HHS Public Access

Author manuscript

Annu Rev Biomed Data Sci. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Annu Rev Biomed Data Sci. 2021 July ; 4: 123–144. doi:10.1146/annurev-biodatasci-092820-114757.

Ethical Machine Learning in Healthcare

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

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;

[2]Microsoft Research, Cambridge, Massachusetts 02143, USA

[3]Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, Stanford, California 94305, USA

[4]Vector Institute, Toronto, Ontario M5G 1M1, Canada

[5]Department of Technology, Culture, and Society, Tandon School of Engineering, New York University, Brooklyn, New York 11201, USA

[6]Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Abstract

The use of machine learning (ML) in healthcare raises numerous ethical concerns, especially as models can amplify existing health inequities. Here, we outline ethical considerations for equitable ML in the advancement of healthcare. Specifically, we frame ethics of ML in healthcare 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 postdeployment considerations. We close by summarizing recommendations to address these challenges.

Keywords

machine learning; bias; ethics; health; healthcare; health disparities

1. INTRODUCTION

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 (2–5). 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

iychen@mit.edu .

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.

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, healthcare models designed to optimize referrals to long-term care management programs for millions of patients have been found to exclude Black patients with similar health conditions compared to white patients from care management programs (8).

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

A growing body of literature wrestles with the social implications of ML and technology. Some of this work, termed critical data studies, is from a social science perspective (9, 10), whereas other work leads with a technical and computer science perspective (11–13). 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 (14–16), despite the potential life-or-death impacts of ML models (8, 17).

Algorithmic fairness:

the study of definitions and methods related to the justice of models

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 research and clinical care practices. How these principles could inform the ML model development ethical 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 healthcare between groups often reflect existing societal disparities for disadvantaged groups. We consider other bioethics principles such as beneficence and nonmaleficence, but focus them primarily on groups of patients rather than on individuals.

Bioethics:

the study of ethical issues emerging from advances in biology and medicine

Ethical pipeline:

the model development process and the corresponding ethical considerations

Justice:

the principle that obligates equitably distributed benefits, risks, costs, and resources

Beneficience:

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

Nonmaleficence:

the principle that forbids harm or injury to the patient, either through acts of commission or omission

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. Problem selection: This disparity occurs in part during problem selection because maternal mortality is an understudied problem (25).
2. Data collection: 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. Outcome definition: Once data are collected, the choice of outcome definition can obscure underlying issues, e.g., differences in clinical practice. General model outcome definitions for maternal health complications might overlook conditions specific to Black mothers, e.g., fibroids (28).
4. Algorithm development: 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. Postdeployment considerations: Finally, after a model is trained, postdeployment 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.

Model outcome:

the output of interest for predictive models

Confounding:

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

Risk score:

a calculated number denoting the likelihood of adverse event

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 explaining 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, as well as 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 so-called 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 populations in the Global South. While efforts like the "All Of Us" Research Program (34) and 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 North–South 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, 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, until 2019 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 (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 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 United States 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 healthcare 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, the authors noted that model performance was lower in female patients since they make up 6.38% of patients in the training data (54). Other work has indicated that this issue cannot be simply addressed by pretraining 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 that ameliorates issues of imbalance in downstream tasks, and it suggests that we must solve this problem at the root, be it with more balanced comprehensive data, specialty learning algorithms, or combinations thereof. 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 (16, 56).

Training data:

information that a ML model fits to and learns patterns from

In this section, we discuss common biases in data collection. We consider two types of processes that result in a loss of data. First, we examine processes that affect what kind of information is collected (heterogenous data loss) across varying input types. Examples include clinical trials with aggressive inclusion criteria or social media data that reflect those with access to devices hosting social media apps. Second, we examine processes that affect whether an individual's information is collected due to the individual's population type (population-specific data losses), often across data input categories. For example, undocumented immigrants may fear deportation if they participate in healthcare systems.

Population-specific data loss:

the process whereby data can be lost in collection due to the features of the population

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 records (EHRs), administrative health data, and social media data.

Heterogeneous data loss:

the process whereby data can be lost in collection due to the data type

Data noise:

meaningless information added to data that obscures the underlying information of the data

Missingness:

the manner in which data are absent from a sample of the population

3.1.1. Randomized controlled trials.—RCTs 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.

Randomized controlled trial (RCT):

a study in which subjects are allocated by chance to receive one of several interventions

3.1.2. Electronic health records.—Much recent work in ML also leverages large EHR data. EHR data are a complex reflection of patient health, healthcare 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 healthcare 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 EHRs (63).

Electronic health record (EHR):

digital version of a patient's clinical history that is maintained by the provider over time

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 healthcare access has not been systematically studied in India, primarily due to a lack of reliable data (66).

3.1.3. Administrative health records.—In addition to RCTs and EHRs, healthcare 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 concerning which populations are followed and who is able to participate. Translating such research into practice is a crucial part of maintaining healthcare 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.

Intervention:

a treatment, procedure, or other action taken to prevent or treat disease or improve health in other ways

3.1.4. Social media data.—Data from social media platforms and search-based research naturally consist of only individuals with internet access (72). Even small choices like limiting samples to those from desktop versus mobile platforms constitute 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 to users who made more 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 accurately captured due to errors in data collection and systemic discrimination. Larger genomics datasets often target European populations, producing genetic risk scores that 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 nonconforming individuals, undocumented migrants, and pregnant women.

3.2.1. 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, they are not digitized and often contain 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).

3.2.2. Transgender and gender-nonconforming individuals.—The healthcare needs and experiences of transgender and gender-nonconforming 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-nonconforming individuals. Apart from health documentation concerns, transgender people are often concerned about their basic physical safety when reporting their identities. In the United States, 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).

3.2.3. Undocumented immigrants.—Safety concerns are important in data collection for undocumented migrants, where sociopolitical 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).

3.2.4. Pregnant women.—Despite pregnancy being neither rare nor an illness, the United States continues to experience rising maternal mortality and morbidity rates. In the United States, 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, in order to reduce the likelihood of a breach of confidentiality. For example, the US Centers for Disease Control and Prevention 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 healthcare 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 of 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 due to abnormalities in clinical care: For example, 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.

Diagnosis code:

a label in patient records of disease occurrence, which may be subject to misclassification, used primarily for billing purposes

We discuss social justice considerations in two examples of commonly modeled healthcare outcomes: clinical diagnosis and healthcare 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 result 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

Additionally, it is important to consider the healthcare system in which disease labels are logged. For example, healthcare 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 underreport (81) and overreport (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 indicate either 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 from clinicians that are strongly correlated with positive labels, known also as silver-standard labels (88). Clinical analysis of sources of error in disease labels can also guide improvements (89) and identify affected groups (56).

4.2. Healthcare Costs

Developers of clinical models may choose to predict healthcare costs, meaning the ML model seeks to predict which patients will cost the healthcare provider more in the future. Some model developers may use healthcare 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 needs will limit future costs. Others may explicitly want to understand patients who will have high healthcare costs to reduce the total cost of healthcare (90). However, because socioeconomic factors affect both access to healthcare and access to financial resources, these models may yield predictions that exacerbate inequities.

For model developers seeking to optimize for health needs, healthcare costs can deviate from health needs 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 healthcare costs as a predictive outcome led to racial disparities in

patient allocation to the program (8). Healthcare costs can differ from health needs on an institutional level due to under-insurance 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 healthcare. 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 needs, the disparities are solely a result of differences in healthcare access and systemic discrimination (92).

Addressing issues that arise from the use of healthcare 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 healthcare costs, e.g., the number of chronic diseases as a measure of health needs. If a model developer is most concerned with cost, it is possible to correct for health disparities in predicting healthcare 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

Algorithm development considers the construction of the underlying computation for the ML model and presents a major vulnerability and opportunity for ethical ML in healthcare. Just as data are not neutral, algorithms are also not neutral. A disproportionate amount of power lies with research teams that, after determining the research questions, make decisions about critical components of an algorithm such as the loss function (46, 94). 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 but instead minimize 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).

Loss function:

the relation that determines the error between algorithm output and a given label, which the algorithm uses to optimize

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 the definition of “fairness” itself.

Deployment:

the process through which an ML model is integrated into an existing production environment

Tuning parameters:

algorithm components used for prediction that are tuned toward solving an optimization problem

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).

Sensitive attribute:

a specified patient feature (e.g., race, gender) that is considered important for fairness considerations

Test data:

unseen information that a model predicts on and is evaluated against

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 nonasthmatic 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 in a seemingly straightforward prediction task by altering the data 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 train reliable models because features can have a mediating or moderating effect (posttreatment 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 (99, 103) 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).

5.2. Feature Selection

With large-scale digitization of EHRs 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 naive inclusion of race information could exacerbate disparities in maternal mortality. This ambiguity calls into question race-based correction in scores like VBAC (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 initialization (e.g., start with an empty set of features or a full set of features) and processing steps (e.g., deletion versus 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 fact that the final set depends 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 colinear 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, whereby incorporation of automated procedures is blended with investigator knowledge (111).

Stepwise regression:

a method of estimation whereby each feature is sequentially considered by addition or subtraction to the existing feature set

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.

Generalizability:

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

5.4. Performance Metrics

There are many commonly used performance metrics for model evaluation, such as area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPRC), and calibration (113). However, the appropriate metrics to optimize depend on the 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 a weak true positive rate) but it also does not describe the impact of the model across selected groups. Furthermore, even so-called objective metrics and scores can be deeply flawed and lead to over- or undertreatment of minorities if blindly applied (114). Note that robust reporting of results should include an explicit statement of other nonoptimized metrics, including the original intended use case, the training cohort and case, or the level of model uncertainty.

Performance metric:

score or other quantitative representation of a model’s quality and ability to achieve goals

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

5.5. 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 (93, 118, 119).

Group fairness:

a principle whereby predefined patient groups should receive similar model performance

Recent work has focused on identifying and mitigating violations of fairness definitions in healthcare settings. While most of these algorithms have emerged outside the field of healthcare, 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 have been 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).

6. POSTDEPLOYMENT 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 above that may lead to chronic over- or undertreatment (106). Here we outline considerations for robust deployment by highlighting the need for careful performance reporting and 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 model auditing postdeployment, 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 by specifying groups that must have equal performance (126). Additionally, there is the possibility that models may help to debias current clinical care by reducing known biases against minorities (127) and disadvantaged majorities (128).

Model auditing:

the postdeployment inspection of model performance on groups and outcomes

6.2. Model Generalizability

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 degrade 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).

Data artifact:

a flaw in data caused by equipment, techniques, or conditions that is unrelated to model output

Several algorithms have recently been proposed to account for distribution shifts 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 that of Reference 136, may require a clear indication of which distributions in a healthcare pipeline are expected to change, and may 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 are imperative deployment considerations 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) bears responsibility for the regulation of healthcare ML models. As there does not exist comprehensive guidance for healthcare 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 regulations draw a distinction between models that are trained and then frozen prior to clinical deployment 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: An RCT 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 such as 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 they 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 modeling pipeline, clear disclosures should be made about imbalanced datasets, and researchers should engage with domain experts to ensure that data reflecting the needs of underserved and understudied populations 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 preprocessing, 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 preanalysis 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, as well as what caveats are included.
5. Audits should be designed to identify specific harms and should be 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 ML 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, potentially giving researchers an opportunity to debias current clinical care (150), measure racial disparities and mistrust in end-of-life care (82), and improve known biases against minorities (127) and disadvantaged majorities (128). Ultimately, the responsibility for ethical models and behavior lies with a broad community, but it

begins with technical researchers fulfilling an obligation to engage with patients, clinical researchers, staff, and advocates to build ethical models.

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 (National Institutes of Health) Director's New Innovator Award (DP2MD012722) (to S.R.), a CIFAR (Canadian Institute for Advanced Research) AI Chair at the Vector Institute (to M.G.), and a Microsoft Research grant (to M.G.).

LITERATURE CITED

1. Topol EJ. 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med25:44–56 [PubMed: 30617339]
2. Ferryman K, Winn RA. 2018. Artificial intelligence can entrench disparities—Here's what we must do. The Cancer Letter, 11. 16. https://cancerletter.com/articles/20181116_1/
3. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, et al.2019. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med25:1337–40 [PubMed: 31427808]
4. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. 2020. A review of challenges and opportunities in machine learning for health. AMIA Summits Transl. Sci. Proc2020:191–200 [PubMed: 32477638]
5. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. 2019. Practical guidance on artificial intelligence for health-care data. Lancet Digital Health1:e157–59 [PubMed: 33323184]
6. Chen IY, Szolovits P, Ghassemi M. 2019. Can AI help reduce disparities in general medical and mental health care?AMA J. Ethics21:167–79
7. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 110–20. New York: Assoc. Comput. Mach.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science366:447–53 [PubMed: 31649194]
9. Boyd D, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform. Commun. Soc15:662–79
10. Dalton CM, Taylor L, Thatcher J. 2016. Critical data studies: a dialog on data and space. Big Data Soc. 3(1). 10.1177/2053951716648346
11. Zliobaite I2015. A survey on measuring indirect discrimination in machine learningarXiv:1511.00148 [cs.CY]
12. Barocas S, Hardt M, Narayanan A. 2018. Fairness and machine learning. Online Book, fairmlbook.org. http://www.fairmlbook.org
13. Corbett-Davies S, Goel S. 2018. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv:1808.00023 [cs.CY]
14. Chen I, Johansson FD, Sontag D. 2018. Why is my classifier discriminatory? In Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems (NIPS 2018), ed. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, pp. 3539–50. https://proceedings.neurips.cc/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf
15. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. 2018. Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med169:866–72 [PubMed: 30508424]
16. Ustun B, Liu Y, Parkes D. 2019. Fairness without harm: decoupled classifiers with preference guarantees. Proc. Mach. Learn. Res97:6373–82
17. Benjamin R2019. Assessing risk, automating racism. Science366:421–22 [PubMed: 31649182]
18. Veatch RM, Guidry-Grimes LK. 2019. The Basics of Bioethics. New York: Routledge. 4th ed.
19. Vayena E, Blasimme A, Cohen IG. 2018. Machine learning in medicine: addressing ethical challenges. PLOS Med. 15:e1002689 [PubMed: 30399149]

20. Kaye J2012. The tension between data sharing and the protection of privacy in genomics research. Annu. Rev. Genom. Hum. Genet13:415–31
21. Powers M, Faden R. 2006. Social Justice: The Moral Foundations of Public Health and Health Policy. New York: Oxford Univ. Press
22. Berg CJ, Atrash HK, Koonin LM, Tucker M. 1996. Pregnancy-related mortality in the United States, 1987–1990. Obstet. Gynecol88:161–67 [PubMed: 8692494]
23. Roberts DE. 1999. Killing the Black Body: Race, Reproduction, and the Meaning of Liberty. New York: Vintage Books
24. Berry DR. 2017. The Price for their Pound of Flesh: The Value of the Enslaved, from Womb to Grave, in the Building of a Nation. Boston: Beacon
25. Fisk N, Atun R. 2009. Systematic analysis of research underfunding in maternal and perinatal health. BJOG116:347–56 [PubMed: 19187366]
26. Howell EA, Egorova N, Balbierz A, Zeitlin J, Hebert PL. 2016. Black-white differences in severe maternal morbidity and site of care. Am. J. Obstet. Gynecol214:122.e1–122.e7 [PubMed: 26283457]
27. Creanga AA, Bateman BT, Mhyre JM, Kuklina E, Shilkrut A, Callaghan WM. 2014. Performance of racial and ethnic minority-serving hospitals on delivery-related indicators. Am. J. Obstet. Gynecol211:647.e1–647.e16 [PubMed: 24909341]
28. Eltoukhi HM, Modi MN, Weston M, Armstrong AY, Stewart EA. 2014. The health disparities of uterine fibroid tumors for african american women: a public health issue. Am. J. Obstet. Gynecol210:194–99 [PubMed: 23942040]
29. Hoffman KM, Trawalter S, Axt JR, Oliver MN. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. PNAS113:4296–301 [PubMed: 27044069]
30. Creanga AA, Berg CJ, Ko JY, Farr SL, Tong VT, et al.2014. Maternal mortality and morbidity in the United States: Where are we now?J. Women's Health23:3–9
31. Vidyasagar D2006. Global notes: the 10/90 gap disparities in global health research. J. Perinatol26:55–56 [PubMed: 16281051]
32. Pierson L, Millum J. 2019. Grant reviews and health research priority setting: Do research funders uphold widely endorsed ethical principles?Paper presented at Global Health Bioeth. Conf., Oxford, 1–2 July
33. Von Philipsborn P, Steinbeis F, Bender ME, Regmi S, Tinnemann P. 2015. Poverty-related and neglected diseases–an economic and epidemiological analysis of poverty relatedness and neglect in research and development. Glob. Health Action8:25818 [PubMed: 25623607]
34. All Us Res. Prog. Investig. 2019. The "All of Us" research program. New Engl. J. Med381:668–76 [PubMed: 31412182]
35. 23andme. 2019. 23andme's call for collaborations to study underrepresented populations. 23andme-Blog, Feb. 28.https://blog.23andme.com/23andme-research/23andmes-call-for-collaborations-to-study-underrepresented-populations/
36. Tsosie KS, Yracheta JM, Dickenson D. 2019. Overvaluing individual consent ignores risks to tribal participants. Nat. Rev. Genet20:497–98 [PubMed: 31308520]
37. Farooq F, Strouse JJ. 2018. Disparities in foundation and federal support and development of new therapeutics for sickle cell disease and cystic fibrosis. Blood132:4687–87
38. Park M2010. NCAA genetic screening rule sparks discrimination concerns. CNN, 8. 4. https://www.cnn.com/2010/HEALTH/08/04/ncaa.sickle.genetic.screening/index.html
39. Rouse C2009. Uncertain Suffering: Racial Health Care Disparities and Sickle Cell Disease. Berkeley: Univ. Calif. Press
40. Chakradhar S2018. Discovery cycle. Nat. Med24:1082–86 [PubMed: 30069040]
41. Eisenberg V, Weil C, Chodick G, Shalev V. 2018. Epidemiology of endometriosis: a large population-based database study from a healthcare provider with 2 million members. BJOG125:55–62 [PubMed: 28444957]
42. Pierson E, Althoff T, Thomas D, Hillard P, Leskovec J.2021. Daily, weekly, seasonal and menstrual cycles in women's mood, behaviour and vital signs. Nat. Human BehavIn press

43. Hillard PJA. 2014. Menstruation in adolescents: What do we know? And what do we do with the information?J. Pediatr. Adolesc. Gynecol27:309–19 [PubMed: 25438706]
44. Am. Acad. Pediatr. Comm. Adolesc., Am. Coll. Obstet. Gynecol. Comm. Adolesc. Health Care. 2006. Menstruation in girls and adolescents: using the menstrual cycle as a vital sign. Pediatrics118:2245–50 [PubMed: 17079600]
45. NIH (Natl. Inst. Health). 2020. NIH offers its first research project grant (R01) on sex and gender. In the Spotlight, Oct. 8.https://orwh.od.nih.gov/in-the-spotlight/all-articles/nih-offers-its-first-research-project-grant-r01-sex-and-gender
46. Kasy M, Abebe R. 2020. Fairness, equality, and power in algorithmic decision making. Work. Pap., Oct. 8. https://maxkasy.github.io/home/files/papers/fairness_equality_power.pdf
47. Hofstra B, Kulkarni VV, Galvez SMN, He B, Jurafsky D, McFarland DA. 2020. The diversity–innovation paradox in science. PNAS117:9284–91 [PubMed: 32291335]
48. West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT. 2013. The role of gender in scholarly authorship. PLOS ONE8:e66212 [PubMed: 23894278]
49. Pierson E2017. Demographics and discussion influence views on algorithmic fairness. arXiv:1712.09124 [cs.CY]
50. Hoppe TA, Litovitz A, Willis KA, Meseroll RA, Perkins MJ, et al.2019. Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. Sci. Adv5:eaaw7238 [PubMed: 31633016]
51. Ginther DK, Schaffer WT, Schnell J, Masimore B, Liu F, et al.2011. Race, ethnicity, and NIH research awards. Science333:1015–19 [PubMed: 21852498]
52. CIHI (Can. Inst. Health Info.). 2020. Proposed standards for race-based and indigenous identity data collection and health reporting in Canada. Data Stand., Can. Inst. Health Info., Ottawa, Ont.
53. Léonard MN. 2014. Census and racial categorization in France: invisible categories and color-blind politics. Humanit. Soc38:67–88
54. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, et al.2019. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature572:116–19 [PubMed: 31367026]
55. McDermott MBA, Nestor B, Kim E, Zhang W, Goldenberg A, et al.2020. A comprehensive evaluation of multi-task learning and multi-task pre-training on EHR time-series data. arXiv:2007.10185 [cs.LG]
56. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 151–59. New York: Assoc. Comput. Mach.
57. Rothwell PM. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?"Lancet365:82–93 [PubMed: 15639683]
58. Courtright K2016. Point: Do randomized controlled trials ignore needed patient populations? Yes. Chest149:1128–30 [PubMed: 27157212]
59. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, et al.2007. External validity of randomised controlled trials in asthma: To whom do the results of the trials apply?Thorax62:219–23 [PubMed: 17105779]
60. Stuart EA, Bradshaw CP, Leaf PJ. 2015. Assessing the generalizability of randomized trial results to target populations. Prev. Sci16:475–85 [PubMed: 25307417]
61. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. 2013. Strategies for handling missing data in electronic health record derived data. eGEMS1(3):7
62. Agniel D, Kohane IS, Weber GM. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ361:k1479 [PubMed: 29712648]
63. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. 2019. Feasibility of using real-world data to replicate clinical trial evidence. JAMA Netw. Open2:e1912869 [PubMed: 31596493]
64. Ferryman K, Pitcan M. 2018. Fairness in precision medicine. Res. Proj., Data & Societyhttps://datasociety.net/research/fairness-precision-medicine/

65. Hing E, Burt CW. 2009. Are there patient disparities when electronic health records are adopted?J. Health Care Poor Underserved20:473–88 [PubMed: 19395843]
66. Kapoor M, Agrawal D, Ravi S, Roy A, Subramanian S, Guleria R. 2019. Missing female patients: an observational analysis of sex ratio among outpatients in a referral tertiary care public hospital in India. BMJ Open9:e026850
67. Haneuse SJA, Shortreed SM. 2017. On the use of electronic health records. In Methods in Comparative Effectiveness Research, ed. Gatsonis C, Morton SC, pp. 469–502. New York: Chapman & Hall/CRC
68. Wing C, Simon K, Bello-Gomez RA. 2018. Designing difference in difference studies: best practices for public health policy research. Annu. Rev. Public Health39:453–69 [PubMed: 29328877]
69. Callahan EJ, Hazarian S, Yarborough M, Sánchez JP. 2014. Eliminating LGBTIQQ health disparities: the associated roles of electronic health records and institutional culture. Hastings Center Rep. 44:S48–52
70. López MM, Bevans M, Wehrlen L, Yang L, Wallen G. 2017. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. J. Racial Ethnic Health Disparities4:812–18
71. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, et al.2015. Accuracy of race, ethnicity, and language preference in an electronic health record. J. Gen. Intern. Med30:719–23 [PubMed: 25527336]
72. Dredze M2012. How social media will change public health. IEEE Intell. Syst27:81–84
73. Abebe R, Hill S, Vaughan JW, Small PM, Schwartz HA. 2019. Using search queries to understand health information needs in Africa. In Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, pp. 3–14. Palo Alto, CA: AAAI
74. Giorgi S, Preoţiuc-Pietro D, Buffone A, Rieman D, Ungar L, Schwartz HA. 2018. 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. 1167–72. Stroudsburg, PA: Assoc. Comput. Linguist.
75. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BMDaly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet51:584–91 [PubMed: 30926966]
76. Jamison DT, Feacham RG, Makgoba MW, Bos ER, Baingana FK, et al.2006. Disease and Mortality in Sub-Saharan Africa. Washington, DC: World Bank. 2nd ed.
77. James S, Herman J, Rankin S, Keisling M, Mottet L, Anafi M. 2016. The report of the 2015 US transgender survey. Washington, DC: Natl. Cent. Transgend. Equal.
78. Fountain C, Bearman P. 2011. Risk as social context: immigration policy and autism in California. Sociol. Forum26:215–40
79. Collier AY, Molina RL. 2019. Maternal mortality in the United States: updates on trends, causes, and solutions. NeoReviews20:e561–74 [PubMed: 31575778]
80. Tiwari C, Beyer K, Rushton G. 2014. The impact of data suppression on local mortality rates: the case of CDC WONDER. Am. J. Public Health104:1386–88 [PubMed: 24922161]
81. Kesselheim AS, Brennan TA. 2005. Overbilling versus downcoding—the battle between physicians and insurers. New Engl. J. Med352:855–57 [PubMed: 15745973]
82. Boag W, Suresh H, Celi LA, Szolovits P, Ghassemi M. 2018. Racial disparities and mistrust in end-of-life care. arXiv:1808.03827 [stat.AP]
83. Canto JG, Goldberg RJ, Hand MM, Bonow RO, Sopko G, et al.2007. Symptom presentation of women with acute coronary syndromes: myth versus reality. Arch. Intern. Med167:2405–13 [PubMed: 18071161]
84. Bugiardini R, Ricci B, Cenko E, Vasiljevic Z, Kedev S, et al.2017. Delayed care and mortality among women and men with myocardial infarction. J. Am. Heart Assoc6:e005968 [PubMed: 28862963]
85. Rose S2016. A machine learning framework for plan payment risk adjustment. Health Serv. Res51:2358–74 [PubMed: 26891974]

86. Geruso M, Layton T. 2015. Upcoding: evidence from medicare on squishy risk adjustment. J. Pol. Econ128(3). 10.1086/704756
87. Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. 2013. Learning with noisy labels. In Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NIPS 2013), ed. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, pp. 1196–204. https://papers.nips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf
88. Halpern Y, Horng S, Choi Y, Sontag D. 2016. Electronic medical record phenotyping using the anchor and learn framework. J. Am. Med. Inform. Assoc23:731–40 [PubMed: 27107443]
89. Oakden-Rayner L2020. Exploring large-scale public medical image datasets. Acad. Radiol27:106–12 [PubMed: 31706792]
90. Tamang S, Milstein A, Sørensen HT, Pedersen L, Mackey L, et al.2017. Predicting patient 'cost blooms' in Denmark: a longitudinal population-based study. BMJ Open7:e011580
91. Cook BL, McGuire TG, Zaslavsky AM. 2012. Measuring racial/ethnic disparities in health care: methods and practical issues. Health Serv. Res47:1232–54 [PubMed: 22353147]
92. Cook BL, Zuvekas SH, Carson N, Wayne GF, Vesper A, McGuire TG. 2014. Assessing racial/ethnic disparities in treatment across episodes of mental health care. Health Serv. Res49:206–29 [PubMed: 23855750]
93. Zink A, Rose S. 2020. Fair regression for health care spending. Biometrics76:973–82 [PubMed: 31860120]
94. Guillory D2020. Combating anti-blackness in the AI community. arXiv:2006.16879 [cs.CY]
95. Lohaus M, Perrot M, von Luxburg U. 2020. Too relaxed to be fair. Proc. Mach. Learn. Res119:6360–69
96. Sagawa S, Koh PW, Hashimoto TB, Liang P. 2020. Distributionally robust neural network. Paper presented at the Eighth International Conference on Learning Representations (ICLR 2020), Apr. 26–May 1. https://openreview.net/pdf?id=ryxGuJrFvS
97. Joshi S, Koyejo O, Kim B, Ghosh J. 2018. xGEMS: generating examplars to explain black-box models. arXiv:1806.08867 [cs.LG]
98. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. 2015. 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. 1721–30. New York: Assoc. Comput. Mach.
99. Hernán MA, Robins JM. 2010. Causal Inference: What If. Boca Raton, FL: Chapman & Hall/CRC
100. Glymour C, Zhang K, Spirtes P. 2019. Review of causal discovery methods based on graphical models. Front. Genet10:524 [PubMed: 31214249]
101. Van der Laan MJ, Rose S. 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer-Verlag
102. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al.2018. Double/debiased machine learning for treatment and structural parameters. Econom. J21(1):C1–68
103. Miao W, Geng Z, Tchetgen Tchetgen EJ. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika105:987–93 [PubMed: 33343006]
104. Franks A, D'Amour A, Feller A. 2019. Flexible sensitivity analysis for observational studies without observable implications. J. Am. Stat. Assoc115(532):1730–76
105. Little MA, Badawy R. 2019. Causal bootstrapping. arXiv:1910.09648 [cs.LG]
106. Vyas DA, Eisenstein LG, Jones DS. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med383:874–82 [PubMed: 32853499]
107. Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, et al.2007. Development of a nomogram for prediction of vaginal birth after cesarean delivery. Obstet. Gynecol109:806–12 [PubMed: 17400840]
108. Thompson B1995. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educ. Psychol. Meas55(4):525–34
109. Harrell FE Jr. 2015. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. New York: Springer. 2nd ed.

110. James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statistical Learning: With Applications in R. New York: Springer-Verlag
111. Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, et al.2020. Concept bottleneck models. Proc. Mach. Learn. Res119:5338–48
112. Sagawa S, Raghunathan A, Koh PW, Liang P. 2020. An investigation of why overparameterization exacerbates spurious correlations. Proc. Mach. Learn. Res119:8346–56
113. Flach PA. 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In Proceedings of the 20th International Conference on Machine Learning, ed. Fawcett T, Mishra N, pp. 194–201. Palo Alto, CA: AAAI
114. Vyas DA, Eisenstein LG, Jones DS. 2020.Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. New Engl. J. Med383:874–82 [PubMed: 32853499]
115. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–26. New York: Assoc. Comput. Mach.
116. Dwork C, Ilvento C. 2018. Fairness under composition. arXiv:1806.06122 [cs.LG]
117. Chouldechova A, Roth A. 2020. A snapshot of the frontiers of fairness in machine learning. Commun. ACM63:82–89
118. Calders T, Karim A, Kamiran F, Ali W, Zhang X. 2013. Controlling attribute effect in linear regression. In 2013 IEEE 13th International Conference on Data Mining, pp. 71–80. Los Alamitos, CA: IEEE Comput. Soc.
119. Zafar MB, Valera I, Rogriguez MG, Gummadi KP. 2017. Fairness constraints: mechanisms for fair classification. Proc. Mach. Learn. Res54:962–70
120. Montz E, Layton T, Busch AB, Ellis RP, Rose S, McGuire TG. 2016. Risk-adjustment simulation: Plans may have incentives to distort mental health and substance use coverage. Health Aff. 35:1022–28
121. McGuire TG, Zink AL, Rose S. 2020. Simplifying and improving the performance of risk adjustment systems. Work. Pap., Natl. Bur. Econ. Res., Cambridge, MA
122. Helmreich RL. 2000. On error management: lessons from aviation. BMJ320:781–85 [PubMed: 10720367]
123. Murayama KM, Derossis AM, DaRosa DA, Sherman HB, Fryer JP. 2002. A critical evaluation of the morbidity and mortality conference. Am. J. Surg183:246–50 [PubMed: 11943120]
124. Herrera-Perez D, Haslam A, Crain T, Gill J, Livingston C, et al.2019. Meta-research: a comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. eLife8:e45183 [PubMed: 31182188]
125. Creager E, Madras D, Pitassi T, Zemel R. 2019. Causal modeling for fairness in dynamical systems. arXiv:1909.09141 [cs.LG]
126. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, et al.2020. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circ. Arrhythm. Electrophysiol13:e007988 [PubMed: 32064914]
127. Inst. Med.2002. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: Natl. Acad. Press
128. Perez CC. 2019. Invisible Women: Exposing Data Bias in a World Designed for Men. New York: Abrams
129. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLOS Med. 15:e1002683 [PubMed: 30399157]
130. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. PNAS117:12592–94 [PubMed: 32457147]
131. Seyyed-Kalantari L, Liu G, McDermott M, Ghassemi M. 2020. CheXclusion: fairness gaps in deep chest X-ray classifiers. arXiv:2003.00827 [cs.CV]

132. Nestor B, McDermott M, Boag W, Berner G, Naumann T, et al.2019. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. Proc. Mach. Learn. Res106:381–405

133. Bissoto A, Fornaciali M, Valle E, Avila S. 2019. (De)constructing bias on skin lesion datasets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. https://openaccess.thecvf.com/content_CVPRW_2019/html/ISIC/Bissoto_DeConstructing_Bias_on_Skin_Lesion_Datasets_CVPRW_2019_paper.html

134. Kundu RV, Patterson S. 2013. Dermatologic conditions in skin of color: part I. Special considerations for common skin disorders. Am. Family Phys87:850–56

135. Rabanser S, Günnemann S, Lipton Z. 2019. Failing loudly: an empirical study of methods for detecting dataset shift. In Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems (NIPS 2019), ed. Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, pp. 1396–408. https://papers.nips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf

136. Subbaswamy A, Saria S. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics21:345–52 [PubMed: 31742354]

137. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. 2017. Calibration drift in regression and machine learning models for acute kidney injury. J. Am. Med. Inform. Assoc24:1052–61 [PubMed: 28379439]

138. Saleh S, Boag W, Erdman L, Naumann T. 2020. Clinical collabsheets: 53 questions to guide a clinical collaboration. Proc. Mach. Learn. Res126:783–812

139. Madaio MA, Stark L, Wortman Vaughan J, Wallach H. 2020. 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, Pap. 318.New York: Assoc. Comput. Mach.

140. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, et al.2019. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–29. New Yorl: Assoc. Comput. Mach.

141. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, et al.2018. Datasheets for datasets. arXiv:1803.09010 [cs.DB]

142. FDA (US Food Drug Admin.). 2021. Artificial intelligence and machine learning in software as a medical device. Web Resour., FDA, Silver Spring, MD. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

143. Ferryman K2020. Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. J. Am. Med. Inform. Assoc27(12):2016–19 [PubMed: 32951036]

144. Sullivan HR, Schweikart SJ. 2019. Are current tort liability doctrines adequate for addressing injury caused by AI?AMA J. Ethics21:160–66

145. Liu X, Rivera SC, Faes L, Di Ruffano LF, Yau C, et al.2019. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat. Med25:1467–68 [PubMed: 31551578]

146. Coravos A, Chen I, Gordhandas A, Stern AD. 2019. We should treat algorithms like prescription drugs. Quartz, Feb. 19.https://qz.com/1540594/treating-algorithms-like-prescription-drugs-could-reduce-ai-bias/

147. Parikh RB, Obermeyer Z, Navathe AS. 2019. Regulation of predictive analytics in medicine. Science363:810–12 [PubMed: 30792287]

148. Mohamed S, Png MT, Isaac W. 2020. Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol33:659–84

149. Lyndon A, McNulty J, VanderWal B, Gabel K, Huwe V, Main E. 2015. Cumulative quantitative assessment of blood loss. In CMQCC Obstet. Hemorrhage ToolkitVers. 2, pp. 80–85. Stanford, CA: Calif. Matern. Qual. Care Collab.https://www.cmqcc.org/content/cumulative-quantitative-assessment-blood-loss

150. Chen IY, Joshi S, Ghassemi M. 2020. Treating health disparities with artificial intelligence. Nat. Med26:16–17 [PubMed: 31932779]

FUTURE QUESTIONS

1. How can we combat urgent global health crises that exacerbate existing patterns of health injustices?
2. How can we encourage machine learning (ML) 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 from, e.g., well-known institutions, privileged patients, and wealthy multinational corporations to the patients most in need?

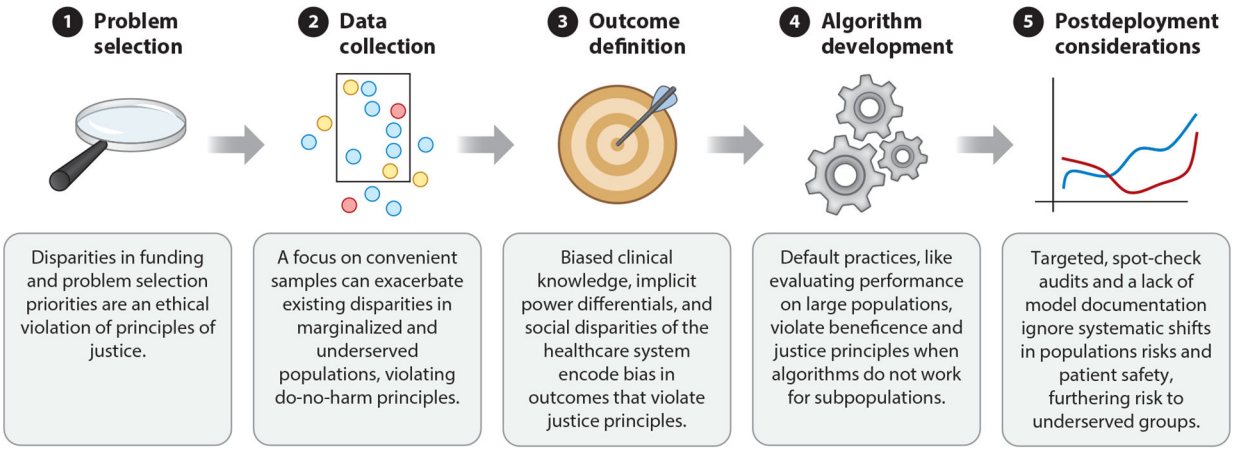

Figure 1.

We motivate the five steps in the ethical pipeline for healthcare model development. Each stage contains considerations for machine learning 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.

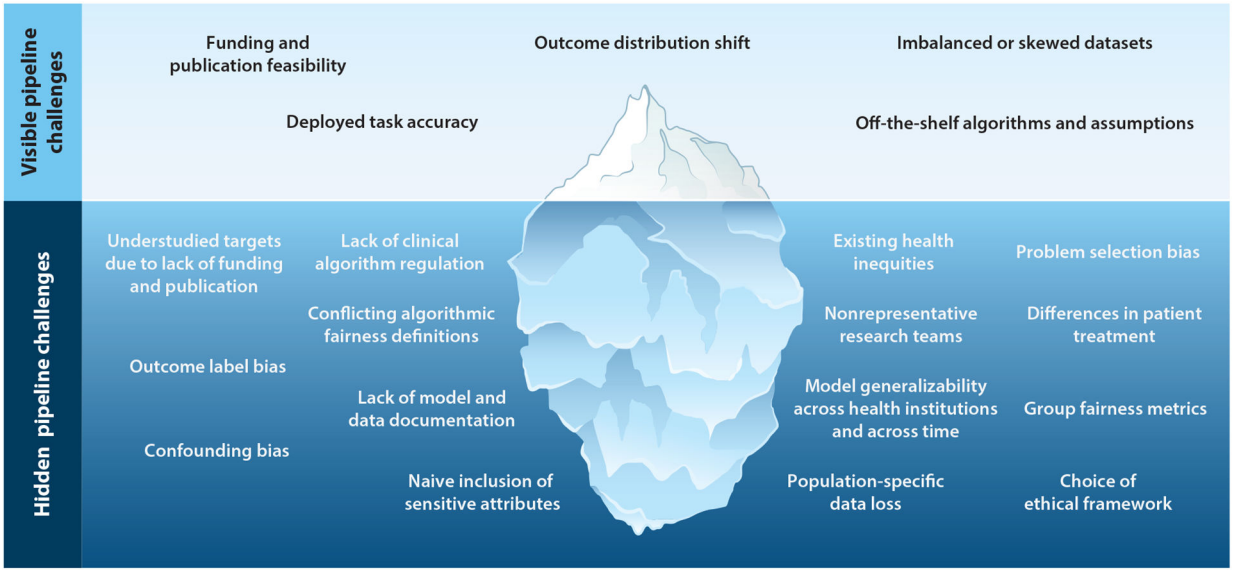

Figure 2.

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