we outline considerations for robust deployment by highlighting the need for careful performance reporting, auditing generalizability, documentation, and regulation.

6.1. Quantifying Impact

Unlike in other settings with high-stakes decisions, e.g., aviation, clinical staff performance is not audited by an external body (122). Instead, clinicians are often a self-governing body, relying on clinicians themselves to determine when a colleague is underperforming or in breach of ethical practice principles, e.g., through such tools as surgical morbidity and mortality conferences (123). Clinical staff can also struggle to keep abreast of what current best practice recommendations are, as these can change dramatically over time; one study found that more than 400 previously routine practices were later contradicted in leading clinical journals (124).

Hence, it is important to measure and address the downstream impact of models though audits for bias and examination of clinical impact (6). Regular "auditing" post-deployment, i.e., detailed inspection of model performance on various groups and outcomes, may reveal the impact of models on different populations (8) and identify areas of potential concern. Some recent work has targeted causal models in dynamic systems in order to reduce the severity of bias (125). Others have targeted bias reduction through model construction with explicit guarantees about balanced performance (16), or specifying groups which must have equal performance (126). Additionally, there is the possibility that models may help to de-bias current clinical care by reducing known biases against minorities (127) and disadvantaged majorities (128).

Model auditing: The post-deployment inspection of model performance on groups and outcomes

6.2. Model Generalizability

Generalizability: The ability of a model to apply in a setting different from the one in which it was trained

Data artifact: A flaw in data caused by equipment, techniques, or conditions that is unrelated to model output

As has been raised in previous sections, a crucial concern with model deployment is generalization. Any shifts in data distributions can significantly impact model performance when the settings for development and for deployment differ. For example, chest X-ray diagnosis models can have high performance on test data drawn from the same hospital but degrade rapidly on data from another hospital (129). Other work in gender bias on chest X-ray data has demonstrated both that small proportions of female chest x-rays degrades diagnostic performance accuracy in female patients (130), and that this is not simply addressed in all cases by adding in more female X-rays (131). Even within a single hospital, models trained on data from an initial EHR system data deteriorated significantly when tested on data from a new EHR system (132). Finally, data artifacts that induce strong priors in what patterns ML models are sensitive to have the potential to perpetrate harms when used without awareness (133). For example, patients with dark skin can have morphological variation and disease manifestations that are not easily detected under the defaults that are set by predominantly white-skinned patients (134).

Several algorithms have recently been proposed to account for distribution shift in data (135, 136). However, these algorithms have significant limitations as they typically require assumptions about the nature or amount of distributional shift an algorithm can accommodate. Some, like (136), may require a clear indication of which distributions in a health care pipeline are expected to change, and develop models for prediction accordingly. Many of these assumptions may be verifiable. If not, periodically monitoring for data shifts (137), and potentially retraining models when performance deteriorates due to such shifts is an imperative deployment consideration with significant ethical implications.

6.3. Model and Data Documentation

Clear documentation enables insight into the model development and data collection. Good model documentation should include clinically specific features of model development that can be assessed and recorded beforehand, such as logistics within the clinical setting, potential unintended consequences, and trade-offs between bias and performance (138). In addition to raising ethical concerns in the pipeline, the process of co-designing "checklists" with clinical practitioners formalizes ad-hoc procedures and empowers individual advocates (139). Standardized reporting of model performance—such as the one-page summary "model cards" for model reporting (140)—can empower clinical practitioners to understand model limitations and future model developers to identify areas of improvement. Similarly, better documentation of the data supporting initial model training can help expose sources of discrimination in the collected data. Modelers could use "datasheets" for datasets to detail the conditions of data collection (141).

6.4. Regulation

In the United States, the Food and Drug Administration (FDA) has responsibility for the regulation of health care ML models. As there does not exist comprehensive guidance for health care model research and subsequent deployment, the opportunity is ripe to create a comprehensive framework to audit and regulate models. Currently, the FDA's proposed ML-specific modifications to the software as a medical device (SaMD) regulations draw a distinction between models that are trained and then frozen prior to clinical deployment

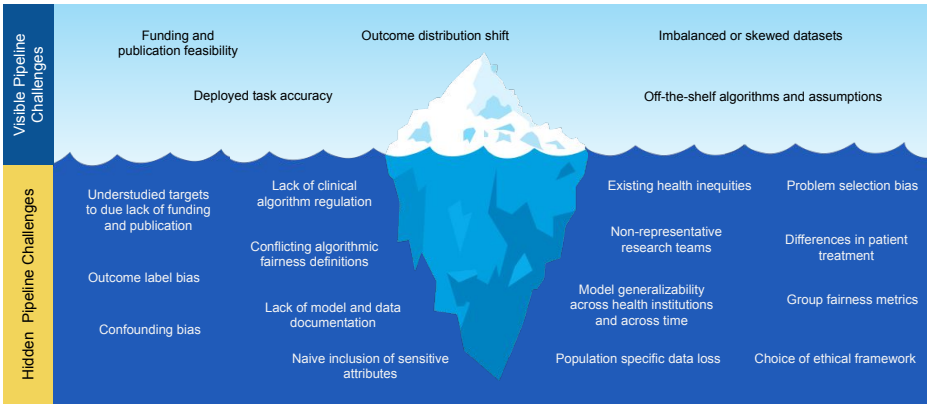


Figure 2

The model development pipeline contains many challenges for ethical machine learning for health care. We highlight both visible and hidden challenges.

and models that continue to learn on observed outcomes. Although models in the latter class can leverage larger, updated datasets, they also face additional risk due to model drift and may need additional audits (142). Such frameworks should explicitly account for health disparities across the stages of ML development in health, and ensure health equity audits as part of postmarket evaluation (143). We also note that there are many potential legal implications, e.g., in malpractice and liability suits, that will require new solutions (144).

Researchers have proposed additional frameworks to guide clinical model development, which could inspire future regulation. ML model regulation could draw from existing regulatory frameworks: a randomized controlled trial for ML models would assess patient benefit compared to a control cohort of standard clinical practice (145), and a drug development pipeline for ML models would define a protocol for adverse events and model recalls (146). The clinical interventions accompanying the clinical ML model should be analyzed to contextualize the use of the model in the clinical setting (147).

7. RECOMMENDATIONS

In this review, we have described the ethical considerations at each step of the ML model development pipeline we introduced. While most researchers will address known challenges like deployed task accuracy and outcome distribution shift, they are unlikely to be aware of the full magnitude of the hidden challenges such as existing health inequities or outcome label bias. As seen in **Figure 2**, many hidden pipeline challenges can go unaddressed in a typical ML health project, but have serious ethical repercussions. With these challenges in mind, we propose five general recommendations that span the pipeline stages.

1. Problems should be tackled by diverse teams and using frameworks that increase the probability that equity will be achieved. Further, historically understudied problems are important targets to practitioners looking to perform high-impact work.
2. Data collection should be framed as an important front-of-mind concern in the ML modelling pipeline including clear disclosures about imbalanced datasets, and researchers should engage with domain experts to ensure that data reflecting under-

served and understudied populations' needs are gathered.
3. Outcome choice should reflect the task at hand and should preferably be unbiased. If the outcome label has ethical bias, the source of inequity should be accounted for in ML model design, leveraging literature that attempts to remove ethical biases during pre-processing, or with use of a reasonable proxy.
4. Reflection on the goals of the model is essential during development, and should be articulated in a pre-analysis plan. In addition to technical choices like loss function, researchers must interrogate how, and whether, a model should be developed to best answer a research question, and what caveats are included.
5. Audits should be designed to identify specific harms, and paired with methods and procedures. Harms should be examined group-by-group, rather than at a population level. ML ethical design "checklists" are one possible tool to systematically enumerate and consider such ethical concerns prior to declaring success in a project.

Finally, we note that machine learning also could and should be harnessed to create shifts in power in health care systems (148). This might mean actively selecting problems for the benefit of underserved patients, designing methods to target systemic interventions for improved access to care and treatments, or enforcing evaluations with the explicit purpose of preserving patient autonomy. In one salient example, the state of California reduced disparities in rates of obstetric hemorrhage (and therefore maternal mortality for women of color) by weighing blood loss sponges, i.e., making access to treatment consistent and unbiased for all women (149). Models could similarly be harnessed to learn and recommend consistent rules, giving researchers a potential opportunity to de-bias current clinical care (150), measure racial disparities and mistrust in end-of-life care (82), improve known biases against minorities (127) and disadvantaged majorities (128). Ultimately, the responsibility for ethical models and behavior lies with a broad community, but begins with technical researchers fulfilling an obligation to engage with patients, clinical researchers, staff, and advocates to build ethical models.

FUTURE QUESTIONS

1. How can we combat urgent global health crises that exacerbate existing patterns of health injustices?
2. How can we encourage model developers to build ethical considerations into the pipeline from the very beginning? Currently, when egregious cases of injustice are discovered only after clinical impact has already occurred, what can developers do to engage?
3. How can evaluation and audits of ML systems be translated into meaningful clinical practice when, in many countries, clinicians themselves are subject to only limited external evaluations or audits?
4. When, if ever, should sensitive attributes like race be used in analysis? How should we incorporate socially constructed features into models and audits?
5. How can ML be used to shift power, e.g., from well-known institutions, privileged patients, and wealthy multinational corporations to the patients most in need?

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Rediet Abebe for helpful discussions and contributions to an early draft, and Peter Szolovits, Pang Wei Koh, Leah Pierson, Berk Ustun, and Tristan Naumann for useful comments and feedback. This work was supported in part by an NIH Director's New Innovator Award DP2MD012722 (SR), a CIFAR AI Chair at the Vector Institute (MG), and Microsoft Research (MG).

LITERATURE CITED

1. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.
2. K. Ferryman and R. A. Winn, "Artificial intelligence can entrench disparitiesheres what we must do," *The Cancer Letter*, Nov 2016.
3. J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, *et al.*, "Do no harm: a roadmap for responsible machine learning for health care," *Nature medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.
4. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020.
5. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "Practical guidance on artificial intelligence for health-care data," *The Lancet Digital Health*, vol. 1, no. 4, pp. e157–e159, 2019.
6. I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can ai help reduce disparities in general medical and mental health care?," *AMA journal of ethics*, vol. 21, no. 2, pp. 167–179, 2019.
7. H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 110–120, 2020.
8. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
9. D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.
10. C. M. Dalton, L. Taylor, and J. Thatcher, "Critical data studies: A dialog on data and space," *Big Data & Society*, vol. 3, no. 1, p. 2053951716648346, 2016.
11. I. Zliobaite, "A Survey on Measuring Indirect Discrimination in Machine Learning," *arXiv preprint arXiv:1511.00148*, 2015.
12. S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.
13. S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.
14. I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?," in *Advances in Neural Information Processing Systems*, pp. 3539–3550, 2018.
15. A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in

machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, pp. 866–872, 2018.
16. B. Ustun, Y. Liu, and D. Parkes, "Fairness without harm: Decoupled classifiers with preference guarantees," in *International Conference on Machine Learning*, pp. 6373–6382, 2019.
17. R. Benjamin, "Assessing risk, automating racism," *Science*, vol. 366, no. 6464, pp. 421–422, 2019.
18. R. M. Veatch and L. K. Guidry-Grimes, *The basics of bioethics*. Routledge, 2019.
19. E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLoS medicine*, vol. 15, no. 11, p. e1002689, 2018.
20. J. Kaye, "The tension between data sharing and the protection of privacy in genomics research," *Annual review of genomics and human genetics*, vol. 13, pp. 415–431, 2012.
21. M. Powers, R. R. Faden, R. R. Faden, *et al.*, *Social justice: the moral foundations of public health and health policy*. Oxford University Press, USA, 2006.
22. C. J. Berg, H. K. Atrash, L. M. Koonin, and M. Tucker, "Pregnancy-related mortality in the united states, 1987–1990," *Obstetrics & Gynecology*, vol. 88, no. 2, pp. 161–167, 1996.
23. D. E. Roberts, *Killing the black body: Race, reproduction, and the meaning of liberty*. Vintage, 1999.
24. D. R. Berry, *The price for their pound of flesh: The value of the enslaved, from womb to grave, in the building of a nation*. Beacon Press, 2017.
25. N. Fisk and R. Atun, "Systematic analysis of research underfunding in maternal and perinatal health," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 116, no. 3, pp. 347–356, 2009.
26. E. A. Howell, N. Egorova, A. Balbierz, J. Zeitlin, and P. L. Hebert, "Black-white differences in severe maternal morbidity and site of care," *American journal of obstetrics and gynecology*, vol. 214, no. 1, pp. 122–e1, 2016.
27. A. A. Creanga, B. T. Bateman, J. M. Mhyre, E. Kuklina, A. Shilkrut, and W. M. Callaghan, "Performance of racial and ethnic minority-serving hospitals on delivery-related indicators," *American journal of obstetrics and gynecology*, vol. 211, no. 6, pp. 647–e1, 2014.
28. H. M. Eltoukhi, M. N. Modi, M. Weston, A. Y. Armstrong, and E. A. Stewart, "The health disparities of uterine fibroid tumors for african american women: a public health issue," *American journal of obstetrics and gynecology*, vol. 210, no. 3, pp. 194–199, 2014.
29. K. M. Hoffman, S. Trawalter, J. R. Axt, and M. N. Oliver, "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites," *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4296–4301, 2016.
30. A. A. Creanga, C. J. Berg, J. Y. Ko, S. L. Farr, V. T. Tong, F. C. Bruce, and W. M. Callaghan, "Maternal mortality and morbidity in the united states: where are we now?," *Journal of women's health*, vol. 23, no. 1, pp. 3–9, 2014.
31. D. Vidyasagar, "Global notes: the 10/90 gap disparities in global health research," *Journal of Perinatology*, vol. 26, no. 1, pp. 55–56, 2006.
32. L. Pierson and J. Millum, "Grant reviews and health research priority setting: Do research funders uphold widely endorsed ethical principles?," *Working Paper*, 2020.
33. P. Von Philipsborn, F. Steinbeis, M. E. Bender, S. Regmi, and P. Tinnemann, "Poverty-related and neglected diseases–an economic and epidemiological analysis of poverty relatedness and neglect in research and development," *Global Health Action*, vol. 8, no. 1, p. 25818, 2015.
34. A. of Us Research Program Investigators, "The "all of us" research program," *New England Journal of Medicine*, vol. 381, no. 7, pp. 668–676, 2019.
35. a. under, "23andme's call for collaborations to study underrepresented populations," Mar 2019.
36. K. S. Tsosie, J. M. Yracheta, and D. Dickenson, "Overvaluing individual consent ignores risks to tribal participants," *Nature Reviews Genetics*, vol. 20, no. 9, pp. 497–498, 2019.

37. F. Farooq and J. J. Strouse, "Disparities in foundation and federal support and development of new therapeutics for sickle cell disease and cystic fibrosis," *Blood*, vol. 132, pp. 4687–4687, 2018.
38. M. Park, "Ncaa genetic screening rule sparks discrimination concerns," Aug 2010.
39. C. Rouse, *Uncertain suffering: racial health care disparities and sickle cell disease*. Univ of California Press, 2009.
40. S. Chakradhar, "Discovery cycle," *Nature Medicine*, vol. 24, no. 8, pp. 1082–1086, 2018.
41. V. Eisenberg, C. Weil, G. Chodick, and V. Shalev, "Epidemiology of endometriosis: a large population-based database study from a healthcare provider with 2 million members," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 125, no. 1, pp. 55–62, 2018.
42. E. Pierson, T. Althoff, D. Thomas, P. Hillard, and J. Leskovec, "The menstrual cycle is a primary contributor to cyclic variation in womens mood, behavior, and vital signs," *Working Paper*, 2019.
43. P. J. A. Hillard, "Menstruation in Adolescents: What Do We Know? and What Do We Do with the Information?," *Journal of Pediatric and Adolescent Gynecology*, vol. 27, no. 6, pp. 309–319, 2014.
44. American Academy of Pediatrics, Committee on Adolescence, American College of Obstetricians and Gynecologists, Committee on Adolescent Health Care, "Menstruation in girls and adolescents: using the menstrual cycle as a vital sign," *Pediatrics*, vol. 118, no. 5, p. 2245, 2006.
45. "Nih offers its first research project grant (r01) on sex and gender."
46. M. Kasy and R. Abebe, "Fairness, equality, and power in algorithmic decision making," tech. rep., Working paper, 2020.
47. B. Hofstra, V. V. Kulkarni, S. M.-N. Galvez, B. He, D. Jurafsky, and D. A. McFarland, "The diversity–innovation paradox in science," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 9284–9291, 2020.
48. J. D. West, J. Jacquet, M. M. King, S. J. Correll, and C. T. Bergstrom, "The role of gender in scholarly authorship," *PloS one*, vol. 8, no. 7, p. e66212, 2013.
49. E. Pierson, "Demographics and discussion influence views on algorithmic fairness," *arXiv preprint arXiv:1712.09124*, 2017.
50. T. A. Hoppe, A. Litovitz, K. A. Willis, R. A. Meseroll, M. J. Perkins, B. I. Hutchins, A. F. Davis, M. S. Lauer, H. A. Valantine, J. M. Anderson, *et al.*, "Topic choice contributes to the lower rate of nih awards to african-american/black scientists," *Science advances*, vol. 5, no. 10, p. eaaw7238, 2019.
51. D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, and R. Kington, "Race, ethnicity, and nih research awards," *Science*, vol. 333, no. 6045, pp. 1015–1019, 2011.
52. "Proposed standards for race-based and indigenous identity data," July 2020.
53. M. d. N. Léonard, "Census and racial categorization in france: Invisible categories and color-blind politics," *Humanity & society*, vol. 38, no. 1, pp. 67–88, 2014.
54. N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.
55. M. B. A. McDermott, B. Nestor, E. Kim, W. Zhang, A. Goldenberg, P. Szolovits, and M. Ghassemi, "A comprehensive evaluation of multi-task learning and multi-task pre-training on ehr time-series data," 2020.
56. L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, 2020.
57. P. M. Rothwell, "External validity of randomised controlled trials:to whom do the results of this trial apply?," *The Lancet*, vol. 365, no. 9453, pp. 82–93, 2005.
58. K. Courtright, "Point: Do randomized controlled trials ignore needed patient populations?

yes," *Chest*, vol. 149, no. 5, pp. 1128–1130, 2016.
59. J. Travers, S. Marsh, M. Williams, M. Weatherall, B. Caldwell, P. Shirtcliffe, S. Aldington, and R. Beasley, "External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?," *Thorax*, vol. 62, no. 3, pp. 219–223, 2007.
60. E. A. Stuart, C. P. Bradshaw, and P. J. Leaf, "Assessing the generalizability of randomized trial results to target populations," *Prevention Science*, vol. 16, no. 3, pp. 475–485, 2015.
61. B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *Egems*, vol. 1, no. 3, 2013.
62. D. Agniel, I. S. Kohane, and G. M. Weber, "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study," *Bmj*, vol. 361, 2018.
63. V. L. Bartlett, S. S. Dhruva, N. D. Shah, P. Ryan, and J. S. Ross, "Feasibility of using real-world data to replicate clinical trial evidence," *JAMA network open*, vol. 2, no. 10, pp. e1912869–e1912869, 2019.
64. K. Ferryman and M. Pitcan, "Fairness in precision medicine," *Data & Society*, 2018.
65. E. Hing and C. W. Burt, "Are there patient disparities when electronic health records are adopted?," *Journal of health care for the poor and underserved*, vol. 20, no. 2, pp. 473–488, 2009.
66. M. Kapoor, D. Agrawal, S. Ravi, A. Roy, S. Subramanian, and R. Guleria, "Missing female patients: an observational analysis of sex ratio among outpatients in a referral tertiary care public hospital in india," *BMJ open*, vol. 9, no. 8, p. e026850, 2019.
67. S. J. A. Haneuse and S. M. Shortreed, *On the use of electronic health records*. Chapman and Hall/CRC New York, NY, 2017.
68. C. Wing, K. Simon, and R. A. Bello-Gomez, "Designing difference in difference studies: best practices for public health policy research," *Annual review of public health*, vol. 39, 2018.
69. E. J. Callahan, S. Hazarian, M. Yarborough, and J. P. Sánchez, "Eliminating lgbtiqq health disparities: the associated roles of electronic health records and institutional culture," *Hastings Center Report*, vol. 44, no. s4, pp. S48–S52, 2014.
70. M. M. López, M. Bevans, L. Wehrlen, L. Yang, and G. Wallen, "Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities," *Journal of racial and ethnic health disparities*, vol. 4, no. 5, pp. 812–818, 2017.
71. E. V. Klinger, S. V. Carlini, I. Gonzalez, S. S. Hubert, J. A. Linder, N. A. Rigotti, E. Z. Kontos, E. R. Park, L. X. Marinacci, and J. S. Haas, "Accuracy of race, ethnicity, and language preference in an electronic health record," *Journal of general internal medicine*, vol. 30, no. 6, pp. 719–723, 2015.
72. M. Dredze, "How social media will change public health," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 81–84, 2012.
73. R. Abebe, S. Hill, J. W. Vaughan, P. M. Small, and H. A. Schwartz, "Using search queries to understand health information needs in africa," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 3–14, 2019.
74. S. Giorgi, D. Preoţiuc-Pietro, A. Buffone, D. Rieman, L. Ungar, and H. A. Schwartz, "The remarkable benefit of user-level aggregation for lexical-based population-level predictions," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1167–1172, 2018.
75. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Clinical use of current polygenic risk scores may exacerbate health disparities," *Nature genetics*, vol. 51, no. 4, pp. 584–591, 2019.
76. D. T. Jamison *et al.*, *Disease and mortality in sub-Saharan Africa*. World Bank Publications, 2006.
77. S. James, J. Herman, S. Rankin, M. Keisling, L. Mottet, and M. Anafi, "The report of the 2015 us transgender survey," 2016.

78. C. Fountain and P. Bearman, "Risk as social context: immigration policy and autism in california," in *Sociological Forum*, vol. 26, pp. 215–240, Wiley Online Library, 2011.
79. Y. C. Ai-ris and R. L. Molina, "Maternal mortality in the united states: updates on trends, causes, and solutions," *Neoreviews*, vol. 20, no. 10, pp. e561–e574, 2019.
80. C. Tiwari, K. Beyer, and G. Rushton, "The impact of data suppression on local mortality rates: the case of cdc wonder," *American journal of public health*, vol. 104, no. 8, pp. 1386–1388, 2014.
81. A. S. Kesselheim and T. A. Brennan, "Overbilling vs. downcodingthe battle between physicians and insurers," *New England Journal of Medicine*, vol. 352, no. 9, pp. 855–857, 2005.
82. W. Boag, H. Suresh, L. A. Celi, P. Szolovits, and M. Ghassemi, "Racial disparities and mistrust in end-of-life care," *arXiv preprint arXiv:1808.03827*, 2018.
83. J. G. Canto, R. J. Goldberg, M. M. Hand, R. O. Bonow, G. Sopko, C. J. Pepine, and T. Long, "Symptom presentation of women with acute coronary syndromes: myth vs reality," *Archives of internal medicine*, vol. 167, no. 22, pp. 2405–2413, 2007.
84. R. Bugiardini, B. Ricci, E. Cenko, Z. Vasiljevic, S. Kedev, G. Davidovic, M. Zdravkovic, D. Miličić, M. Dilic, O. Manfrini, *et al.*, "Delayed care and mortality among women and men with myocardial infarction," *Journal of the American Heart Association*, vol. 6, no. 8, p. e005968, 2017.
85. S. Rose, "A machine learning framework for plan payment risk adjustment," *Health services research*, vol. 51, no. 6, pp. 2358–2374, 2016.
86. M. Geruso and T. Layton, "Upcoding: Evidence from medicare on squishy risk adjustment," tech. rep., National Bureau of Economic Research, 2015.
87. N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in neural information processing systems*, pp. 1196–1204, 2013.
88. Y. Halpern, S. Horng, Y. Choi, and D. Sontag, "Electronic medical record phenotyping using the anchor and learn framework," *Journal of the American Medical Informatics Association*, vol. 23, no. 4, pp. 731–740, 2016.
89. L. Oakden-Rayner, "Exploring large-scale public medical image datasets," *Academic Radiology*, vol. 27, no. 1, pp. 106–112, 2020.
90. S. Tamang, A. Milstein, H. T. Sørensen, L. Pedersen, L. Mackey, J.-R. Betterton, L. Janson, and N. Shah, "Predicting patient cost blooms in denmark: a longitudinal population-based study," *BMJ open*, vol. 7, no. 1, p. e011580, 2017.
91. B. L. Cook, T. G. McGuire, and A. M. Zaslavsky, "Measuring racial/ethnic disparities in health care: methods and practical issues," *Health services research*, vol. 47, no. 3pt2, pp. 1232–1254, 2012.
92. B. L. Cook, S. H. Zuvekas, N. Carson, G. F. Wayne, A. Vesper, and T. G. McGuire, "Assessing racial/ethnic disparities in treatment across episodes of mental health care," *Health services research*, vol. 49, no. 1, pp. 206–229, 2014.
93. A. Zink and S. Rose, "Fair regression for health care spending," *Biometrics*, vol. 76, pp. 973–982, 2020.
94. D. Guillory, "Combating anti-blackness in the ai community," *arXiv preprint arXiv:2006.16879*, 2020.
95. M. Lohaus, M. Perrot, and U. von Luxburg, "Too relaxed to be fair," in *International Conference on Machine Learning*, 2020.
96. S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *ICLR 2020*, 2019.
97. S. Joshi, O. Koyejo, B. Kim, and J. Ghosh, "xgems: Generating examplars to explain black-box models," *arXiv preprint arXiv:1806.08867*, 2018.
98. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of*

the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1721–1730, 2015.
99. M. A. Hernán and J. M. Robins, "Causal inference," 2010.
100. C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.
101. M. J. Van der Laan and S. Rose, *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.
102. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.
103. W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, vol. 105, no. 4, pp. 987–993, 2018.
104. A. Franks, A. DAmour, and A. Feller, "Flexible sensitivity analysis for observational studies without observable implications," *Journal of the American Statistical Association*, pp. 1–33, 2019.
105. M. A. Little and R. Badawy, "Causal bootstrapping," *arXiv preprint arXiv:1910.09648*, 2019.
106. D. A. Vyas, L. G. Eisenstein, and D. S. Jones, "Hidden in plain sightreconsidering the use of race correction in clinical algorithms," 2020.
107. W. A. Grobman, Y. Lai, M. B. Landon, C. Y. Spong, K. J. Leveno, D. J. Rouse, M. W. Varner, A. H. Moawad, S. N. Caritis, M. Harper, *et al.*, "Development of a nomogram for prediction of vaginal birth after cesarean delivery," *Obstetrics & Gynecology*, vol. 109, no. 4, pp. 806–812, 2007.
108. B. Thompson, "Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial," 1995.
109. F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
110. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.
111. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," *ICML*, 2020.
112. S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," *ICML*, 2020.
113. P. A. Flach, "The geometry of roc space: understanding machine learning metrics through roc isometrics," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 194–201, 2003.
114. D. A. Vyas, L. G. Eisenstein, and D. S. Jones, "Hidden in plain sight reconsidering the use of race correction in clinical algorithms," *New England Journal of Medicine*, vol. 0, no. 0, p. null, 0.
115. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.
116. C. Dwork and C. Ilvento, "Group fairness under composition," in *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*, 2018.
117. A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Communications of the ACM*, vol. 63, no. 5, pp. 82–89, 2020.
118. T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *2013 IEEE 13th international conference on data mining*, pp. 71–80, IEEE, 2013.
119. M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, pp. 962–970, 2017.
120. E. Montz, T. Layton, A. B. Busch, R. P. Ellis, S. Rose, and T. G. McGuire, "Risk-adjustment simulation: plans may have incentives to distort mental health and substance use coverage,"

Health Affairs, vol. 35, no. 6, pp. 1022–1028, 2016.
121. T. G. McGuire, A. L. Zink, and S. Rose, "Simplifying and improving the performance of risk adjustment systems," tech. rep., National Bureau of Economic Research, 2020.
122. R. L. Helmreich, "On error management: lessons from aviation," *Bmj*, vol. 320, no. 7237, pp. 781–785, 2000.
123. K. M. Murayama, A. M. Derossis, D. A. DaRosa, H. B. Sherman, and J. P. Fryer, "A critical evaluation of the morbidity and mortality conference," *The American journal of surgery*, vol. 183, no. 3, pp. 246–250, 2002.
124. D. Herrera-Perez, A. Haslam, T. Crain, J. Gill, C. Livingston, V. Kaestner, M. Hayes, D. Morgan, A. S. Cifu, and V. Prasad, "Meta-research: A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals," *Elife*, vol. 8, p. e45183, 2019.
125. E. Creager, D. Madras, T. Pitassi, and R. Zemel, "Causal modeling for fairness in dynamical systems," *arXiv preprint arXiv:1909.09141*, 2019.
126. P. A. Noseworthy, Z. I. Attia, L. C. Brewer, S. N. Hayes, X. Yao, S. Kapa, P. A. Friedman, and F. Lopez-Jimenez, "Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis," *Circulation: Arrhythmia and Electrophysiology*, vol. 13, no. 3, p. e007988, 2020.
127. U. Treatment, "Confronting racial and ethnic disparities in health care," *Washington, DC, Institute of Medicine*, 2002.
128. C. C. Perez, *Invisible women: Exposing data bias in a world designed for men*. Random House, 2019.
129. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.
130. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12592–12594, 2020.
131. L. Seyyed-Kalantari, G. Liu, M. McDermott, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," *arXiv preprint arXiv:2003.00827*, 2020.
132. B. Nestor, M. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi, "Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks," *Machine Learning for Healthcare*, 2019.
133. A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(de) constructing bias on skin lesion datasets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.
134. R. V. Kundu and S. Patterson, "Dermatologic conditions in skin of color: part i. special considerations for common skin disorders," *American family physician*, vol. 87, no. 12, pp. 850–856, 2013.
135. S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, eds.), pp. 1396–1408, Curran Associates, Inc., 2019.
136. A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health ai.," *Biostatistics (Oxford, England)*, vol. 21, no. 2, p. 345, 2020.
137. S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, "Calibration drift in regression and machine learning models for acute kidney injury," *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1052–1061, 2017.
138. S. Saleh, W. Boag, L. Erdman, and T. Naumann, "Clinical collabsheets: 53 questions to guide a clinical collaboration,"
139. M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-designing checklists to

understand organizational challenges and opportunities around fairness in ai," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.
140. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.
141. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.
142. C. for Devices and R. Health, "Artificial intelligence and machine learning in software," Jan 2020.
143. K. Ferryman, "Addressing health disparities in the fdas ai and machine learning regulatory framework.," *Journal of the American Medical Informatics Association*, To Appear.
144. H. R. Sullivan and S. J. Schweikart, "Are current tort liability doctrines adequate for addressing injury caused by ai?," *AMA journal of ethics*, vol. 21, no. 2, pp. 160–166, 2019.
145. X. Liu, S. C. Rivera, L. Faes, L. F. Di Ruffano, C. Yau, P. A. Keane10, H. Ashrafian11, A. Darzi11, S. J. Vollmer, and J. Deeks, "Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed," *Nat. Med*, vol. 25, pp. 1467–1468, 2019.
146. A. Coravos, I. Chen, A. Gordhandas, and A. D. Stern, "We should treat algorithms like prescription drugs," *Quartz*, 2019.
147. R. B. Parikh, Z. Obermeyer, and A. S. Navathe, "Regulation of predictive analytics in medicine," *Science*, vol. 363, no. 6429, pp. 810–812, 2019.
148. S. Mohamed, M.-T. Png, and W. Isaac, "Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence," *Philosophy & Technology*, pp. 1–26, 2020.
149. A. Lyndon, S. Francisco, J. Mcnulty, *et al.*, "Cumulative quantitative assessment of blood loss," *CMQCC Obstetric Hemorrhage Toolkit Version*, vol. 2, pp. 80–85, 2015.
150. I. Y. Chen, S. Joshi, and M. Ghassemi, "Treating health disparities with artificial intelligence," *Nature Medicine*, vol. 26, no. 1, pp. 16–17, 2020.