Dataset Bias in Diagnostic AI systems: Guidelines for Dataset Collection and Usage

Julie Vaughn*
Massachusetts Institute of Technology
Cambridge, Massachusetts
juliev@mit.edu

Avital Baral**
Massachusetts Institute of Technology
Cambridge, Massachusetts
abaral@mit.edu

Mayukha Vadari
Massachusetts Institute of Technology
Cambridge, Massachusetts
mvadari@mit.edu

William Boag
Massachusetts Institute of Technology
Cambridge, Massachusetts
wboag@mit.edu

ABSTRACT

In the last few years, the FDA has begun to recognize *de novo* pathways (new approval processes) for approving AI algorithms as medical devices. A major concern is that the review process does not adequately test for bias in these models. There are many ways in which bias can arise in deployed AI models, including during data collection, training, and model deployment. In this paper, we adopt a framework for categorizing medical dataset bias in a fine-grained manner, which enables informed, targeted interventions for each issue appropriately. From there, we propose policy recommendations to the FDA and NIH to promote the deployment of more equitable AI diagnostic systems.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; **Medical information policy**; *Medical technologies.*

KEYWORDS

datasets, bias, policy, healthcare, fairness

ACM Reference Format:
Julie Vaughn, Avital Baral, Mayukha Vadari, and William Boag. 2020. Dataset Bias in Diagnostic AI systems: Guidelines for Dataset Collection and Usage. In *Proceedings of CHIL '20: ACM The ACM Conference on Health, Inference, and Learning (CHIL '20).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In 2016, researchers in Germany built a neural network to identify skin cancer (melanoma) cases based on clinical images [26]. The network was trained on over 100,00 skin images, and was able to detect 95% of melanoma cases in a new set of data accurately, outperforming a panel of 58 certified dermatologists who were collectively 88.9% accurate. This research was published in Annals of Oncology and was heralded as a sign that diagnostic artificial intelligence (AI) should be incorporated into clinical practice. However, there was a problem: more than 95% of the data used to train the model depicted white skin [36]. Given that the model was trained on largely homogeneous data, the algorithm is not likely to generalize to a more diverse population.

This case is not exceptional by any means, as bias is incredibly pervasive in the field of medical diagnostics. For example, a Toronto-based startup built an auditory test for detecting Alzheimer's, but it only worked for fluent English speakers of a speakers of a specific Canadian dialect [22]. In addition, many commonly-used facial recognition systems are up to 20% better at identifying lighter-skinned individuals, and were even worse at identifying darker-skinned women [19]. There is clearly immense need for regulatory measures in the development of AI systems, as these systems become increasingly more integrated into society. Experts predict that artificial intelligence will completely revolutionize the field of diagnostics, allowing doctors to make diagnoses exponentially faster and more accurately [50] [15]. Implemented correctly, AI could make healthcare universally more accessible. However, it could also worsen already deeply-entrenched healthcare disparities. In particular, we may increase disparities observed along demographic lines such as worse outcomes amongst racial and ethnic minorities and low-income people. A recent study has shown that a currently deployed algorithm used to manage the health of populations shows strong racial bias because of poorly-chosen health outcome measures [40]. We are at a turning point in the field of diagnostics - a turning point where we have the chance to create AI systems that are both highly accurate and that combat institutionalized inequality rather than perpetuate it.

There are numerous potential sources of bias within the currently known framework for understanding AI. These sources of bias can be roughly categorized into data collection, data aggregation, and interpretations of model results [15] [49]. In this context, the term "bias" refers to an unintended or potentially harmful property of the data.

The US currently has no legislative framework for determining bias in datasets in general. We hope that the following recommendations to mitigate bias in medical AI can serve as a model for other domains as well (for example, prohibiting hiring discrimination that occurs as a result of hiring/recruiting algorithms). Notably,

*Both authors contributed equally to this research.

CHIL '20, April 02–04, 2020, Toronto, ON

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

medical data is among the most sensitive and well-protected forms of data, and therefore serves as a useful case study in balancing pathways for innovation with legislative oversight.

Our recommendations are as follows:

(1) Design research incentives to diversify medical dataset generation through the NIH and NSF.
(2) Standardize an FDA regulation process (SaMD pathway) to evaluate algorithm robustness to real-world data before deployment.
(3) Create a standard pathway and incentives for hospitals and other organizations to publish anonymized datasets for academic and industrial use.

The rest of this paper is organized as follows. We begin by discussing the current regulation of ML for clinical diagnosis and the need for increased research diversity in Section 2. Next, in Section 3, we review the framework for categorizing bias in medical data generation, and then discuss the relevant stakeholders in Section 4. We present our policy recommendations in Section 5. Finally, we discuss limitations in Section 6 and offer a concluding summary in Section 7.

2 CURRENT POLICY FOR AI DIAGNOSIS

As artificial intelligence matures and becomes more widely-used, researchers and other stakeholders look to AI and machine learning methods specifically to automate and standardize the diagnostics process and take advantage of increasingly large datasets to improve detection and treatment of diseases. There is also a large body of research to suggest that AI could improve the patient-doctor relationship by automating routine medical tasks, thereby allowing doctors more time to focus on being present and empathetic with the patient [50].

The FDA (Food and Drug Administration, which is the federal agency responsible for approving new drugs and medical devices) has already approved over 30 AI-based healthcare algorithms, including several diagnostic algorithms [20] [4]. However, numerous limitations and concerns impact the clinical implementation of these algorithms [51].

The FDA continues to approve increasingly more AI-based programs (referred to as "Software as a Medical Device" (SaMD)) each year. There were two approvals in all of 2017, but by mid-2018 the FDA was approving them as frequently as once or twice a month [38]. These FDA approvals also reflect increasingly less reliance on physicians. In 2018, the FDA permitted the marketing of a SaMD to detect diabetic retinopathy (a diabetes-related eye problem that can cause blindness) without the need for a clinician, called IDx-DR [3].

The device was approved under the De Novo (i.e, novel and unprecedented) approval pathway, and tested with 900 patient retinal images from 10 primary care sites [16]. The FDA has therefore actively demonstrated an interest in determining new regulatory guidelines for artificial intelligence as a medical device. The diversity of patient populations in this dataset was described to some degree: it was shown to be relatively high, compared to other US-based datasets (37% of the population was non-white, and genders were represented approximately equally), with some exceptions of specific groups; namely, American Natives and Asian Americans were less significantly less represented in the dataset, despite the high prevelance of diabetes in these subpopulations [16]. Given that we know that diabetes may affect each group's biomarkers differently [27] – and it is unknown how ethnicity may affect nuances of diabetic retinopathy screening – we believe it may be prudent to include IDx-DR results alongside diabetes blood biomarkers stratified by ethnicity.

Earlier this year, the FDA proposed a set of guidelines for regulating under what conditions an AI-based SaMD need to be re-approved when updated [4]. There are mentions of the need for quality assurance in data that is used to train algorithms, but little emphasis on bias. The former FDA commissioner, Dr. Scott Gottlieb, released a statement this April explaining that old regulatory frameworks for SaMDs are not flexible and appropriate for algorithms that can learn and adapt to real-world data as they are used [5]. As of 2020, to our knowledge, no deployed clinical AI systems update in real-time yet. Perhaps real-time data adaption may be useful for clinical ML one day, but whenever that is, we believe that algorithms should necessarily be trained and tested with a large quantity of diverse data as a certification before deployment.

We encounter a related concern when examining how to effectively test these algorithms: for academia and research in general, patient-level clinical datasets are relatively scarce. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) protects identifiable medical data from public usage, and patients must consent to the use of their data in a research setting, unless the data has been de-identified. A significant amount of AI research is conducted with a small number of accessible datasets. For example, MIMIC-III, a large ICU dataset curated by researchers from MIT's Computer Science and Artificial Intelligence Lab (CSAIL), has already been cited over 900 times since its publication in 2016 [31]. Similarly, the ISIC archive dataset on melanoma detection is used very often in melanoma AI-diagnostic research [30]. The problem most research being done any small number of datasets is that any biases – be it racial, gender, geographic, age-based, etc – which are present would be magnified from the large number of studies performed. For instance, MIMIC-III comes from one hospital in the city of Boston, which is not representative of many regions in many ways. It is therefore very important that we validate algorithms with a variety of data, with the hope being that it is more likely for a given representation-based bias to cancel out in the aggregate of many diverse studies [8].

In general, medicine has historically catered to the needs of demographic groups centered by societal hierarchies, which in European and North American contexts have been white male-bodied people (often those with enough financial and social resources to be able to seek medical care). Known genetics datasets are largely homogeneous and white [35], and racial disparities have persisted in a variety of disease-specific contexts despite activism efforts [6]. It is therefore within the best interests of activists representing these communities to encourage research that pertains specifically to marginalized groups, perhaps through the creation of federal research funding opportunities for research that seeks to create datasets and algorithms pertinent to minority populations.

3 BIAS OVERVIEW

Artificial Intelligence (AI) is a blanket term that refers to software that can aid in human cognition. If a training set is biased, then the model that the AI develops will be overfit to the biased data, and will not be able to handle other data very well, because it has never been exposed to it. Imagine training an AI to detect broken arms, branding it as a "broken bone detector," and then using it to detect a broken rib. Everyone understands that strategy would fail because the patterns that the AI has learned in order to differentiate images of broken arms from uninjured arms will not generalize to distinguishing broken ribs from intact ribs. Not every example of bias is as obvious as that example, but even effects that may seem small to some people can have large impacts both on system performance and on downstream outcomes.

Throughout the field of AI-based diagnostics systems, there are relatively few medical datasets that researchers may gain access to and train algorithms with. Medical datasets are inherently subjected to bias, as we will argue in this section. For example, consider how datasets on HIV/AIDS infection rates in the early years of the U.S. epidemic were necessarily biased by medical institutions seeking to primarily identify white men in urban centers as the main target of the disease, when in reality, the disease disproportionately affected people of color [18]. More recently, studies have shown that several wearable health devices meant to measure heart rate and energy expenditure (such as the Fitbit Surge and Apple Watch) are biased on the basis of skin color [47]. However, it is important to consider exactly what is meant by the term bias, and how different kinds of bias impact our understanding of how to correct for it [49].

Our discussion will focus on the following sources of bias, proposed by Suresh et al. 2019: [49]:

- **representation bias**: when the training data is not representative of the data used in practice.
- **measurement bias**: when the way that the training data is measured or collected introduces inaccuracies in the model.
- **aggregation bias**: when combining data improperly creates results that are inconsistent across different data types.
- **historical bias**: when a dataset accurately reflects real-world, but its usage perpetuates existing societal bias

In Figure 1, we can ascertain how these varieties of bias arise during the overall process of data collection and model implementation. This paper will focus primarily on issues of data generation and collection, though the authors believe there should also be sufficient attention devoted to mitigating bias in model training and evaluation as well.

3.1 Representation Bias

Returning to the skin cancer detection algorithm [26], the model was trained on 95% white skin and is less likely to generalize to darker-skinned populations. To fix this, the researchers could add additional data with a variety of skin tones, and re-train the algorithm on this more diverse dataset. This would hopefully result in much better performance on these cases. This is an example of representation bias in medical data collection: when the training dataset population for a model is not representative of its real-world usage [23]. Representation bias can often arise from the availability heuristic: datasets are hard enough to create and distribute as it is,

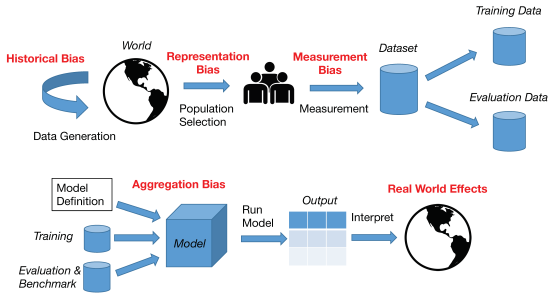

Figure 1: Types of bias within the machine learning pipeline.

especially when generating a more representative cohort would either entail downsampling the existing records to achieve population representation or to spend more money to actively seek out patients that have slipped through the cracks. On the other hand, because datasets are so rare, the value of creating a new one that many researchers will use should incur additional responsibility to take representation bias seriously.

3.2 Measurement Bias

Consider the case of gender and coronary artery disease (CAD) diagnosis. For a long time, the guidelines for diagnosing CAD were based primarily on the symptoms experienced by men. Though women share many of the same traditional risk factors with men, they have some unique differences in pathophysiology that can result in misdiagnosis if the same criteria are applied universally [25]. For instance, frequently, women do not experience chest pain alongside CAD to the same extent as men, and physicians are less likely to diagnose them with CAD. If we build a model based on a dataset that relies heavily on chest pain and other primarily male-bodied criteria as main indicators of CAD, we will have a model that will be woefully inaccurate for female-bodied people. This is a case of measurement bias, because the quality of data for male and female-bodied individuals is different. Another example of measurement bias would be if we used an inaccurate indicator for CAD diagnosis, such as treating a high cholesterol level as the primary indicator of CAD rather than as a feature affecting the diagnostic outcome.

Electronic Health Records (EHR) are often vary in quality within different subpopulations; this difference in quality could be described as measurement bias [10]. This is a well-documented phenomenon [53] [28]. One study found that EHR data is most complete for severely ill patients (with severity of illness measured independently from the EHR data). Specifically, doctors tended to record laboratory results for severely ill pneumonia patients more often than for healthy patients undergoing the same tests [29]. This makes sense, in a hospital context, given the many demands on a physician's time and resources. However, it means that clinical results based on EHR data is necessarily biased to the most ill patients rather than accurately capturing the spectrum of manifestations of a disease [29]. This difference in data quality across groups with

different illness severities is therefore introducing measurement bias into any classifier built on said EHR data.

3.2.1 Adversarial Attacks. In a similar vein, we can also consider the fact that data can be altered in order to essentially "game" the outcome of AI systems; there is a danger that diagnostic AI systems can be tricked in practice if practitioners feed slightly altered data into the system. This alteration can take the form of noise (i.e. small, pixel-level changes) that is imperceptible to the human eye, but takes advantage of the specificity of AI training systems to cause the system to misdiagnose the image [7]. Because the healthcare system is naturally adversarial in some areas (e.g. whether to reimburse an insurance claim, whether an insurer should receive a cost-sharing subsidy for serving a sicker-than-average population, etc), this kind of attack may prove to be especially concerning.

3.3 Aggregation Bias

Although the other forms of bias refer to a characteristic of a dataset, aggregation bias refers to a process: the application of a one-size-fits-all model to a diverse dataset (even if the training data had no other problems). Aggregation bias (also referred to as "Hidden Stratification" of sub-populations) occurs when a large umbrella label is composed of many heterogeneous sub-groups but are nonetheless all modelled as one. An example of this lies in fitting a one-size-fits-all model to study diabetes in a population, even when it is known that certain blood biomarkers for diabetes diagnosis differ vastly across ethnicities [27]. Moving outside typical notions of "bias," another example of hidden stratification was identified in chest X-Ray classification: although a model achieved respectable performance when classifying patients who did vs didn't have pneumothorax (i.e. collapsed lung), the more clinically meaningful task was identifying patients whose collapsed lungs were not yet treated, and on that task, the model performed 10% worse because it optimized for the much more common case of already-treated pneumothorax [39]. The given label of "pneumothorax" was an umbrella term that contained many different sub-groups.

A static dataset doesn't inherently have aggregation bias, but rather the identified harm comes from inappropriate use of the said data. However, there is indeed a characteristic which is intimately related to aggregation bias: heterogeneity. By definition, this bias can only happen on hetergenous populations, where one-size-fits-all solutions cannot work for everyone. Similarly, when a dataset is heterogenous, there is a high likelihood that models fit on the data will have this bias, unless the engineer exercises constant vigilance towards safety and fairness.

3.4 Historical Bias

Unfortunately, it's also possible that a dataset may be representative of the true population and measured correctly, yet it still reflects pre-existing social inequalities. In 2019, researchers discovered that an algorithm being used to decide the care for millions of patients was heavily underestimating the health needs of the sickest black patients [41]. The algorithm was not intended to be biased; it was trained to forecast future spending (as a proxy for patient sickness), but unfortunately there is an existing systemic bias in the healthcare industry, wherein less money is spent on a black patient than a white patient at the same severity of illness. It is imperative that we take into account historical (i.e. real-world) biases when interpreting and building models on medical data, and that we make sure that resources are allocated fairly and responsibly when basing assumptions on algorithmic outputs. Unlike representation bias, this variety of bias is not easily solved with the addition of more, thoughtfully collected data; the problem is that the data reflects a deeper underlying injustice.

There is also an overarching societal bias that we must consider when building AI-based diagnostic tools: the fact that medical research trends in general reflect the needs of those with the most power and influence. Therefore, it follows that there is a scarcity of medical research that addresses the needs of historically oppressed populations. Furthermore, we must consider how issues of healthcare access and patient empowerment affect both data collection and use of the algorithm in practice. This may be best implemented by empowering diverse oversight and planning committees to be ensure the ethics and and accountability of the research projects.

4 STAKEHOLDERS

In establishing guidelines for dataset collection for AI-enabled diagnostics, we must consider the interests of the various stakeholders, as well as how these interests may at times conflict with each other. We want to explain both concrete strategies for ensuring appropriate dataset diversity, and the reasoning behind each of these recommendations, to ensure that relevant stakeholders understand why our recommendations are important and should be followed. Specifically, we will consider the following stakeholders: patients, governmental agencies and regulatory bodies, medical corporations (including pharmaceutical and medical devices companies), insurers, doctors and their representatives, and researchers.

First, we must consider the people who are in many ways the ultimate stakeholders: the patients who will be diagnosed through emerging AI systems. Typically, patients want to have AI-enabled diagnostic systems be as accurate and unbiased as possible, while also minimizing the costs (financial or otherwise[1]) of such systems, so that they are widely accessible to those who may need to utilize them. Governmental regulatory bodies often serve as proxies for protecting consumers, as they create and enforce regulations to ensure patient safety. Patients should already be willing to support these guidelines, as the proposed rules were designed with specifically their interests in mind. We believe that the most organized way to engage with patient stakeholders is through patient advocacy organizations, such as the American Heart Association and the National Organization for Rare Disorders. Many of these organizations are represented in the National Health Council, a nonprofit umbrella organization [9].

Another set of key stakeholders are governmental regulatory bodies, chiefly the FDA (Food and Drug Administration). It is the mission of the FDA to ensure that new medications and medical devices are safe to use and meet testing standards before being available on the market. The FDA provides rules and guidance on the clinical trial process, which has been required for new drugs since the 1970s. The FDA is especially concerned with "Adherence to the principles of good clinical practice (GCP), including human subject

[1]Sometimes patients indicate that they cannot come in to the clinic more than once because they cannot take time off from work.

protection (HSP)" [1]. Relevant to our goal of ensuring the diversity of datasets used for AI-enabled diagnostics, the FDA recently released a set of guidelines [17] aimed at enhancing diversity in clinical trial populations, which is currently in the public comment phase. The guidelines are non-binding and include such recommendations as broadening clinical trial population recruitment pools through advertising their existence through a wider set of channels and specifically targeting underrepresented populations, as well as putting in place "inclusive retention policies", meaning ways of ensuring that diverse groups of patients stay in clinical trials once they have started participating in them. This demonstrates that they already support establishing guidelines to increase diversity in datasets for AI healthcare tools. The FDA has the power to regulate and approve AI-enabled diagnostic models, and they have the duty to ensure that every SaMD is safe and effective.

Another set of stakeholders that we will need to consider in our analysis are pharmaceutical companies and other corporations that operate in the medical space. Medical corporations serve an important role in providing necessary supplies for the diagnosis and management of diseases, and are motivated by sales of these supplies. Our current system does not incentivize medical corporations to ensure equal access to their services, especially when operating in conjunction with fee-for-service payment models. We believe that it is unlikely that industry groups and for-profit corporations would "self-regulate" appropriately to enforce dataset collection guidelines and audits on themselves. We believe that many of these biases – by virtue of applying to minority populations – represent market failures where sufficient effort towards diversity will yield diminishing financial returns. A benefit of increasing the diversity of available medical datasets is that this information will help companies develop more personalized drugs. The granularity of new datasets developed as a result of our recommendations may also help companies better serve the needs of diverse market segments that were previously overlooked.

Most healthcare provider decisions are heavily influenced by insurance coverage in the U.S., and thus insurers will ultimately determine the success or failure of AI diagnostic tools. Insurers must agree to reimburse the use of these models in hospitals in order for them to be adopted. We can look to telemedicine as a cautionary tale for Diagnostic AI: a seemingly good idea, yet most hospitals and clinics never adopted at scale. That said, in capitated healthcare systems, such as the VA (where the healthcare system is incentivized to keep patients healthy and does not receive additional compensation by doing additional services), telemedicine is widely used. This demonstrates the disconnect between a tool that is effective and a tool that is widely used. Insurers main concerns will be regarding model accuracy and liability; because the software is being treated as a medical device by regulators, we think it is likely for insurers to reimburse for them as well.

We must also consider doctors, the medical associations who represent them, such as the American Medical Association (AMA), and hospital systems where these doctors practice medicine. Hospitals are likely to be in favor of these AI-based diagnostic models, as they will – in theory – standardize adoption and reliability of clinical systems. Doctors will be on the frontlines of using AI-enabled diagnostics systems with their patients to augment their practice. Ideally the tools will return more time and energy to devote to truly empathizing and caring for their patients instead of bookkeeping (a common complaint among medical professionals) [50]. It is vital that doctors' interactions with AI be extremely thoughtfully considered, in order to avoid burdening doctors with clunky systems and creating overreliance on algorithms that lack human common sense [52]. It is also important to ensure that the guidelines developed by the FDA are easily understood by doctors and other medical professionals, so that doctors understand that these systems are beneficial for the patients.

Finally, in making our recommendations we want to emphasize the importance of researchers as stakeholders (in research universities, research institutions, and private company settings), who will be assembling the datasets for AI-enabled diagnoses. We expect that the prestige-based incentive described in Section 5 will help encourage researchers to develop these diverse datasets, especially if it is tied to special grants.

5 RECOMMENDATIONS

Our primary goal in this paper is to recommend a set of guidelines for reducing bias in diagnostic AI systems. These varieties of biases, as covered in Section 3, are representation bias, measurement bias, aggregation bias, and historical bias. The general approaches we would like to take in combating these kinds of bias are as follows:

From a pragmatic perspective, managing each kind of bias requires a different kind of intervention. Representation bias is perhaps the most straightforward to address, in that there is clearly a need for the population in a medical data training set to be representative of the population with which a diagnostic algorithm will be used. Measurement bias is more difficult to detect from a regulatory standpoint; it requires a high degree of domain expertise to know whether the collection of data for diagnosis is oversimplifying a disease or that risk factors are measured incorrectly. Aggregation bias similarly requires expertise in medicine and social determinants of health, and a technical understanding of best practices in selecting models and training data. Finally, correcting for historical bias and the idea of encouraging more equitable research is a nuanced and sensitive issue of social change. Please see Table 1 for a summary of these varieties of bias and our corresponding recommendations.

5.1 Addressing Representation and Measurement Bias

To combat representation bias, we would like to encourage training on diverse datasets for data used in AI-enabled diagnostics systems. Our recommendations will be primarily aimed at governmental regulatory bodies, chiefly the FDA. First, we suggest that the FDA draft a similar variety of document to the current guidelines on encouraging diversity in clinical studies (which are currently under public comment) [2]. The datasets should include breakdowns by demographic (e.g. gender, age, ethnicity), in keeping with guidelines set forth by the FDA for transparency of representation in clinical datasets [14]. This document would also include a suggestion to use a standard data format, such as FHIR, to make it easier for these regulatory bodies to test the AI models without needing to convert the data to the format that the developers chose to use for this particular system. This should also help compare the quality of data across sub-populations to assess potential measurement bias.

In order to encourage the formation of such datasets, we would recommend that the NIH and the medical research community in general institute incentives to increase open data sharing. The main evidence-based incentive that has been investigated to date in this area is badges for data sharing [46]. This practice involves journals awarding badges to researchers that have curated high quality datasets that others may benefit from, and was proven effective at increasing the number of data sharing publications in the journal [33]. Many experts in this area suggest that an effective incentive system may involve the use of prestige and recognition [46]; for instance, the NIH could have a page of top contributing institutions, and include author names and affiliations immediately alongside public datasets in the DASH dataset webpage (prestige is a powerful motivator, and has a relatively low implementation cost). We would also specifically recommend larger grants for the development of datasets with higher representations of historically underprivileged groups.

Secondly, we suggest that the FDA develop private data repositories with diverse patient profiles, and test AI-based diagnostic models before they are used in clinical practice. Testing models with these datasets will help determine if the model suffers from measurement and representation bias based on previously-utilized insufficient training data. The easiest way to collect this data would be to put out an RFP requesting this sort of dataset while keeping it private (to ensure developers of diagnostic models cannot train on the testing data beforehand), which would necessarily follow all of the FDA's guidelines. These RFPs would be released based on known trends in the field (e.g., based on the knowledge that there is a high interest in developing algorithms for automating radiology tasks right now, the FDA would release an RFP for chest X-Ray datasets) and in anticipation of applications from private developers. They could develop the dataset in-house, but this would be more difficult, as the FDA as a whole does not have much of a background in data collection. The use of a common data format will make this significantly easier, but the FDA will need to develop conversion programs from the standard format to other formats for any AI models that do not use the chosen standard format. Alternatively, the FDA could simply not approve any models that do not use the standard data format, which would simplify their process. We would further recommend that the FDA draw on these datasets to improve the robustness of these models before they enter the market.

5.2 Addressing Aggregation Bias

As discussed in Section 3, aggregation bias is different from is closely tied to heterogeneity in the dataset. Recent technical work in hidden stratification has identified methods for counteracting these harms on both the supply side (Schema Completion by the dataset creator) and demand side (Error Auditing by a regulator, such as the FDA) [39]. We believe that the best way to reduce this insidious bias is with a combination of both approaches.

On the dataset creation side, we recommend the creation of a industry-standards "best practices" checklist for dataset creation. This checklist will embrace transparency, and encourage Schema Completion for issues that should be foreseeable to the creator. As it stands, there is little incentive for hospitals to release medical datasets for researchers to use. Because of the additional hurdles that this could impose on an already undervalued role of releasing data, the FDA and NIH should institute incentives for medical institutions to curate more medical datasets with said transparency checklists akin to "Datasheets for Datasets" [21] and "Model Cards for Model Reporting" [37]. We envision that this checklist would include factors such as demographic breakdown, quantification of heterogeneity, IRB approval notes, recommended stratification on subpopulations, mechanisms used in medical data collection (e.g. sensors and human notes), and other clinically relevant considerations. We would also want to include more nuanced data, such as where and over what time scale the dataset was collected, as well as any limitations, legal matters, or ethical considerations that the collectors wish to share.

We also recommend that the FDA ensure strong Error Auditing during the SaMD approval process. To catch non-obvious errors like the "untreated pneumothorax" case, we strongly encourage having practicing clinicians (or other appropriate stakeholders) spot check randomly selected output of the model, similar to what was done in the Hidden Stratification paper [39].

5.3 Addressing Historical and Societal Bias

To address historical bias, we recommend that the NIH, NSF, and other federal grant-giving bodies provide additional research incentives for the gathering of datasets and studies that provide insights into the experiences of minority populations specifically, and into diagnostic nuances that are confounded by factors such as ethnicity, socioeconomic status, sexual orientation, and gender. This could, for example, take the form of an NIH Request For Applications (RFA) in the area of bias research, diverse dataset collection and model building [43].

In the practice of data collection, we also recognize the potential for unethical practices. It is critically important that these incentives promote doing research *with* these communities, and not just research *about* them. In 2019, Google – with the desire to train their Facial Recognition AI on a diverse set of faces – hired contractors to photograph people experiencing homeless and took their pictures [45]. Although in a narrow sense, this would help that Facial Recognition tool better handle diverse users, this approach was wrong. In *Data Feminism*, Catherine D'Ignazio and Lauren Klein introduce their framework for understanding data science and research through the lens of power [13]:

- Data science by whom?
- Data science about whom?
- Data science with whose values?

When outcomes are consistently unjust, redesigning the process becomes the only way to build trust in a system. For this reason, these incentives should be structurally designed to empower impacted communities, perhaps by giving those stakeholders the authority to award grants to projects that would be address their needs. Trials should be designed and conducted with informed consent, ethics, and justice in mind (and in accordance with legislation surrounding the National Research Act of 1974 [42]).

Table 1: Bias Types and Corresponding Recommendations

Bias Type	Example	Recommendation
Representation Bias: The training set does not accurately represent real-world data.	Training a melanoma classifier to detect cancer for patients with white skin only, and then expecting it to perform well on darker skin colors.	1. Establish best practices in dataset collection 2. Encourage diverse dataset development through NIH RFPs 3. FDA approval pathways should involve testing on diverse held-out data
Measurement Bias: The way that the training data is measured/collected introduces inaccuracies in the model.	Inconsistencies in the quality of EHR data for a given patient in different hospitals introduces bias from different ways of measuring patient outcomes.	1. Establish best practices in dataset collection 2. Encourage diverse dataset development through NIH RFPs 3. FDA approval pathways should involve testing on diverse held-out data 4. Involve domain experts early on in the process of designing medical AI, in order to avoid improper collection
Aggregation Bias: Multiple datasets are improperly combined in a single model.	Combining all blood biomarkers for diabetes across different ethnicities, and expecting it to perform well for all ethnicities despite biomarker differences between groups.	1. Require a limited-deployment stage in approving SaMD where the algorithm is tested 2. Involve conversations between domain experts and data scientists to identify potential issues with aggregation
Historical/Societal Bias: Datasets accurately reflect our current reality, and therefore existing societal inequality.	Developing an algorithm that uses ZIP code as a feature in predicting hospital length of stay, for use in assigning a case manager to those who are predicting to stay a shorter amount of time. Results in discrimination against socioeconomically disadvantaged patients.	1. Encourage diverse dataset development through NIH RFPs 2. Developers should work with ethicists and all stakeholders to consider the implications of using particular dataset features

5.4 Privacy and Consent

Additional research is necessary to capture the appropriate balance between dataset information and protection of patient privacy (a good test would be: does this information about the dataset allow us to trace the data back to specific patient groups?). However, the current technical solutions to the de-identification problem are a good starting point for our recommendations. The MIMIC dataset used a rule-based de-identification system. "The de-identification process for structured data required the removal of all eighteen of the identifying data elements listed in HIPAA, including fields such as patient name, telephone number, address, and dates. In particular, dates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals … Time of day, day of the week, and approximate seasonality were conserved during date shifting" [31]. Recurrent Neural Nets, which are a subset of neural network methods, have been shown to outperform manual de-identification rules in de-identifying EHR data [12]. NeuroNER [11], developed in the same lab as MIMIC, provides an interface for non-experts to label data points to train a named-entity recognition (NER) neural network such as the one cited above. Such neural networks are used to identify sensitive patient data in health records, and could be used to implement our recommendations. The NIH and other relevant bodies may consider introducing recent developments in privacy technology for medical AI - namely, the concepts of remote execution, differential privacy, end-to-end encryption, and secure multi-party communication. These concepts would also be valuable to private entities (e.g. hospitals and healthcare organizations) developing secure gateways to their own patient information. Overall, there is a rich background for technological tools for patient de-identification in health-related datasets, and we advocate for the architects of solutions based on our recommendations to make use of these tools.

By capturing as much information about the data collection process and recommended usage as possible, we can easily trace instances of bias to specific data collection factors. We would also be able to quickly identify the potential limitations of a given dataset (for example, that the data may have been inaccurate because of the use of a particular medical device). This would involve some amendments to be included in submissions to the NIH's Data and Specimen Hub (DASH) [44].

6 DISCUSSION

In this paper, we argue for targeted incentives, regulations, and data collection/auditing to address a multitude of related problems in dataset bias for AI Diagnostics.

One concern with these recommendations is these additional guidelines on datasets will serve as restrictions on innovation, and that it will be harder for smaller companies to enter the healthcare AI space as a result. This argument is often introduced whenever the government considers new regulations for industry.

However, the evidence does not show this to be true. In the biotechnology industry, for example, when the FDA increased pre-approval scrutiny in medical devices in 1976, there was a drop in innovation, but it was temporary [48]. Market innovation rebounded after a few years, as companies learned to adjust to the new regulations. While in some cases there was a decrease in innovation with new regulations, this only occurred with regulations that required significant changes in technology, or if there was a lot of uncertainty regarding future regulation [48]. Sometimes regulation can even allow companies to improve their operations. For example, the European Union's GDPR (General Data Protection Regulation) required many tech companies to significantly change how they handled data. This resulted in companies keeping better track of their data and how it flows, forcing different departments to communicate with each other more, which was beneficial for future innovation in these companies[34].

Another concern is that the implementations of our recommendations – particularly the data collection – could be a challenging feat for a governmental agency to pull off. In order for the FDA to successfully implement the technical systems that will handle the collection of sensitive patient data, they must prioritize security and robustness for the design of these systems. The most famous example of the US government failing to deliver on a large IT project is the botched rollout of healthcare.gov in 2013. The main cause of this failure is that different government contractors were working in silos on different parts of the website, and there was no effective structure in place to properly relay feedback and critical issues from one silo to another [24]. Since then, many agencies have learned from that failure, and have introduced reforms to breakdown silos and ensure better communication throughout that agency. The FDA is currently executing its Technology Modernization Action Plan [14], which will:

(1) Modernize their technical infrastructure.
(2) Enhance their capabilities to develop technology products to support its regulatory mission.
(3) Communicate and Collaborate with stakeholders to drive technological progress that is interoperable across the system and delivers value to consumers and patients.

While proposals to reform government contract procurement fall outside the scope of our paper, we point to such proposals [32] to help mitigate failure in government digital endeavors.

7 CONCLUSION

In this proposal, we have outlined the recent history of how bias in datasets used for clinical trials – and more specifically in machine learning-based diagnostics – create unfair and inaccurate outcomes. There is urgency in developing a coherent set of policy recommendations for addressing this emerging issue. We have also enumerated the different, surprising ways bias can arise in the healthcare data. It is critical to address these problems in a targeted way, with different policy interventions for different forms of dataset bias. Through a combination of incentives, regulations, and increased transparency, we believe that many of the serious problems in biased AI-based diagnostics can be addressed. We urge policymakers and other relevant stakeholders to consider recommendations and address the issue of bias in AI healthcare datasets.

ACKNOWLEDGMENTS

We would like to acknowledge the support of the communication instructors at MIT - most notably, Dr. Michael Trice, in crafting this paper.

In addition, we acknowledge the support of the course staff for 6.805 (Internet Policy): Dr. R. David Edelman, Prof. Daniel Weitzner, Prof. Michael Fischer, and Prof. Hal Abelson for suggesting the paper topic and working with us to refine our ideas.

REFERENCES

[1] Food Drug Administration. 2019. Clinical Trials and Human Subject Protection. https://www.fda.gov/science-research/science-and-research-special-topics/clinical-trials-and-human-subject-protection
[2] Food Drug Administration. 2019. Enhancing the Diversity of Clinical Trial Populations -Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry DRAFT GUIDANCE. https://www.fda.gov/media/127712/download
[3] Food Drug Administration. 2019. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye
[4] Food Drug Administration. 2019. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) -Discussion Paper and Request for Feedback. https://www.fda.gov/media/122535/download
[5] Food Drug Administration. 2019. Statement from FDA Commissioner Scott Gottlieb, M.D. on steps toward a new, tailored review framework for artificial intelligence-based medical devices. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-steps-toward-new-tailored-review-framework-artificial
[6] Ayal A. Aizer, Tyler J. Wilhite, Ming-Hui Chen, Powell L. Graham, Toni K. Choueiri, Karen E. Hoffman, Neil E. Martin, Quoc-Dien Trinh, Jim C. Hu, and Paul L. Nguyen. 2014. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer* 120 (02 2014), 1532–1539.
[7] Isaac and Kohane, Hyun Won Chung, and Andrew Beam. 2019. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv* (02 2019).
[8] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. 2019. (De)Constructing Bias on Skin Lesion Datasets. http://openaccess.thecvf.com/content_CVPRW_2019/papers/ISIC/Bissoto_DeConstructing_Bias_on_Skin_Lesion_Datasets_CVPRW_2019_paper.pdf
[9] National Health Council. 2014. Membership Directory. http://www.nationalhealthcouncil.org/about-nhc/membership-directory
[10] Kelly Davio. 2019. Precision Medicine Research Subject to Bias in Datasets and Outcomes. https://www.ajmc.com/focus-of-the-week/precision-medicine-research-subject-to-bias-in-datasets-and-outcomes
[11] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. https://arxiv.org/abs/1705.05487
[12] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24 (12 2016), ocw156. https://academic.oup.com/jamia/article/24/3/596/2769353
[13] Catherine D'Ignazio and Lauren Klein. 2020. *Data Feminism.*
[14] FDA. 2019. FDA's Technology Modernization Action Plan (TMAP). https://www.fda.gov/media/130883/download
[15] Kadija Ferryman and Mikaela Pitcan. 2018. Fairness in Precision Medicine. https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf
[16] U.S. Food and Drug Administration (FDA). [n.d.]. DE NOVO CLASSIFICATION REQUEST FOR IDX-DR. https://www.accessdata.fda.gov/cdrh_docs/reviews/

DEN180001.pdf
[17] Center for Drug Evaluation and Research. 2019. Enhancing the Diversity of Clinical Trial Populations — Eligibility Cr. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial
[18] Kaiser Family Foundation. 2019. Black Americans and HIV/AIDS: The Basics. https://www.kff.org/hivaids/fact-sheet/black-americans-and-hivaids-the-basics/#footnote-391734-1
[19] The Medical Futurist. 2019. A.I. Bias In Healthcare. https://medicalfuturist.com/a-i-bias-in-healthcare/
[20] The Medical Futurist. 2019. FDA Approvals For Smart Algorithms In Medicine In One Giant Infographic. https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/
[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/
[22] Dave Gershgorn. 2018. Health care AI has the same racial and gender biases as its trainers. https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/
[23] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine* 178 (11 2018), 1544. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/
[24] Amy Goldstein. 2016. HHS failed to heed many warnings that HealthCare.gov was in trouble. *The Washington Post* (02 2016). https://www.washingtonpost.com/national/health-science/hhs-failed-to-heed-many-warnings-that-healthcaregov-was-in-trouble/2016/02/22/dd344e7c-d67e-11e5-9823-02b905009f99_story.html
[25] Prabhakaran Gopalakrishnan, Moluk Mirrasouli Ragland, and Tahir Tak. 2009. Gender differences in coronary artery disease: review of diagnostic challenges and current treatment. *Postgraduate medicine* 121 (2009), 60–8. https://www.ncbi.nlm.nih.gov/pubmed/19332963
[26] H A Haenssle, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, L Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* 29 (05 2018), 1836–1842. https://academic.oup.com/annonc/article/29/8/1836/5004443
[27] William H. Herman and Robert M. Cohen. 2012. Racial and Ethnic Differences in the Relationship Between HbA1c and Blood Glucose. *Obstetrical Gynecological Survey* 67 (08 2012), 468–469.
[28] William R. Hersh, Mark G. Weiner, Peter J. Embi, Judith R. Logan, Philip R.O. Payne, Elmer V. Bernstam, Harold P. Lehmann, George Hripcsak, Timothy H. Hartzog, James J. Cimino, and Joel H. Saltz. 2013. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care* 51 (08 2013), S30–S37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3748381/
[29] George Hripcsak, Charles Knirsch, Li Zhou, Adam Wilcox, and Genevieve Melton. 2011. Bias Associated with Mining Electronic Health Records. *Journal of Biomedical Discovery and Collaboration* 6 (2011), 48–52. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149555/
[30] The International Skin Imaging Collaboration (ISIC). 2019. ISIC Archive. https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/isicArchive
[31] Alistair Johnson, Tom Pollard, Lu Shen, Li-Wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. OPEN SUBJECT CATEGORIES Background Summary. (2016). https://lcp.mit.edu/pdf/JohnsonSD2016.pdf
[32] Steven Kelman. 2001. Remaking Federal Procurement. https://sites.hks.harvard.edu/fs/skelman/Remaking%20Federal%20Procurement%20Visions%20paper.pdf
[33] Mallory C. Kidwell, Ljiljana B. Lazarević, Erica Baranski, Tom E. Hardwicke, Sarah Piechowski, Lina-Sophia Falkenberg, Curtis Kennett, Agnieszka Slowik, Carina Sonnleitner, Chelsey Hess-Holden, Timothy M. Errington, Susann Fiedler, and Brian A. Nosek. 2016. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology* 14 (05 2016), e1002456. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4865119/
[34] Sameena Kluck. 2018. Regulations: Stifling Innovation or Forcing Companies to Better Communicate and Innovate? http://www.legalexecutiveinstitute.com/regulations-communicate-innovate-dla-piper/
[35] Jonas Korlach. 2018. We Need More Diversity in Our Genomic Databases. https://blogs.scientificamerican.com/voices/we-need-more-diversity-in-our-genomic-databases/
[36] Harvard Medicine Magazine. 2019. The Importance of Nuance. https://hms.harvard.edu/magazine/artificial-intelligence/importance-nuance
[37] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. *CoRR* abs/1810.03993 (2018). arXiv:1810.03993 http://arxiv.org/abs/1810.03993
[38] Ana Mulero. 2019. FDA Speeds Up Artificial Intelligence Approvals, Review Finds. https://www.raps.org/news-and-articles/news-articles/2019/1/fda-speeds-up-artificial-intelligence-approvals-r
[39] Luke Oakden-Rayner*, Jared Dunnmon*, Gustavo Carneiro, and Christopher Re. 2019. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In *Proceedings of Machine Learning Research (Proceedings of Machine Learning Research)*. PMLR, Vancouver, Canada. https://arxiv.org/pdf/1909.12475.pdf
[40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342 arXiv:https://science.sciencemag.org/content/366/6464/447.full.pdf
[41] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342 arXiv:https://science.sciencemag.org/content/366/6464/447.full.pdf
[42] U.S. Department of Health Human Services (HHS). 2016. The Belmont Report. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
[43] National Institutes of Health. 2019. What Does NIH Look For? | grants.nih.gov. https://grants.nih.gov/grants/what-does-nih-look-for.htm
[44] US National Institutes of Health (NIH). 2019. NICHD DASH - Eunice Kennedy Shriver National Institute of Child Health and Human Development Data and Specimen Hub. https://dash.nichd.nih.gov/datasetExplorer
[45] Ginger Adams Otis and Nancy Dillon. 2019. Google using dubious tactics to target people with 'darker skin' in facial recognition project: sources. https://www.nydailynews.com/news/national/ny-google-darker-skin-tones-facial-recognition-pixel-20191002-5vxpgowknffnvbmy5eg7epsf34-story.html
[46] Anisa Rowhani-Farid, Michelle Allen, and Adrian G. Barnett. 2017. What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review* 2 (05 2017).
[47] Anna Shcherbina, C. Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey Christle, Trevor Hastie, Matthew Wheeler, and Euan Ashley. 2017. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine* 7 (05 2017), 3. https://www.mdpi.com/2075-4426/7/2/3
[48] Luke Stewart. 2010. The Impact of Regulation on Innovation in the United States: A Cross-Industry Literature Review. https://www.itif.org/files/2011-impact-regulation-innovation.pdf
[49] Harini Suresh and John Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. https://arxiv.org/abs/1901.10002
[50] Eric J Topol. 2019. *Deep medicine : how artificial intelligence can make healthcare human again.* Basic Books, March.
[51] Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25 (01 2019), 44–56. https://www.nature.com/articles/s41591-018-0300-7
[52] Robert M Wachter. 2017. *The digital doctor hope, hype, and harm at the dawn of medicine's computer age.* New York Mcgraw-Hill Education.
[53] Nicole G. Weiskopf, Suzanne Bakken, George Hripcsak, and Chunhua Weng. 2017. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGEMs (Generating Evidence Methods to improve patient outcomes)* 5 (09 2017), 14.