Machine Learning in Healthcare: How Strict Data Collection Policies Lead to Misrepresentational Models

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jakub Lipowski

Spring 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

Introduction:

Developments in analytics have shaken the way industries view their key stakeholders, whether it be through uncovering descriptive data to characterize populations, predictive data to determine future outcomes, or prescriptive data to aid in crafting actionable solutions. Furthermore, advances in machine learning algorithms have unlocked a new realm of decision-making capabilities. When it comes to healthcare, implementation of artificial intelligence raises numerous sociotechnical questions and concerns. One particular area that is subject to considerable controversy is the collection and usage of patient data in machine learning models. Machine learning models, especially when applied to healthcare, need to be trained with data that is representative of several demographics so that they are able to fairly and effectively generalize to those demographics (Gianfrancesco et al., 2018). First, it is necessary to develop an appreciation for the key tenants of a successful machine learning model. Subsequently, it is important to understand how existing data systems and policies applied to healthcare in the United States fit into the grand scheme of a data revolution. This research ultimately aims to reconcile the need for representative training during machine learning model development with the limitations that the United States' public policies place on data collection and privacy.

Research Question and Methods:

The research question being considered is: "How can the discrepancy between a need for representative training data and strict data collection polices within an unequal healthcare system be reconciled to minimize the risks of biased machine learning models in the United States?"

The methodology used to answer this sociotechnical question is two-fold. First, regulations concerning protection and accessibility of health data are discussed. Next, case studies identifying instances of flawed machine learning algorithms in healthcare are presented.

The case studies are published in *Smithsonian Magazine* and *WIRED*. The case studies are contextualized by the preceding discussion of health data policy in the United States. Following the presentation of both policies and case studies, the two are connected by understanding how certain policies within the United States healthcare system lend themselves to discriminatory machine learning models that are exemplified in the case studies. The connection between policy and case study is where the bulk of the analysis occurs. Incorporated into the analysis is the concept of risk analysis, which acts as a vehicle for guiding discussion of the research question. Finally, a suggestion for action in the field is presented not only to help answer the research question, but to inspire future action. Key terms in this study include "machine learning," "model", "artificial intelligence", "healthcare", "data", "HIPAA", and "Privacy Rule".

Supportive Background Information:

What is Machine Learning?

Formally defined by IBM, "Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy." Machine learning relies on using data to train an algorithm to make decisions about future instances. Algorithms relying on statistical methods are used to classify pieces of information or predict future events (*What Is Machine Learning?*, 2021). When applied correctly, machine learning has potential to save people and organizations time, financial capital, and uncertainty. The value proposition of machine learning models is that they can continue to improve themselves without being explicitly programmed. Currently, organizations across all industries are adopting artificial intelligence to optimize their operations (*What Is the Definition of Machine Learning? | Expert.Ai | Expert.Ai*, n.d.).

Healthcare and the Cross-Industry Standard Process for Data Mining

First developed in 1996 as a project spearheaded by companies in the European Union, the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework serves as a highly generalizable outline for developing successful machine learning models. CRISP-DM consists of six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment ("CRISP-DM," n.d.). A simple visualization of the framework is depicted in Figure 1.

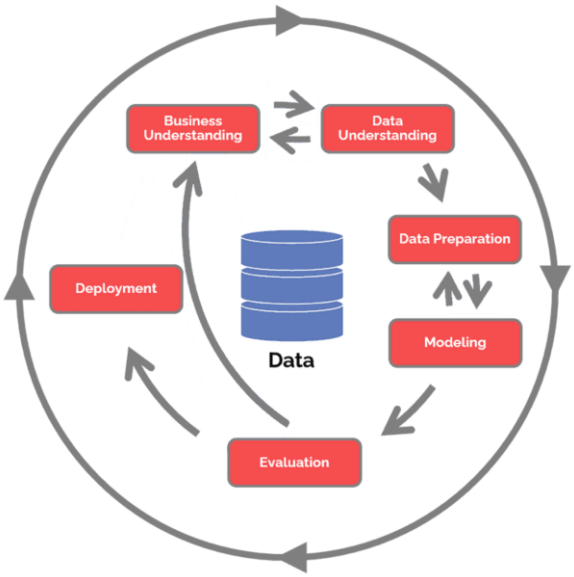

Figure 1: CRISP-DM framework for data-driven, analytical endeavors.

The first step of CRISP-DM in the context of healthcare, business understanding, encompasses understanding of topics including but not limited to personalized medicine, drug discovery, disease diagnosis, pandemics, electronic record keeping, and imaging (*Top 10 Applications of Machine Learning in Healthcare - FWS*, n.d.). Data understanding refers to an appreciation for the kind of data that is necessary to characterize a disease and the demographics it represents. The next three steps, data preparation, modeling, and evaluation refer to the technical aspects of model development and are beyond the scope of this sociotechnical research.

Deployment entails using a machine learning model in a clinical setting to diagnose patients and aid physicians in decision-making protocols. The pain point of machine learning models in healthcare comes at the data understanding stage, while the problems manifest at the deployment stage. The shortcomings in data understanding within the scope of the United States healthcare system are a result of two factors: variable data collection practices and a history of discriminatory regulations.

Variable Data Collection Practices: Health Insurance Portability Accountability Act

The United States implements the Health Insurance Portability and Accountability Act (HIPAA) to protect personally identifiable health-related information stored by healthcare institutions such as health plans, providers, clearinghouses, and select business associates. However, this regulation does not apply to other sources of health data, including businesses developing phone applications and devices like the Apple Watch (Vayena et al., 2018). Although HIPAA helps with uneasiness concerning data privacy, it also makes it extremely difficult to collect accurate data to train machine learning models in healthcare settings. Under HIPAA, researchers developing machine learning models have access to deidentified information, which refers to data that cannot be used to identify a person. However, oftentimes personally identifiable information is necessary for projects in fields such as personal medicine, which require information about a patient down to their genetic code (Nair, 2010). Such incomplete data could be skewed to favor particular geographies, demographics, or origins of data.

Discrimination in Healthcare

Although the United States ended legal segregation with the Civil Rights Act of 1964, the effects of racial bias is still noticeable when it comes to healthcare data (Data et al., 2003).

African Americans, for example, make up 27.5 million of uninsured people in the country, with 45% of the demographic citing cost as being the leading cause for lack of insurance. Additionally, for 42 million African Americans, the average cost of health care premiums is about 20% of their average annual household income (*Racism, Inequality, and Health Care for African Americans*, 2019). To make matters worse, minorities' insurance claims and electronic health records are more exposed to careless errors, causing them to receive suboptimal treatments (Hoffman & Podgurski, 2020). By virtue of having less access to healthcare institutions due to systemic inequalities, certain demographics hold less representation in aggregated health data. Machine learning models cannot be developed to serve demographics for which there is not enough training data, thus perpetuating the existing disparities found in the American healthcare system.

Risk Analysis:

In *Risk Society,* German sociologist Ulrich Beck defines risk as a "systematic way of dealing with hazards and insecurities induced and introduced by modernization itself." Beck is referring to the concept of risk analysis, an important sociotechnical framework that can be used to explain society's caution towards adoption of new technologies. Novel technologies introduce both potential for advancing society and elements of uncertainty about the effects of the technologies and modernization. In an effort to mitigate risk, society often ends up examining itself more closely and therefore changing itself in the process. British sociologist Anthony Giddens coined the term "reflexivity" to describe the phenomena of social self-examination (*Beck's Sociology of Risk: A Critical Assessment - Anthony Elliott, 2002*, n.d.).

Risk analysis is an appropriate lens for analysis of the disparity between data collection practices and the need for representative training data for healthcare models in an increasingly

digitized society. As researchers and scientists work towards the integration of healthcare institutions with artificial intelligence, the general public has developed a skepticism towards adoption. The most frequent users of the American healthcare systems are senior citizens, most of who are not tech savvy (Americans, 2008). As a result, they feel more distrust towards tools like machine learning models being used during treatments. This kind of risk-aversion is part of the process of risk analysis. Additionally, by recognizing that data stores lack representative data from many demographics, the United States is being reflexive by examining the effects of its history on the current healthcare system and the impact it can have as artificial intelligence advances.

However, a common criticism of risk analysis is the idea that it slows progress and defends harmful policies (Eid, 2003). Skepticism towards adoption is certainly one factor that is slowing progress at the intersection of artificial intelligence and healthcare. The other factor is that policymaking is not supportive of large-scale machine learning models in healthcare. As discussed, certain policies introduced by HIPAA aim to avoid risk by easing Americans' minds about data privacy. Risk analysis can be used to consider a balance between protection of health data and its utility in large-scale machine learning model development.

Results and Discussion:

Data collection policies can become more cooperative with machine learning model development in healthcare by eliminating stringent standards of de-identification of patient health data and creating a space for machine learning engineers and physicians to collaborate in the data collection and model development process. It is important to first understand the governing laws that affect collection of health data in the United States. Before discussing how policy can be amended, focusing on instances of data collection gone wrong sets the scope of

issues with health data aggregation. Considering that artificial intelligence in healthcare is expected to grow by an annualized 48% between 2017 and 2023, it is critical that a discussion about data collection policy and instances of machine learning in healthcare be had (Intelligence, n.d.).

Health Data Regulation and Policy

Prior to analyzing the limitations of health data regulation, it is imperative to become familiar with the facts pertaining to key policies governing the healthcare industry. The most relevant piece of legislation to health data regulation is the Health Insurance Portability and Accountability Act (HIPAA). HIPAA is a law introduced in the United States in 1996, designed to protect health information from being distributed without the patient's consent. One of the key tenants of HIPAA is the *Standards for Privacy of Individually Identifiable Health Information*, commonly referred to as the Privacy Rule. The Privacy Rule established a set of national standards to protect certain health information. The information protected by the Privacy Rule includes data relating to a patient's past, present, or future physical or mental health condition, the provision of health care to the individuals, and demographics. A key nuance behind the Privacy Rule is that there must be a reasonable basis to believe that data can be used to identify an individual. De-identified data, however, is permitted to be used in studies. Data is considered de-identified if it does not provide information about the unique identity of an individual. De-identification of data must be completed by a qualified statistician (*Summary of the HIPAA Privacy Rule | HHS.Gov*, n.d.).

While de-identification of data is possible, it can require significant time and resources. Although maintaining privacy is important for the sake of public trust, certain tenants of HIPAA are outdated because they hinder promising, cutting-edge research in the machine learning space

by preventing access to training data. Authors of the Privacy Rule argue that the regulations streamline the research process by creating standards for de-identification. However, the process just adds additional steps that make data less accessible (Rights (OCR), 2012).

Case Study: The DNA We Have Is Too White

Precision medicine uses genomic information along with prescriptive analytics to develop personalized medical conclusions about patients. The problem is that 87% of participants in worldwide genomics research come from European descent. Participants of African descent account for 3% of the data while participants of Hispanic descent account for a measly 0.5%. For comparison, 60.1%, 13.4%, and 18.5% of the United States population consists of Caucasians, African Americans, and Hispanics, respectively (*U.S. Census Bureau QuickFacts*, n.d.). Figure 2 visualizes the discrepancy between representation in DNA data and the actual United States population. The rest of participants are of Southeast Asian ancestry. As a result, advances in personalized medicine are only available to people of European descent.

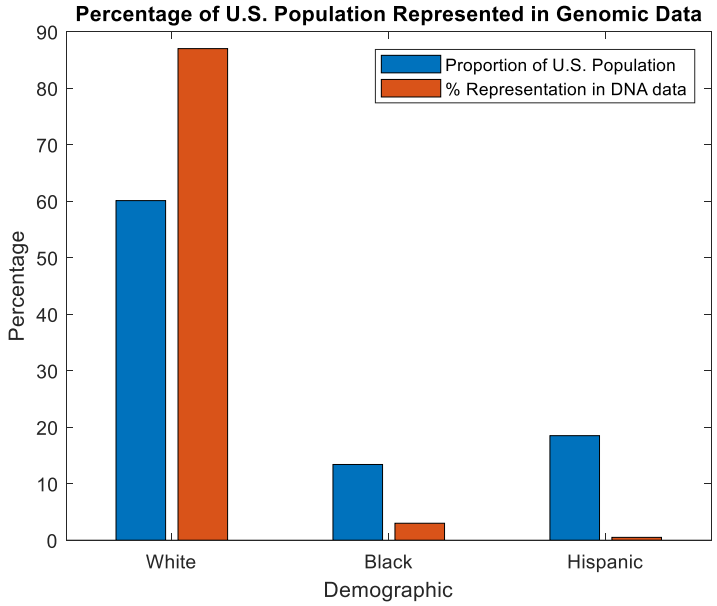

Figure 2: Percentage of United States population represented in genomic data (Lipowski, 2022).

The “All of Us” initiative aims to collect data from underrepresented communities in order to ensure that all people, no matter their ethnicity, are able to reap the benefits of technological advancement in the field of precision medicine. In order to obtain this data, members of the “All of Us” initiative are collaborating with drugstores, blood banks, and labs to obtain the desired genomic data. However, the task at hand is nowhere near as simple as knocking on doors and obtaining data. Minority groups have a higher distrust towards volunteering for medical studies. Incidents like the experiments at the Tuskegee Institute where African Americans were unknowingly infected with syphilis for decades and Henrietta Lacks’ cells that were taken from her without permission to benefit research that is not accessible to members of her own demographic are recent (Stein, 2017). Unsurprisingly, minority groups are reluctant to give up their genomic data. Minorities speculate that their genomic data could be used to develop technologies that will not be made accessible to them in the future (Magazine, n.d.).

Case Study: A Healthcare Algorithm Offered Less Care to Black Patients

Algorithms are becoming increasingly more prevalent in the healthcare space. One machine learning algorithm, developed by a company and implemented by an academic research hospital, was used to determine patient priority when it comes to receiving treatment. The names of the company and hospital remain anonymous. However, the algorithm consistently prioritized white patients over black patients. These were not just routine treatments, these were treatments for chronic illnesses, including liver and kidney transplants. The algorithm operated by assigning patients risk scores. If a patient’s risk score surpassed a threshold for a particular treatment, they were prioritized for receiving that treatment. However, upon closer examination, black patients who received the same score as white patients exhibited many more critical conditions and

symptoms. Risk scores for black patients were generally lower, causing them to be deprioritized in the queue system.

The most surprising aspect of this machine learning model is that race was not used as a feature in the training set (a feature is a variable or characteristic used to predict an outcome in a machine learning model). However, the algorithm utilized cost as a feature instead. Patients with higher incomes could generally afford to reimburse the hospital for higher costs associated with treatment. On average, people of African descent have lower incomes, are more likely to be uninsured, and are less likely to be able to cover costs of treatment. As a result, African Americans were deprioritized by this algorithm due to their inability to pay. Although this particular case study only discusses African Americans, other minority groups likely face the same problems. In the end, implementation of this biased algorithm decreased the proportion of black patients who received treatment from 50% to 20% (Simonite, n.d.).

Discussion

As demonstrated by both case studies, data used to train machine learning algorithms, not the artificial intelligence infrastructure, is the reason for models going awry. It is not necessarily true that the algorithms themselves are racist, but rather the data being used to train the algorithms are reflective of the sociopolitical atmosphere. In the case of the healthcare algorithm that offered less care to black patients, race was not even used to determine risk scores and assign priority to patients. What mattered to the algorithm was patients' ability to pay. Since income is well correlated with race, black people were given less consideration by default due to systemic bias. Figure 3 from the Economic Policy Institute summarizes the disparity between levels of average income by race.

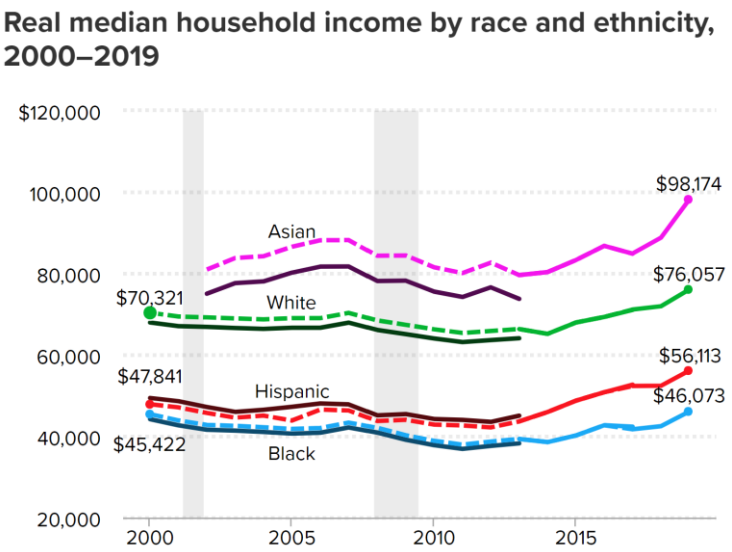

Figure 3: Median household by race (Wilson, 2020).

The healthcare industry's intention of creating a prioritization algorithm roots itself in risk analysis. Physicians are trying to provide care to people who need it most. Ironically, the outcome of the algorithm created unnecessary risk for an entire segment of the United States population. As opposed to viewing machine learning algorithms as tools that curb risk in healthcare, it is urgent that actors in the industry view the algorithms as potential sources of risk. Applying risk analysis to machine learning starts with verifying quality of training data. Considerations include ensuring that there are enough samples, that the samples are representative of the population in terms of demographics, and that data is not compromised by missing values and noise.

Ensuring that training data is picked intentionally in order to develop ethical machine learning models in healthcare is imperative. However, obtaining sufficient data from all demographics is challenging. As a demonstrated by *The DNA We Have Is Too White*, minorities are engaging in risk analysis; they are being skeptical of volunteering their own data because they are unconvinced that they will be able to reap the rewards of their contributions to the scientific community. Lastly, regulation imposed by HIPAA's Privacy Rule makes it difficult to

obtain data that is readily available. The combination of a shortage of representative training data and minority skepticism towards volunteering health data makes HIPAA policy counterproductive to large-scale adoption of machine learning models in healthcare. Two key areas of HIPAA policy should be reconsidered: 1) the necessitation of expert statisticians having to de-identify data for usage and 2) the classification of data relating to a patient's past, present, or future physical or mental health or condition, the provision of health care to the individuals, and demographics as being protected under the Privacy Rule.

There are 38,860 professional statisticians in the United States (*Statisticians*, n.d.). Meanwhile, there are 928,966 active licensed physicians in the United States (*Active Physicians With a U.S. Doctor of Medicine (U.S. MD) Degree by Specialty, 2019*, n.d.). Operating under the assumption that physicians collect health data and statisticians are available to de-identify this data upon request, it becomes apparent that having 24 times more physicians than professional statisticians in the labor market limits the accessibility of data de-identification. Therefore, much of the volumes of data that do end up becoming collected by physicians is not able to be processed for development of training sets to be used in machine learning model development. Although the purpose of requiring expert statisticians to deidentify data is a result of risk analysis, it produces collateral in the form of limited access to machine learning in healthcare. As a result, training should become available to machine learning engineers who are interested in collating training sets to be used in healthcare environments.

Additionally, classifying data relating to a patient's past, present, or future physical or mental health or condition, the provision of health care to the individuals, and demographics as being protected under the Privacy Rule severely limits the potential for leaning technology to make strides across the healthcare field. Eliminating these restrictions would make data all the

more accessible to researchers and useful to the general public. However, it would be reckless to simply disenroll any regulation pertaining to sensitive patient data. The data should be available to trusted and certified professionals who collect the data (physicians) and develop models (machine learning engineers, data scientists, etc.). As the United States and global community enter a data revolution, analytics and machine learning will become increasingly intertwined with healthcare, which stress the importance of cooperation between physicians and data engineers. From the standpoint of risk analysis, creating a joint operation between domain experts and engineers would limit the potential for deployment of problematic machine learning models.

Despite the proposed solution to create a space for collaboration between physicians and engineers, it is important to understand the culture of risk aversion within the United States. When it comes to data collection, risk aversion is apparent on several fronts. The United States government enacts laws like HIPAA to curb the risk of sensitive data being compromised. Minority groups feel that it is risky to surrender their own data due to skepticism. Ironically, the stringent government regulation and systemic bias that deters minorities from volunteering information create an environment that is conducive to the development and deployment of biased, racist, uninterpretable, and non-generalizable machine learning models. If data collection practices are not amended, biased models will widen the pre-existing socioeconomic gap among Americans by favoring the white majority. As shown by the rise of several publications detailing the dangers of implementing biased machine learning models in clinical settings, American society is showing signs of reflexivity – the process of society examining and changing itself while conducting risk analysis.

Ultimately, the healthcare industry needs to pivot its perspective on risk analysis. Currently, the focus is on restricting data to prevent breaches of privacy. After creating a more flexible access to data, focus needs to shift towards analyzing the risk that comes with using incomplete stores of data to train critical machine learning models. Much of the risk pertaining to data quality rests on the fact that the United States has a discriminatory history. Risk mitigation should come in the form of consideration of initiatives like "All of Us" to compile more representative stores of data.

The largest limitation of this research is that even if the proposed changes to HIPAA policy are enacted, systemic bias within the United States healthcare system will persist. Socioeconomic inequalities disproportionately affect people of color. Since these populations are less likely to receive health care in the first place, their data is not collected. The lack of data resulting from discrimination of minorities is beyond the scope of HIPAA regulation and poses a wicked problem that would require copious amount of research and activism. The next step of this research would be to develop a greater understanding of how years of inequality are reflected in the stores of data that are available for model development.

Conclusion:

Development of successful and ethical machine learning models in healthcare can become more of a reality if steps are taken to soften the Privacy Rule while creating a channel of cooperation between physicians who collect data and engineers who develop machine learning models. Although amending United States data privacy laws is a step in the right direction, it is also important to recognize that systemic inequities in the United States healthcare system affect minorities and members of lower socioeconomic classes. Machine learning algorithms are not inherently racist; it is the data that are used to train them that reflects racism, so it is important

for the American healthcare system to examine how its history and laws impact machine learning technologies as the country, and the globe, powers into a digital era.

Works Cited

Active Physicians With a U.S. Doctor of Medicine (U.S. MD) Degree by Specialty, 2019. (n.d.). AAMC. Retrieved March 18, 2022, from https://www.aamc.org/data-reports/workforce/interactive-data/active-physicians-us-doctor-medicine-us-md-degree-specialty-2019

Americans, I. of M. (US) C. on the F. H. C. W. for O. (2008). Health Status and Health Care Service Utilization. In *Retooling for an Aging America: Building the Health Care Workforce*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK215400/

Beck's Sociology of Risk: A Critical Assessment—Anthony Elliott, 2002. (n.d.). Retrieved October 31, 2021, from https://journals.sagepub.com/doi/10.1177/0038038502036002004

CRISP-DM. (n.d.). *Data Science Process Alliance*. Retrieved February 5, 2022, from https://www.datascience-pm.com/crisp-dm-2/

Data, N. R. C. (US) P. on D. C. of R. and E., Melnick, D., & Perrin, E. (2003). IMPROVING RACIAL AND ETHNIC DATA ON HEALTH. In *Improving Racial and Ethnic Data on Health: Report of a Workshop*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK222062/

Eid, M. (2003). Reflexive Modernity and Risk Society. *International Journal of the Humanities, 1*. https://doi.org/10.18848/1447-9508/CGP/v01/58162

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Hoffman, S., & Podgurski, A. (2020). *Artificial Intelligence and Discrimination in Health Care*. 75.

Intelligence, I. (n.d.). *Use of AI in healthcare & medicine is booming – here's how the medical field is benefiting from AI in 2022 and beyond*. Insider Intelligence. Retrieved April 20, 2022, from https://www.insiderintelligence.com/insights/artificial-intelligence-healthcare/

Magazine, S. (n.d.). *The DNA Data We Have Is Too White. Scientists Want to Fix That*. Smithsonian Magazine. Retrieved March 18, 2022, from https://www.smithsonianmag.com/science-nature/gene-bank-too-white-180968884/

Nair, S. R. (2010). Personalized medicine: Striding from genes to medicines. *Perspectives in Clinical Research*, *1*(4), 146–150. https://doi.org/10.4103/2229-3485.71775

Racism, Inequality, and Health Care for African Americans. (2019, December 19). The Century Foundation. https://tcf.org/content/report/racism-inequality-health-care-african-americans/

Rights (OCR), O. for C. (2012, September 7). *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* [Text]. HHS.Gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Simonite, T. (n.d.). A Health Care Algorithm Offered Less Care to Black Patients. *Wired*. Retrieved March 18, 2022, from https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/

Statisticians. (n.d.). Retrieved March 18, 2022, from https://www.bls.gov/oes/current/oes152041.htm

Stein, R. (2017, October 25). Troubling History In Medical Research Still Fresh For Black Americans. *NPR*. https://www.npr.org/sections/health-shots/2017/10/25/556673640/scientists-work-to-overcome-legacy-of-tuskegee-study-henrietta-lacks

Summary of the HIPAA Privacy Rule | HHS.gov. (n.d.). Retrieved March 18, 2022, from https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

Top 10 Applications of Machine Learning in Healthcare—FWS. (n.d.). Retrieved February 5, 2022, from https://www.flatworldsolutions.com/healthcare/articles/top-10-applications-of-machine-learning-in-healthcare.php

U.S. Census Bureau QuickFacts: United States. (n.d.). Retrieved March 20, 2022, from https://www.census.gov/quickfacts/US

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

What is Machine Learning? (2021, November 5). https://www.ibm.com/cloud/learn/machine-learning

What Is the Definition of Machine Learning? | Expert.ai | Expert.ai. (n.d.). Retrieved April 20, 2022, from https://www.expert.ai/blog/machine-learning-definition/