Assessing Underrepresentation in Machine Learning Datasets for Cardiovascular Disease (CVD) Research

by

CJ Dunn, Danilo Correia, and Liam Rathke

Worcester Polytechnic Institute

June 25th, 2022

Approved by Advisor Professor Emmanuel Agu

Index Pg

List of Figures and Tables

Acknowledgements

We would like to thank our Interactive Qualifying Project advisor, Professor Emmanuel O. Agu, who helped us throughout the project period and provided invaluable advice and feedback at every step.

1. Abstract

Cardiovascular diseases (CVDs) are the single most common global cause of death, causing around 30% of all deaths globally. With recent advancements in computer technology, Machine Learning (ML) and artificial intelligence (AI) are now common approaches for CVD research. However, while CVDs manifest differently across race/ethnicity and gender groups, it is unclear whether CVD datasets utilized in research include adequate representation of all races/ethnicities and gender, which could lead to inaccurate results. The research characterizes the representation of various racial, ethnic and gender groups in CVD datasets utilized in CVD ML and AI studies. After analyzing 11 CVD datasets, it was found three datasets which included information on race/ethnicity and gender, all of which were demographically consistent with the US Census. However, the remaining 8 datasets reported on neither the race/ethnicity nor gender of study participants. Additional investigation is necessary to quantify the existence and impact of misrepresentation across demographic groups in CVD research.

In addition to that, a website describing this work was developed, as an easy and accessible way of communicating the content produced during this research. The website went over usability evaluation to ensure that the correct message was being communicated. Participant feedback received was generally encouraging.

2. Introduction

2.1 Background

Cardiovascular diseases (CVDs) include various symptoms, ailments, and illnesses related to heart function. While strokes and heart attacks are well-known CVDs, the disease class also includes many forms of heart disease, and conditions events such as atrial fibrillation. Collectively, CVDs represent the largest cause of death globally; in 2012, an estimated 17.5 million people died as a direct result of CVD [5], rising to 17.7 million by 2017 [3]. The true number could be even larger, since this value is only an estimation from the World Health Organization. In the United States, 30% of deaths are attributed to CVD, and this number rises to 45% for the European Union [4].

Traditionally, CVDs are typically diagnosed, monitored, and tracked by physicians [4] - thus, getting an early diagnosis is contingent on access to a physician with knowledge of CVD warning signs. While there are many risk factors that can predispose an individual to CVDs, studies show that over 70% of cases are directly linked to a subset of causes and comorbidities, such as hypertension, tobacco use, high non-HDL cholesterol, and a poor diet [3]. Other parameters can also reliably predict eventual CVD likeliness: better-educated people are less likely to die of CVDs, and the inverse relationship between education and CVD death rate is stronger than the correlation of wealth/income and CVD diagnosis [3]. In short, it is possible to identify the individuals most likely to succumb to CVD with reasonable accuracy; improving the accuracy of CVD prediction or expanding access to CVD screenings could help larger segments of the population to receive the treatment they need earlier, potentially reducing the substantial CVD death rate.

The data-driven nature of CVD diagnosis suggests a natural link between CVD prediction and a rapidly expanding field in computer science: machine learning. Indeed, with a relatively low barrier to entry, and a large array of CVD-related datasets available for analysis, the number of clinical journals leveraging machine learning algorithms has increased in recent years [6]. As a baseline metric, a search made on May 2, 2022, in the JSTOR research database found 626 articles published in health-focused journals relating to “cardiovascular disease” and “machine learning” between 2011 and 2016, a 53% increase over the 401 articles published between 2001 and 2006.

There are two main practical applications of machine learning in the CVD field [19]. First, machine learning for CVD prediction: collecting datasets including CVD-specific risk factors for populations and creating a model that predicts the individuals in the population most likely to suffer a CVD illness in the future. Many such models exist, and the dataset-driven nature allows for cross-verification across algorithms; for example, an algorithm built from one dataset could be evaluated using a different dataset, validating or invalidating said algorithm’s results, and potentially providing new insights for the second dataset. Machine learning researchers are actively comparing the performance of leading CVD algorithms, with the consensus that machine learning algorithms are better at predicting CVD than other traditional data analyses methods [9]. Second, machine learning for CVD research: using machine learning techniques while researching cures for CVD illnesses. For example, cures for atrial fibrillation, a heart condition that leads to blood clots, are often unreliable, in part because analyzing and interpreting important electrical signals is difficult. Using machine learning, researchers created an algorithm to analyze such electrical signals with over 95% accuracy [2]. In both cases, machine learning techniques offer a promising solution to CVD-related challenges. First, computer algorithms can produce more consistent results, meaning that good machine learning algorithms will be more precise when

predicting CVD. Second, computerized algorithms can be run on large groups automatically, increasing accessibility for and awareness of CVD warning signs.

While the potential for machine learning in medicine, and more specifically, CVD, is great, there are many complications that must be considered. For instance, medical journals employ statisticians to review papers for accuracy before publication, but not all statistical editors are trained in machine learning. Without adequate support for this new technology, machine learning papers submitted to public health journals are not subjected to a usual standard of review [6]. More importantly, there is evidence that CVD warning signs and results vary across racial and ethnic groups [1]. Thus, biased datasets used in machine learning studies could lead researchers to biased conclusions. Although the main focus of many CVD papers leveraging machine learning involves improving algorithms for CVD prediction, instead of applying machine learning directly to research, an algorithm optimized on biased or unrepresentative data could steer public health officials towards suboptimal CVD diagnosis and wrong treatment strategies. In turn, underrepresented groups might receive inadequate diagnoses and care. Underrepresentation of ethnicity groups when compared to the true population data (CENSUS), could lead to skewed algorithms, potentially missing CVDs or having higher error rates for underrepresented groups, and ultimately leading to inequitable treatment for certain demographic groups.

2.2 The goal of the Interactive Qualifying Project (IQP)

- Characterize the demographic breakdown of cardiovascular disease research datasets and determine the extent to which minority groups are underrepresented.
- Review CVD studies to understand underlying factors that might lead to differing representation across demographic groups

2.3 Related Work

Several papers related to CVD and ML identified race and ethnicity as a key factor when predicting cardiovascular disease in individuals. In 2015, researchers discovered that including race and ethnicity significantly improved cardiovascular prediction accuracy (*Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events*) [1]. Two years later, another paper (*Can machine-learning improve cardiovascular risk prediction using routine clinical data?*) found that race and ethnicity were the best predictor of CVD in individuals when using the logistic regression model [5]. However, it, even after such research has been published, race and ethnicity is still not amplified as it should, as it will be discussed over this work.

2.4 Methodology

In this Interactive Qualifying Project, it was aimed to characterize the demographic breakdown of ML datasets utilized in CVD research and understand whether the dataset-driven approach for machine learning research on cardiovascular disease is utilizing biased data with underrepresentation of certain groups. First, a literature review of current materials related to CVD, machine learning, and CVD research using machine learning was conducted. Second, a collection of 11 standard CVD studies with public datasets was established. Third, contact with researchers from CVD and machine learning area to try to acquire additional data that was missing in their studies. Fourth, a data analysis pipeline was created to run statistical analyses on the received CVD datasets. Finally, a compilation of the findings was integrated into an interactive website, creating a more accessible medium to our findings and the issue that we are investigating.

2.5 Summary of Results

Of the 4 datasets collected from 11 studies, only three included race/ethnicity and gender information: the National Health Nutrition, the National Health Nutrition from 2018, and the

Centers of Disease and Control. These three datasets were not demographically consistent with US Census results and showed to be scaled to represent prevalent racial and ethnic groups in the US equally. The remaining, and most popular against our sample, exist. In addition to that, the user testing of the interactive website showed that the final product achieves its objective in transmitting our findings in a more accessible medium.

3. Related Work

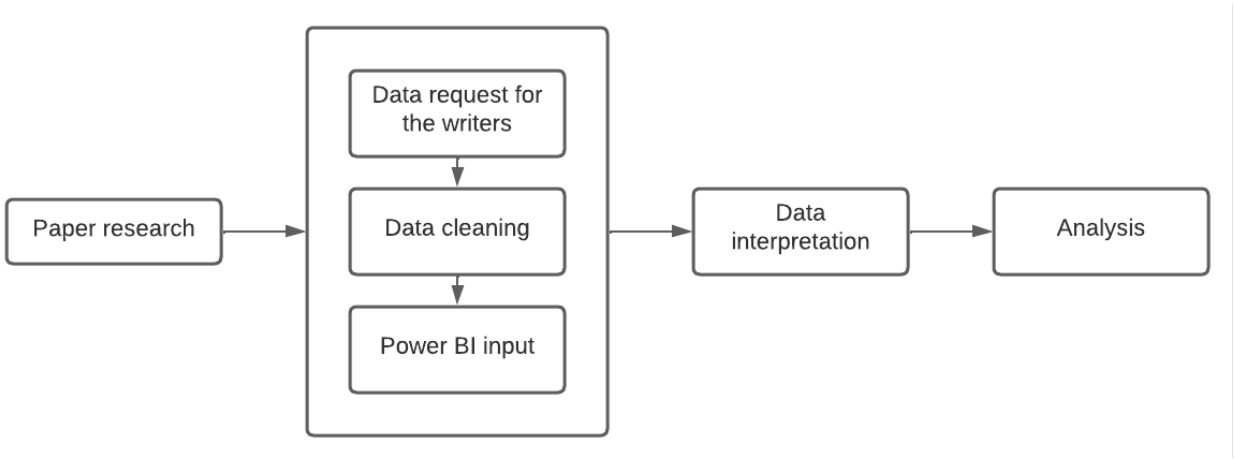
Differences in the way cardiovascular diseases manifest themselves across racial and ethnic groups are already well-known. Specifically, a 2015 study (*Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events*) created a model that predicted cardiovascular risk when considering race and ethnicity alongside other typical heart disease parameters [1]. This model was more accurate than other models that excluded race and solely based their predictions on risk factors. In other words, in order to deliver adequate preventative care for CVD ailments, a nuanced understanding of racial and ethnic background and adequate representation of all groups in ML datasets are critical. Similar to this project, researchers worked with datasets that included additional demographic information, finding racial and ethnic factors extremely useful when making a CVD prediction [26]. However, this study focused on improving CVD prediction algorithms instead of assessing the state of diversity and representation in CVD datasets.

Surprisingly, numerous recent journal articles focused exclusively on CVD prediction using machine learning but have made little or no reference to race or ethnicity at all, in part a result of the common *Heart Disease Data Set*, a dataset collected by the University of California Irvine made specific for ML, which itself does not reference race or ethnicity. However, including race and ethnicity as ML dataset parameters can often yield stronger results [26]. A 2017 journal article [5] analyzed data from a cohort of almost 300,000 to compare the effectiveness of four different machine learning algorithms in CVD prediction: logistic regression, random forest, gradient boosting machines, and neural networks. Ethnicity was the number one predictor for CVD in the logistic regression model, and the number three predictor for CVD using the random forest, gradient boosting machine, and neural network algorithms. Thus, in machine learning studies, racial and ethnic background information is a key component to a good CVD prediction. Like the

previous paper, this study identified the importance of accounting for race and ethnicity under the context of improving prediction algorithms, and did not conduct a comprehensive review of the extent of diversity in cardiovascular disease research using ML.

Much of this research consisted of identifying datasets used within CVD studies; the most relevant datasets are identified and discussed further in section 4.1 (dataset review).

4. Methodology



(Figure 4.1 - methodology flowchart)

This section presents the tasks that were completed in order to accomplish the goal of the research in addition to the methods appropriate to achieve each separate objective. Due to the complexity of factors pertaining to the topic, a multifaceted approach was taken in order to understand the scope of the project. This approach included: 1) Gathering background information by accumulating relevant research papers within the topic of machine learning and artificial intelligence concerning cardiovascular disease. 2) Collecting the datasets used within the accumulated research papers. 3) Email correspondence with authors of accumulated research papers. 4) Developing a data pipeline in order to further break down data. These steps were necessary to develop a complex understanding of the variables affecting our research.

1. Gathering and accumulating research papers. Gathering data began by accumulating studies focused on machine learning as a viable way to further diagnose, perceive, and treat a plethora of different cardiovascular diseases. During the accumulation of datasets, study frameworks such as clinical trials, experimental studies, longitudinal studies, and case studies were all avoided. These frameworks focus on variable manipulation, time variations, or individual progression, whereas the goal of our data gathering was to find the distribution of gender and race

within a dataset. Otherwise, implying that each individual within the dataset was treated the same would be impossible. For this, it was found that correlational studies are the best fit, especially studies that were measuring the difference in using the desired attributes (ethnicity and gender) and not utilizing them in machine learning model trials. Additionally, many studies utilized secondary datasets acquired through national health databases rather than using primary data acquired themselves. After gathering enough papers that could be analyzed within our time frame, it was possible to move into the data gathering of the datasets that were used in the studies found.

2. Collecting the datasets used within the accumulated research papers. During the collection of datasets from machine learning research projects pertaining to cardiovascular disease there were no ethical considerations involved in our choices. The papers we chose datasets from needed to meet certain requirements of being peer reviewed to ensure validity. Additionally, while the reliability of secondary datasets was confirmed by the peer review process as well, truly authenticating medical datasets is troublesome. There are common human errors (encoding) and also negligence (duplicates). Through the process of searching for the best possible dataset, meaning the most commonly used because of its ability to meet authentication standards, it was found that the Cleveland dataset as well as the Spitz Dataset to be the gold standard for machine learning studies relating to cardiovascular disease. Despite these qualifications, neither dataset include the attribute of ethnicity within their scope [14].

3. Email correspondence with authors of accumulated research papers. After gathering enough datasets from the papers that were accumulated in step 1, emails were sent to the authors about their dataset choice and structure, so an analysis regarding the presence of race and gender within these popular studies would be possible. The emails were divided into two different types: dataset request for those papers that did not specify the datasets that were used, and dataset

questions regarding the reason why ethnicity was not part of their scope. With adequate responses and data, a pipeline and analysis started.

4. Developing a data pipeline in order to further break down data. After acquiring a diverse set of datasets from various sources, a data pipeline for analysis. Before analysis, datapoints whose origins were shared were grouped into broader terms. These grouping terms were decided upon after the data points race origins could be traced. For example, “African Americans” were grouped into “of African descent” which was then joined into “Black” after finding this is the most common dataset attribute descriptor. The framework of the data pipeline is a simple comparison breaking down the proportions of race and gender against the proportions of census data for the area of coverage implied by the sample. For samples at the national level, national proportions were used, while for samples on a local scale, local proportions were used. This comparison of experimental versus actual was then further broken down into a range of charts and graphs indicating the representations of our targeted attributes within the data.

Statistical analysis was then done in order to indicate trends for every accounted for attribute. This statistical summary was done for each individual dataset as well as for each attribute by accumulating the range of errors. The purpose of this multifaceted approach was to not only see trends within an individual study, but be able to see trends on a broader scale for gender and race within the field of study.

4.1 CVD Datasets

Dataset	Studies where dataset was used	URL link to the dataset	Description
NHANES 2011-2016	[8], [10],	Link [8] Link [10]	The National Health and Nutrition (2011-

			2016) is an annual survey that collects health and food data from children and adults in the USA. Regarding CVD, it contains multiple data regarding blood pressure, using oscillometric measurements.
NHANES 2018	[10]	Link [10]	The National Health and Nutrition edition of 2018.
CDC	[15]	Link [15],	The Centers od Disease and Control made its own compilation of different datasets related to CVDs, stroke mortality, blood pressure, and etc.
Heart and Disease Data Set	[1], [2*], [3**] [4], [5],	Link [1] Link [2] Link [3] Link [4], Link [5]	The University of California Irvine dataset is one of the most popular CVDs datasets that we found during our research. Its purpose when created was to be used in ML studies.

Note: Not all the papers disclosed their data-source,

Sources marked with "*" disclosed their data-sources through correspondence

Sources marked with "" received data from a source that cited these data-sources**

NHANES 2011-2016:

The National Health and Nutrition (2011-2016) is a survey research program conducted by the National Center of Health and Statistics to assess health and nutritional status of children and adults in the United States [24]. Therefore, this dataset was not originally intended for machine learning. Data was accumulated for medical research purposes, and only later adopted by projects attempting to model CVD in a predictive manner. Although there are dozens of attributes tracked in the dataset such as blood pressure using oscillometer measurements and X-ray data, no studies attempting to make predictive models used more than 11 attributes. The most common attributes included cardiovascular disease indicator attributes and identifying attributes such as age, gender, and race.

NHANES 2018:

The National Health and Nutrition dataset collected from 2017-2018. Equally as the prior, this dataset was also not originally intended for machine learning. It is part of the National Center of Health Statistics survey. As described in the website, the goal of NHANES studies is to produce "vital and health statistics for the Nation." This dataset was created with the same intentions and specifications as NHANES 2011-2016 and was similarly adopted later as a means of making predictive models for CVD. Once again the attributes examined for the purpose of CVD models were CVD indicator variables and identifying attributes such as age, gender, and race.

CDC:

The Centers of Disease Control compiled several datasets from prior research into one database. It was purposefully crafted to be used for artificial intelligence training and modeling. For this reason, the dataset shows an equal distribution of demographics. As a result of the focus on enabling i=i comparisons with US census data, minority groups within this study are massively

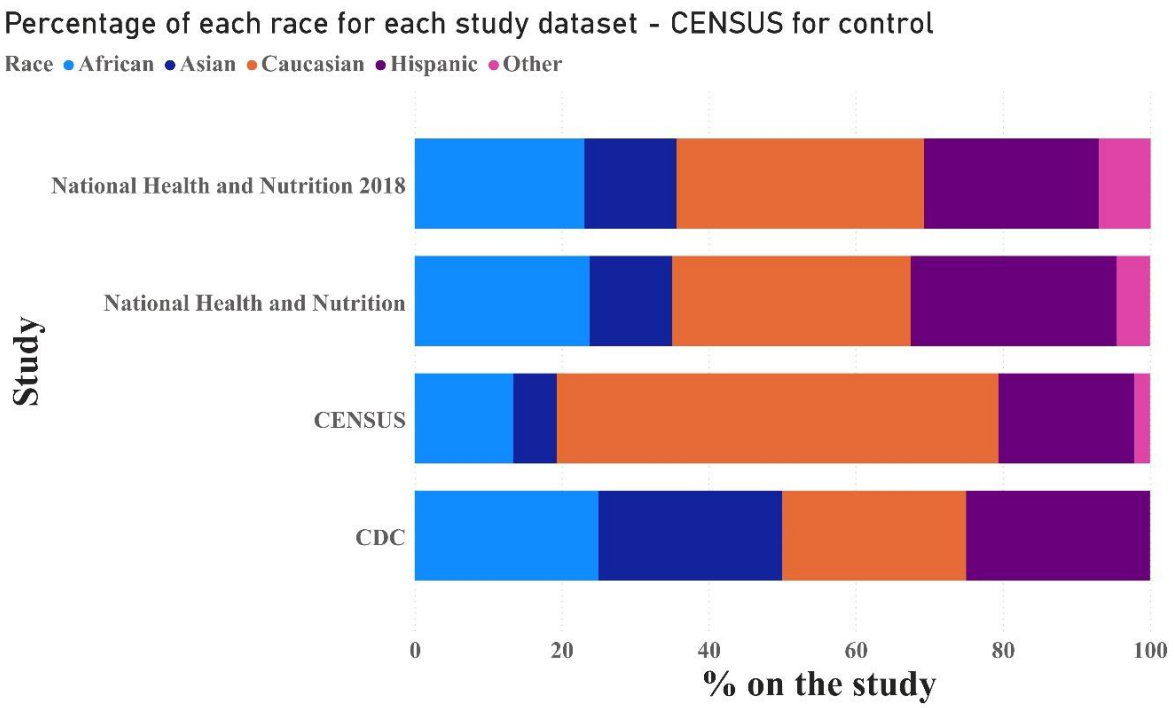
overrepresented. For the purpose of predictive modeling, this data equality allows for the attribute of origins to be a key factor for accurate modeling.

Heart and Disease Data Set:

This dataset is the self-proclaimed most commonly cited data source within studies using machine learning and artificial intelligence to produce models pertaining to cardiovascular disease. This dataset is not included within this paper's data modeling as the attribute of race is not accounted for within this set. This dataset is part of research realized within the V.A. Medical Center, in Cleveland [14].

5. Results

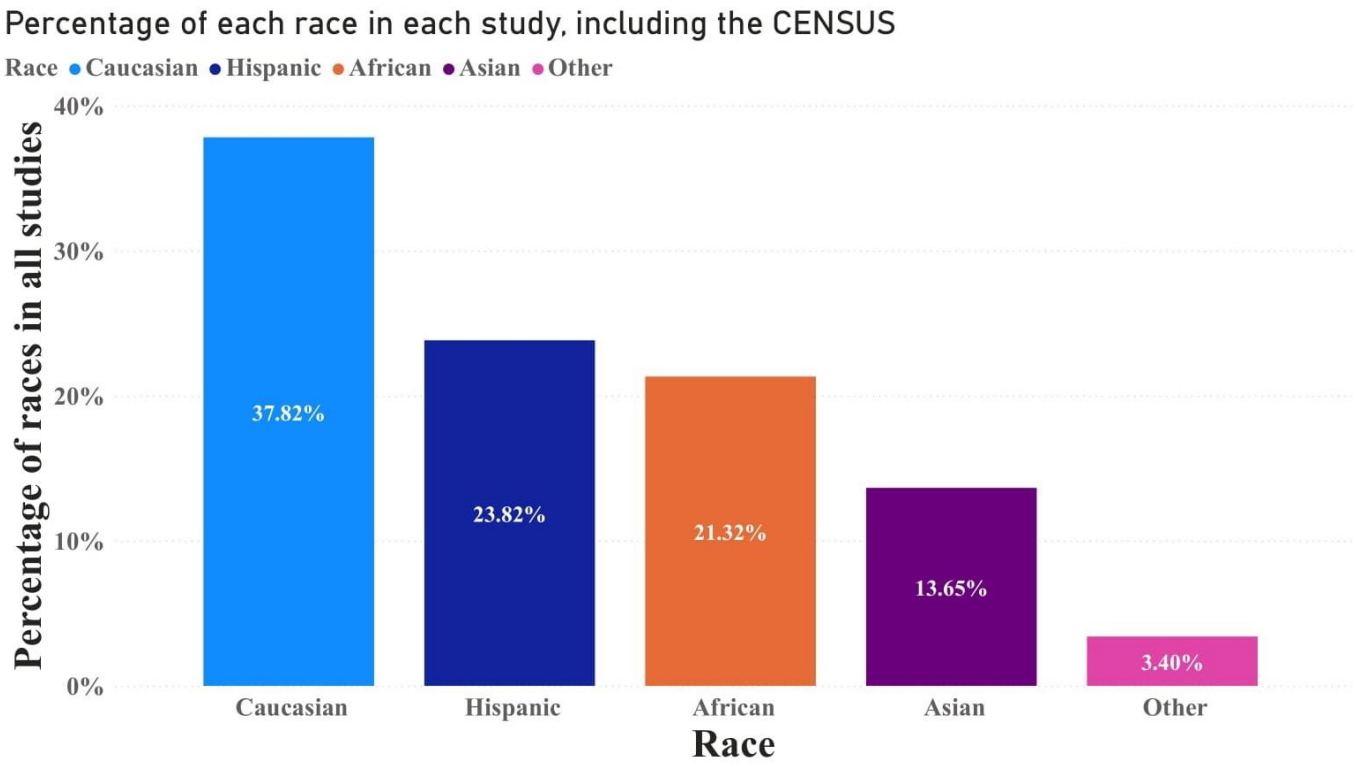
For better analysis and comparison, it was decided to compare the datasets using Power Business Intelligence version 13.0.19635.59 technology to create data visualizations - Power BI is an interactive data visualization software developed by Microsoft, and it is largely used in the industry when making data analysis and interpretation [25]. The goal of the visualizations is to facilitate human perception regarding the difference in proportions for each ethnicity for each dataset, thus relating these differences with society's issues and structure. The results below were generated using Power BI data interpretation and report generator, along with some custom queries and data input.



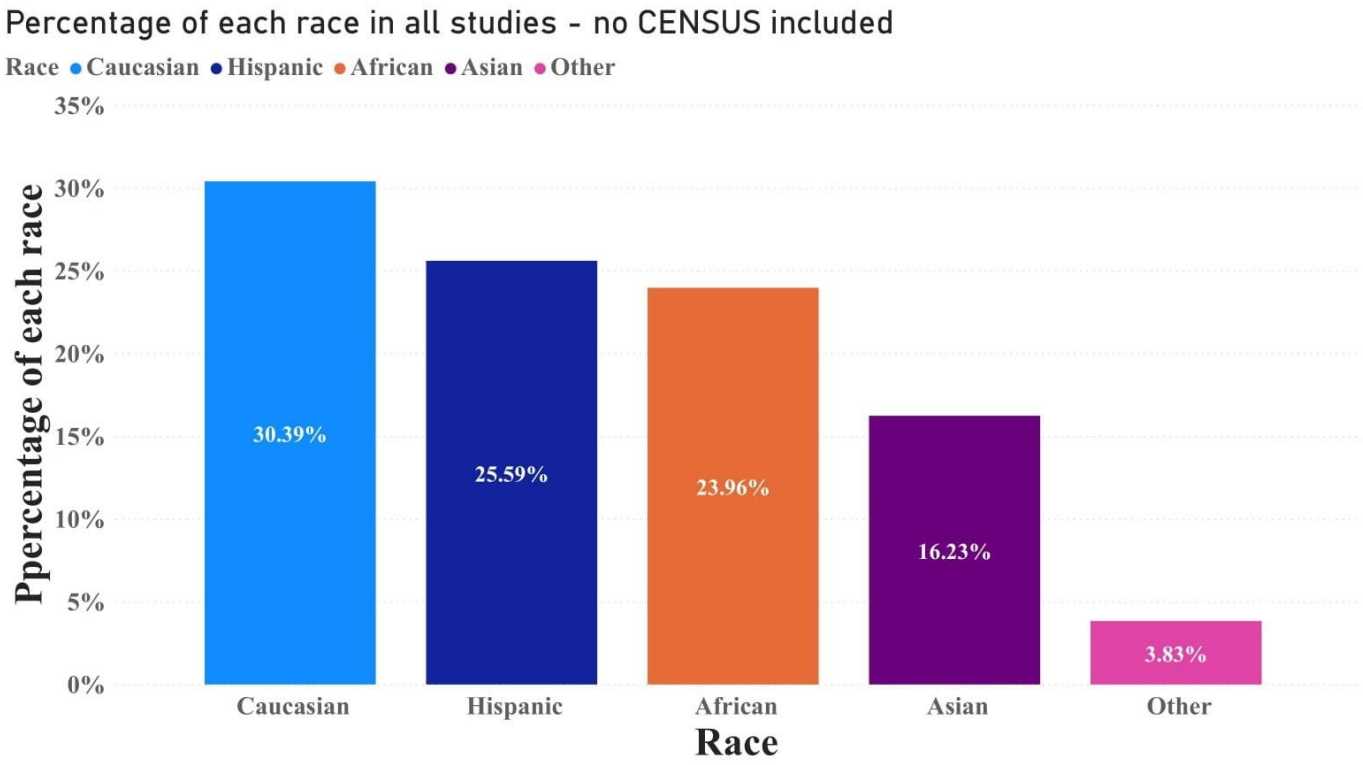
(Figure 5.1 - percentage of each race for each study dataset - CENSUS for control)

In order to identify how studies are statistically biased regarding their datasets, the American CENSUS was used as a control study that represents the real ethnicity distribution in

the US population. It can be noticed that, from figure 5.1, Caucasian and Hispanic populations are the most underrepresented when compared to the general population distribution from the CENSUS. The African descent population, on the other hand, is overrepresented percentage wise. Later in this paper, we will hypothesize on the social, economic, and technical reasons why the datasets are following this pattern of over-representation of these usually underrepresented populations. Figures 5.2 and 5.3 project the average distribution of the ethnicities for all studies, with and without adding the CENSUS to it to the calculation.



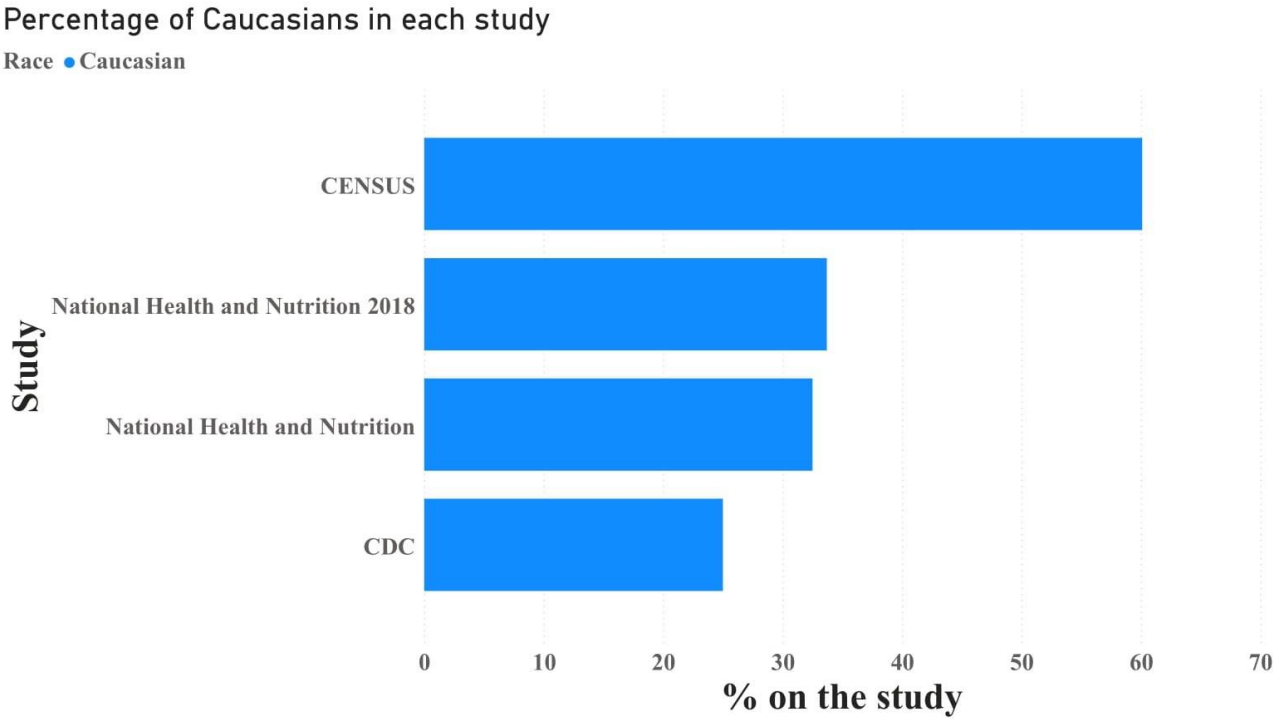
(Figure 5.2 - percentage of each race in each study, including the CENSUS)



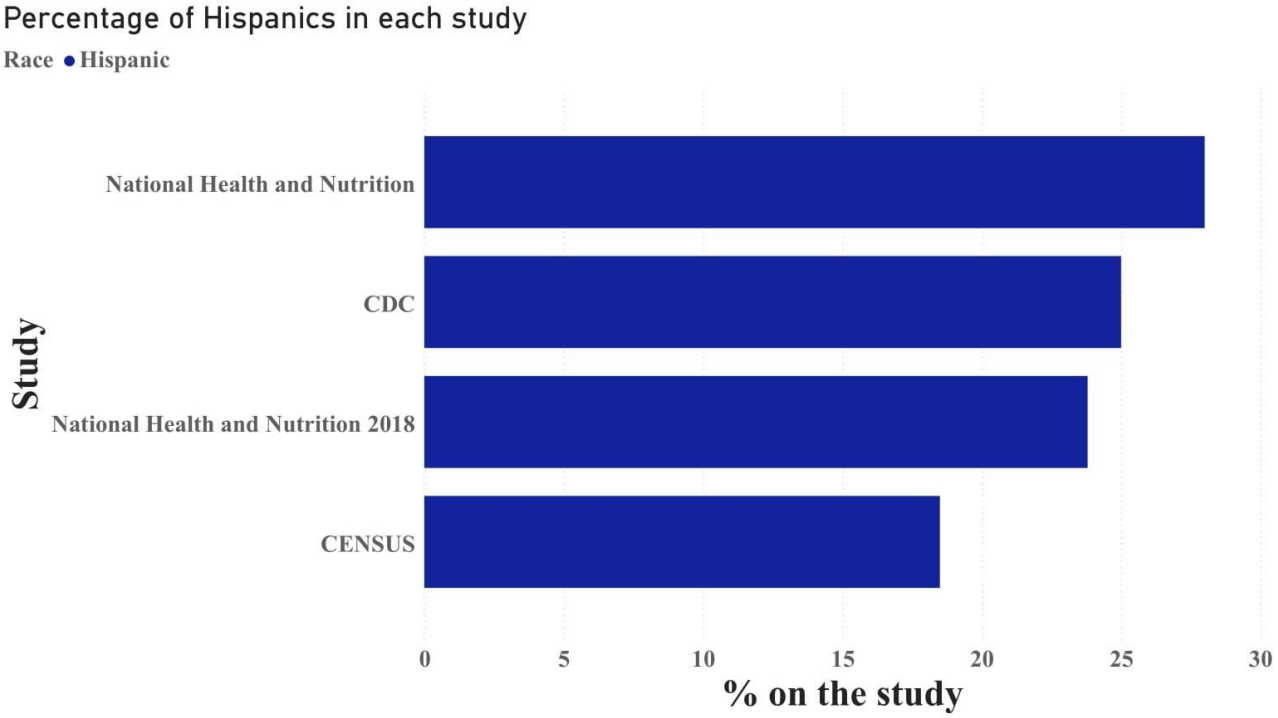
(Figure 5.3 - percentage of each race in all studies - no CENSUS included)

Closely analyzing each population from figures 5.2 and 5.3, it is noticeable that Caucasians constitute more than 35% (figure 5.2), on average, of the sum of populations of all datasets. When the census is taken out of the analysis, Caucasians are now 30.39% of the total population while other races slightly increase (figure 5.3). There is also an almost constant Asian population percentage independently if the census is being considered or not.

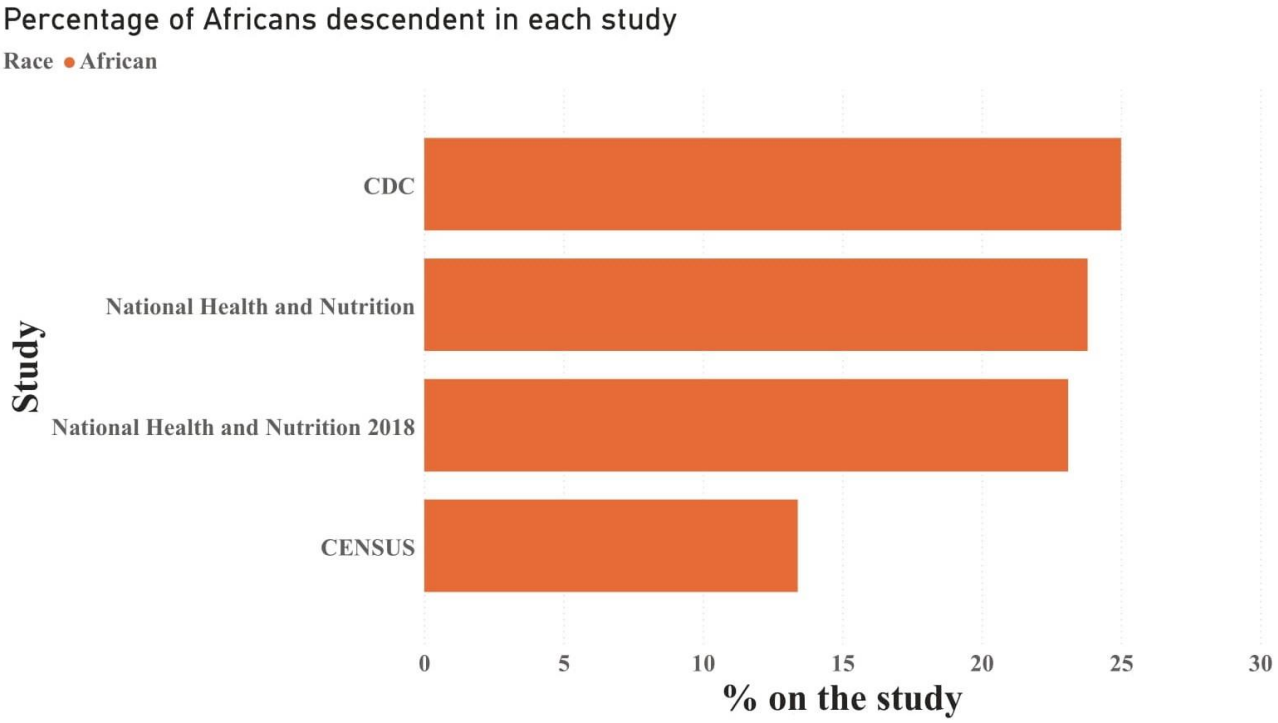
To facilitate better analysis, the percentages for each race for each study can also be broken down:



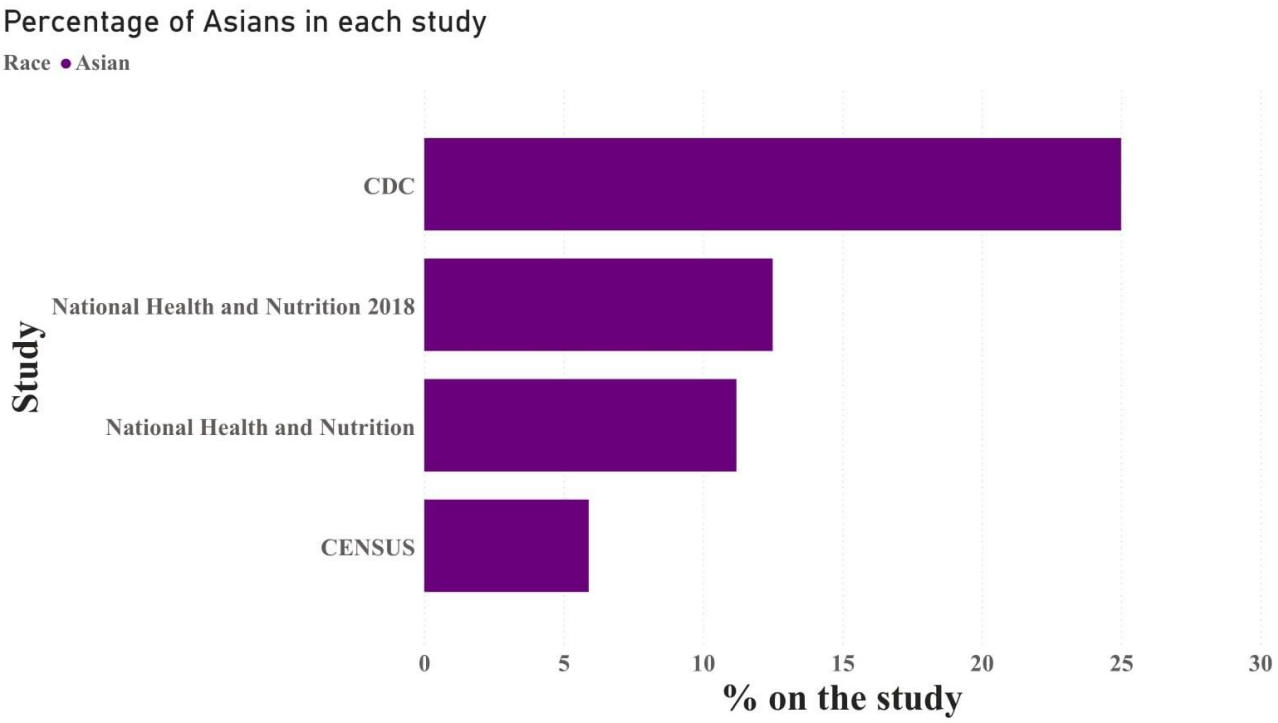
(Figure 5.4 - percentage of Caucasians for each dataset)



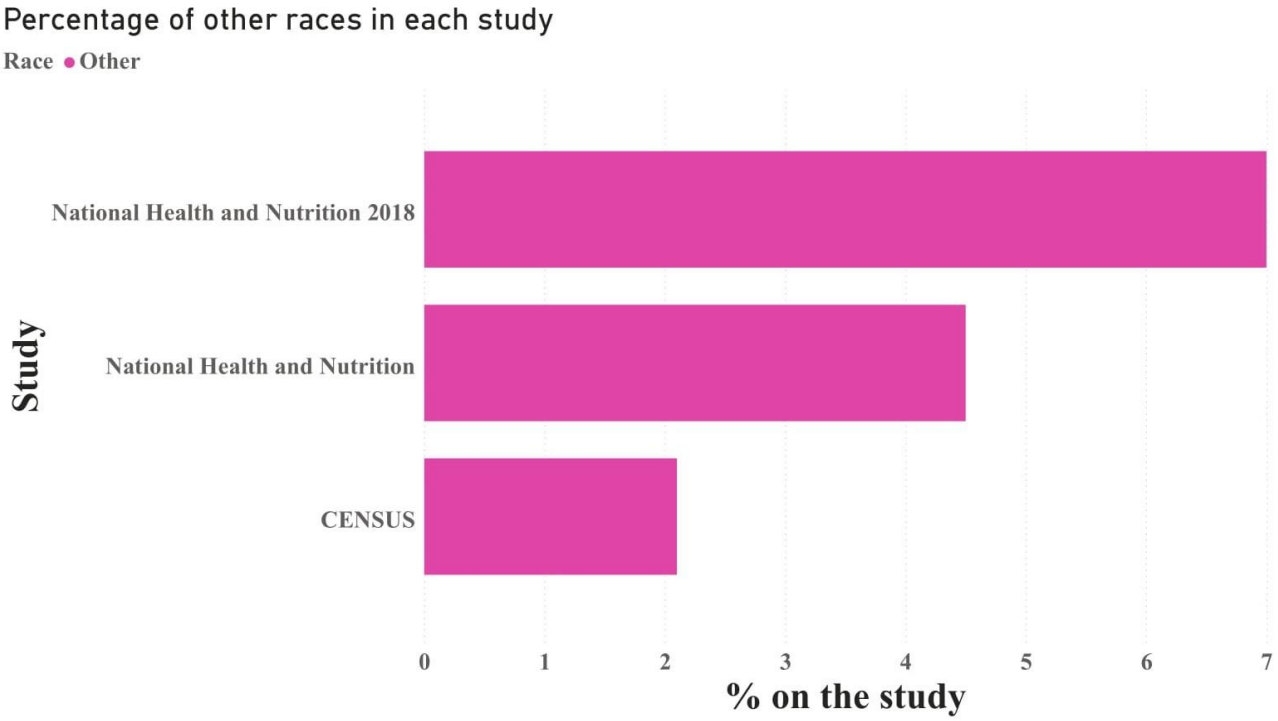
(Figure 5.5 - percentage of Hispanics for each dataset)



(Figure 5.6 - percentage of Africans for each dataset)



(Figure 5.7 - percentage of Asians for each dataset)



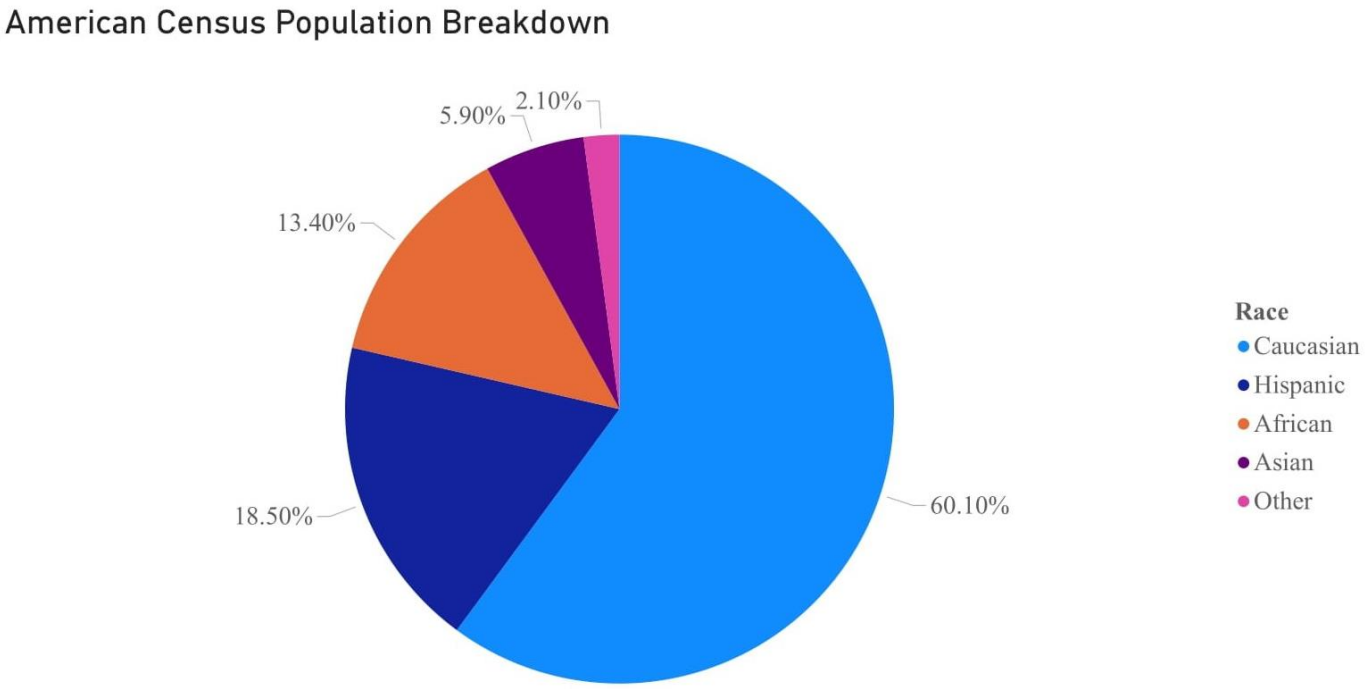
(Figure 5.8 - percentage of subjects of other races for each dataset)

After breaking down the different races for each dataset population, the gap between datasets and the real population is more easily observed. Figure 5.4 showed an average of the Caucasian population per study as 30.1% of the sample, as the real population on census is described as 60.1%. Figure 5.5 shows that the average Hispanic population for the datasets collected is about 25.6%, while the real population from census is portraited as 18.5%. Figure 5.6 shows that the African American population averages 23.9% for the datasets collected, while the census records 13.4%. The Figure 5.7 shows that the average Asian population for the collected datasets is about 16.32%, in the meanwhile the census population is 5.9%. Finally, other races were 5.75% for the datasets, where in the census, they represent 2.1%. For each race, we decided to create the table 5.9 comparing the values of the average percentile used in the studies and the percentile in the US population from the census:

Race	**Average percentile used in the studies datasets**	**Percentage in real population (census)**	**Difference**
Caucasian	30.4	60.1	-29.7
Hispanic	25.6	18.50	7.1
African	23.9	13.4	10.56
Asian	16.23	5.9	10.3
Others	5.75	2.1	3.65

(Table 5.9 - differences in representation of different races)

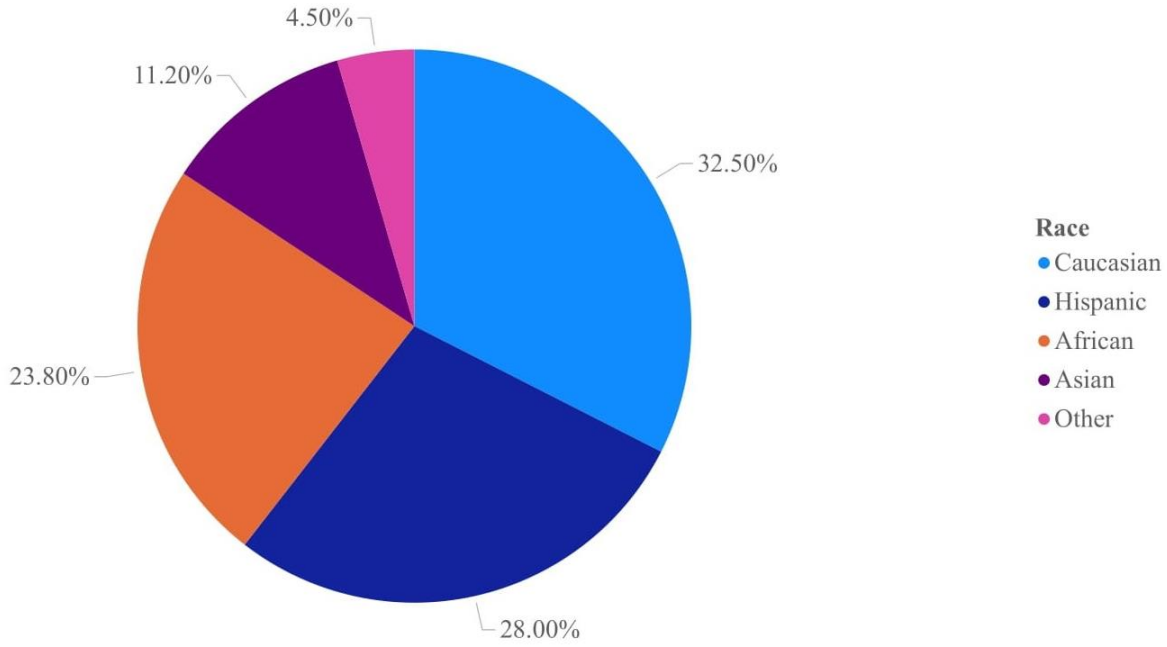
In conclusion, looking at table 5.9, it is evident that there is a tendency to equally distribute the different races for the datasets used for the machine learning models that are mostly used for these types of prediction. This “flattening” is directly correlated with how the algorithm is trained. This will be further discussed in the discussion section.



(Figure 5.10 - American CENSUS for control)

Figure 5.10 is a population breakdown in the United States by race. In this graph “Other” represents a plethora of origins including but not limited to: Pacific Islander, Native American, and multicultural people who identify as such. While these percentages can be represented within this graph the same cannot be said for the graphing of cardiovascular disease datasets which did not contain this information.

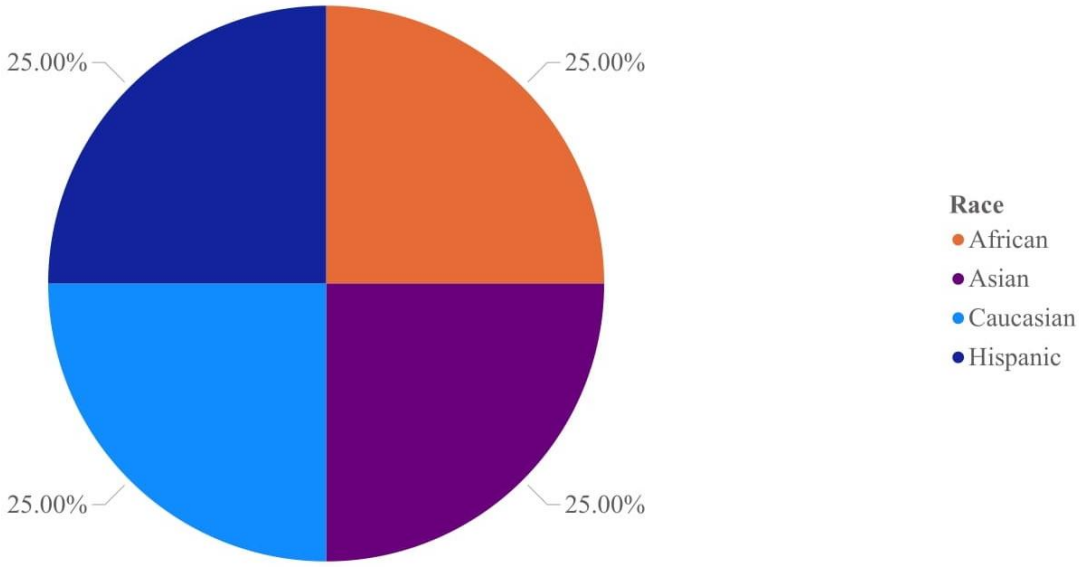
NHANES



(Figure 5.11 - NHANES Study 2011-2016: Race)

This pie-chart is a representation of the races described within the NHANES (National Health and Nutrition Examination Survey) dataset taken from 2011-2016. When compared with the census data for the same time-period, this dataset shows an over-representation of all minority groups by a maximum of 175% (Asians), a minimum of 38% (African Americans), an average of 116.4%

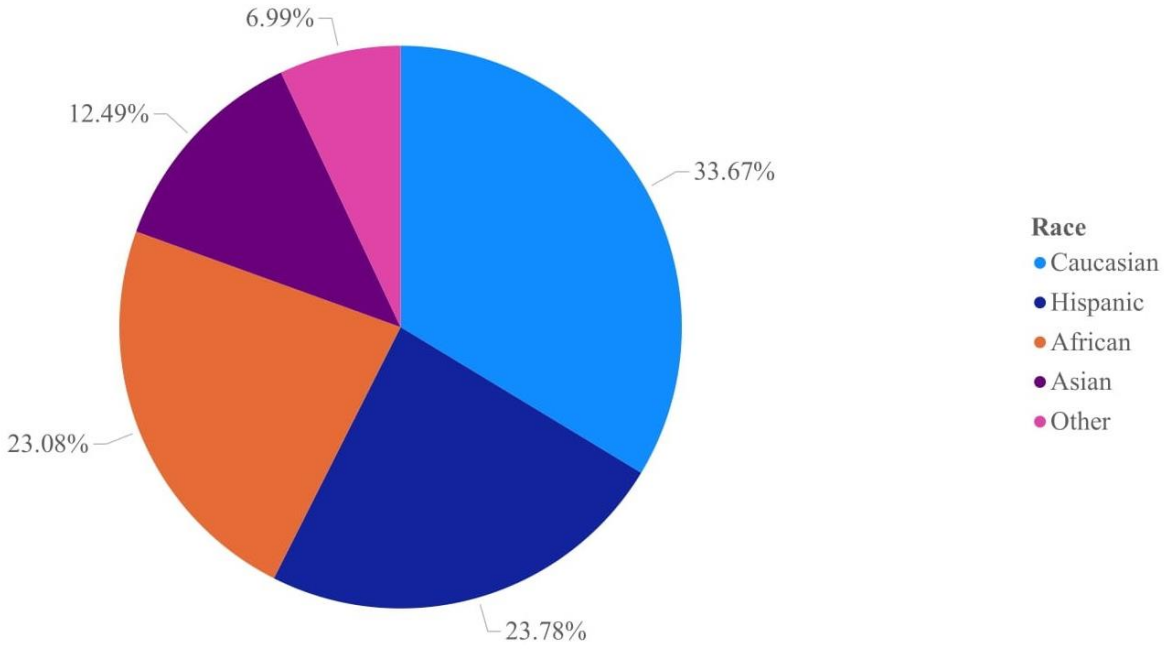
CDC



(Figure 5.12 - CDC Cardiovascular Disease Study 19-20)

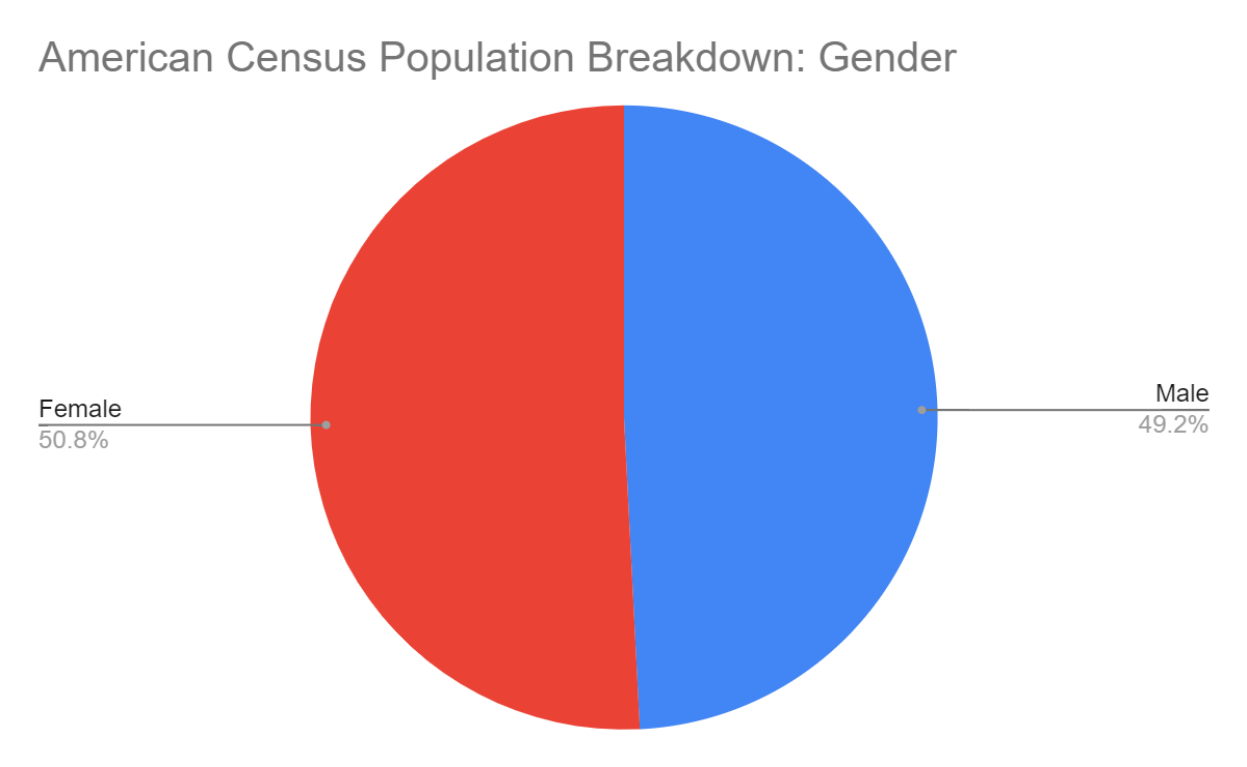
Much like figure 5.3, this study depicts an overrepresentation of minorities within the framework of an American population study. The sample of various ethnicities was stated to have been “made equal on purpose” without further explanation. We believe that this statement implies that this dataset was purposefully crafted with the intention of using it for a type of artificial intelligence model training called classification, which requires all classes to have an equal number of samples.

NHANES 17-18



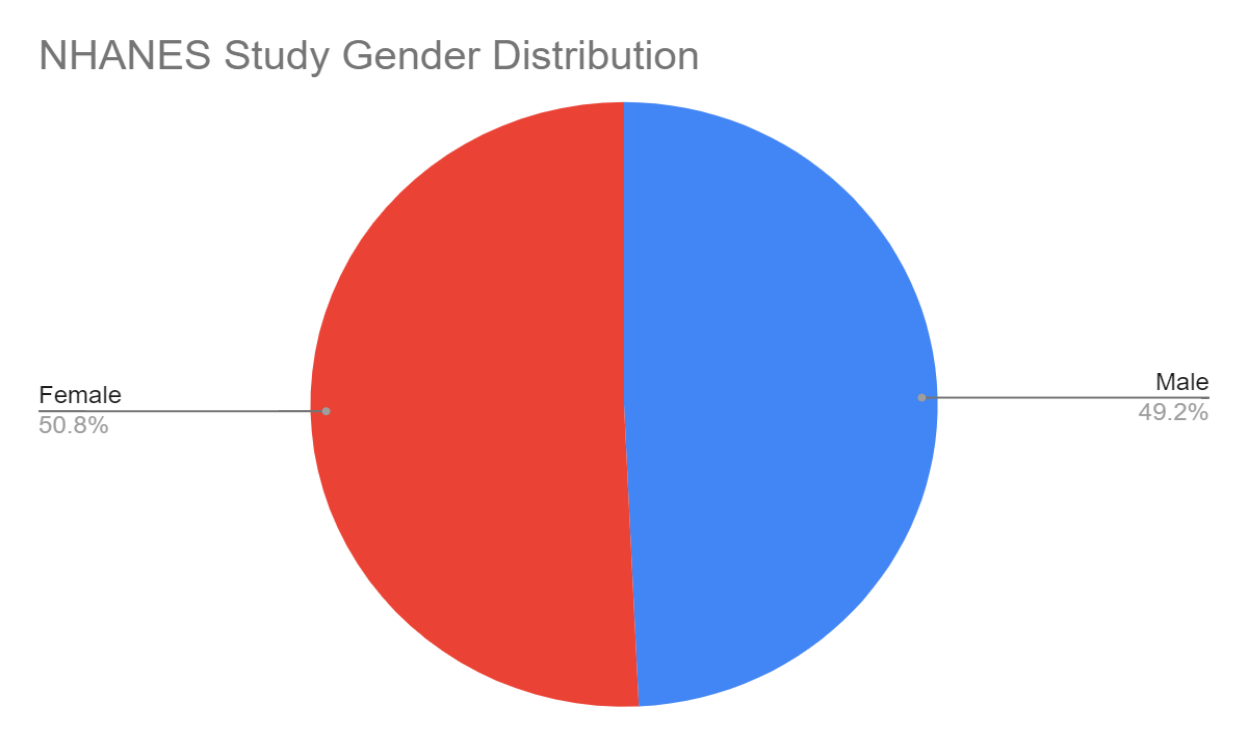
(Figure 5.13 - NHANES Study CVD 17-18)

As shown in figure 5.13, in comparison with census data, The NHANES study CVD 17-18 sample of the same year shows an average over-representation of minority groups at a rate of 112%. This value is skewed by the outlying 232% over representation within the “Other” attribute. This over-representation allows for easier classification within the machine learning models. These models being trained with a “flat” or even representation is stronger than models trained with census consistent data.



(Figure 5.14 - American Census Population Breakdown: Gender: 2016)

Figure 5.6 is the gender distribution of America from the 2016 census. That was used as a control for distribution of samples collected at a national scale during this year.



(Figure 5.15 - NHANES Study Gender Distribution: 2016)

Figure 5.15 is the gender distribution of the NHANES 2016 dataset. The distribution is the exact same as the distribution from the American census of this year. This consistency is valued within the sample when looking to create a predictive model for the other attributes within the dataset. The argument can be made that while this sample represents the national population, it fails to represent the sub-population affected by cardiovascular diseases.

Through investigating the most used heart disease datasets found in machine learning studies the most glaring discovery found was that race and gender go mostly unaccounted for within these datasets and in perpetuity the models themselves. The UCI machine learning repository contains the *Heart Disease Data Set*, which the most commonly cited data source found in machine learning pertaining to CVD [14]. It has 75 attributes, such as cigarettes per day per subject and physical resistance exams such as biking, but only 14 of these attributes (a subset) are typically used within machine learning models. These 14 attributes are: age, sex, chest pain type,

resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and oldpeak. None of these seventy-five or the fourteen attributes pertain to race, although gender was included. Table 5.16 contains the datasets and studies analyzed whose demographic information was found missing:

Study	Dataset	Number of subjects	Reason of the study	Data collected
Genetic Programming for Data Classification. [22]	Heart Disease Data Set	920	The study objective was to present more efficient ways of dealing and resolving decision trees in machine learning.	The study concluded that each node taken (decision) can impact the overall performance. Thus their proposed algorithm showed to be more efficient, bringing a performance gain up to 11%

				when compared to older algorithms.
Diversity in Neural Network Ensembles. [23]	Heart Disease Dataset	920	The study goal was to investigate error diversity in ensembles of classifiers in neural networks of machine learning algorithms using Negative correlation Algorithm.	The study found several paradoxes regarding diversity in algorithms. They also found that using NC in terms of fitness function* can help non inclusive datasets to be more inclusive when running machine learning algorithms.
Prediction of cardiovascular	PhysioNet/Computing in	1297	The goal of the study was to	The study concluded that

diseases by integrating multi-modal features with machine learning methods [10]	Cardiology (CinC) Challenge 2016		create a different algorithm to predict CVDs using electro cardio- gram and phonogram data.	when both data sources are combined in the algorithm, “the genetic algorithm model improves the performance of the initial feature set from 0.890 to 0.934.”
Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients [11]	Own data mining process used.	500	The goal of the project was to prove data refined data mining processes corelating two variables is an effective way of producing data for machine	When diabetes was added to the classifiers, the performance and accuracy of the algorithm had a significant improvement, reaching up to 97.52% precision rate for

			learning algorithms.	positive diabetes cases.
An Automated System for Generating Comparative Disease Profiles and Making Diagnoses	Heart Disease Dataset	920	The goal of the study was to develop an automated diagnose system that could identify a disease with user input.	The tool lets the user "to explicitly specify the disease group and the control group, and thereby focus attention on differences of interest." For diseases as breast cancer, the algorithm had a 93.1% accuracy rate.

(Table 5.16 - summary of studies that did not include demographic information)

*Fitness function: This forces differences among the individuals by adding penalty terms to their fitness functions; a penalty is higher if the individual is similar to others in the population.

6. Discussion

This research analyzed CVD datasets to characterize the diversity of their populations, and highlighted the issues involved with adequate demographic representation within this field. Researchers looking to use medical data to create accurate models are seemingly struggling to use data that meets diversity goals. Three open-source datasets that were utilized to accurately create cardiovascular disease markers as well as their demographic attributes were found. The graphical breakdown provided in the results section shows that some datasets were well suited for machine learning models within this field. The populations were equally distributed facilitating interpretation of data regardless of the statistical strategy used by the machine learning algorithm [18]. These datasets have reinforced that they could be utilized to identify medical markers while also demonstrating that their demographics were adequate. Several minority populations were overrepresented in these datasets relative to the US census.

All the datasets that were found had an under and over representation depending on the population. When comparing the population distribution of the datasets found with the National Census, differences in the percentages of each ethnicity that we highlighted in table 5.9 was noticed. The populations were equally distributed facilitating interpretation of data regardless of the statistical strategy used by the machine learning algorithm [18]. The overrepresentation of minorities is ideal for classification studies of this manner. Yet, a majority of the research being published within this field is still using older instances of data that does not meet the same diversity standards and avoids mentioning demographics at all. This brings us to our next finding.

Data must be scaled in all features for better results in machine learning algorithms. One of the largest causes of bias in machine learning algorithms, especially classification, is bad

distribution of predictive variables that have considerable impact [20]. In order to mitigate problems like that, researchers can both improve the dataset collection into a more inclusive one [20], or train different models for each sub-population as suggested and tested by Ambale Venkatesh that “developed Heart Failure (HF) risk prediction models separately in Black participants and White participants as a strategy to better elucidate the importance of risk factors that may be most relevant to the development of HF across races” [21]. Therefore, the way that the data is organized and distributed directly affects machine learning model results.

Gender was not included in most the datasets found. Within the datasets that were analyzed during this research, gender was not included as a variable for the machine learning models, even though its importance is known in the field [8]. This omitted information or resultant sampling bias may lead to wrong or unexpected outputs from the models [11].

The fifty-seven machine learning models trained on the *Heart Disease Data Set* account for a self-proclaimed “majority of all machine learning endeavors" in this field [14]. Per an accumulation of studies dating back to 2011 describing “profound race-associated disparities among those who are affected by and die from cardiovascular disease”, not publishing demographics within a dataset in the medical field is unexpected [15]. With the general consensus of the medical field being that race and gender play massive roles in health and in perpetuity the health care received by an individual, a dataset that does not disclose its demographics is likely doing so because of its inability to achieve diversity goals within the sample. The *Heart Disease Data Set* was published in 2014 after sampling from the years 2011-2013. Yet today it is still the standard dataset used for modeling without the attributes of gender or race. This could possibly indicate that current research struggles to deal with these attributes as factors within modeling; hence why current models choose to shy away from current data with demographic breakdowns.

Our study had limited access to datasets used in the research paper that were found. Email correspondence was attempted with both the authors of the dataset and the two studies citing this data source most recently, both in the year 2022. After two weeks of waiting, zero responses were received from the studies citing the dataset. Likewise, there was no response from the paper's first author, but within the README the questions posed were answered. According to the author "the data was formed with the intentions of utilizing it for machine learning modeling and finding correlation between cardiovascular disease markers"". There was no mention of race nor gender, yet the author is stating that the data was accumulated for a very specific purpose. One can surmise from this that studies citing this data source with different intentions are doing so knowingly.

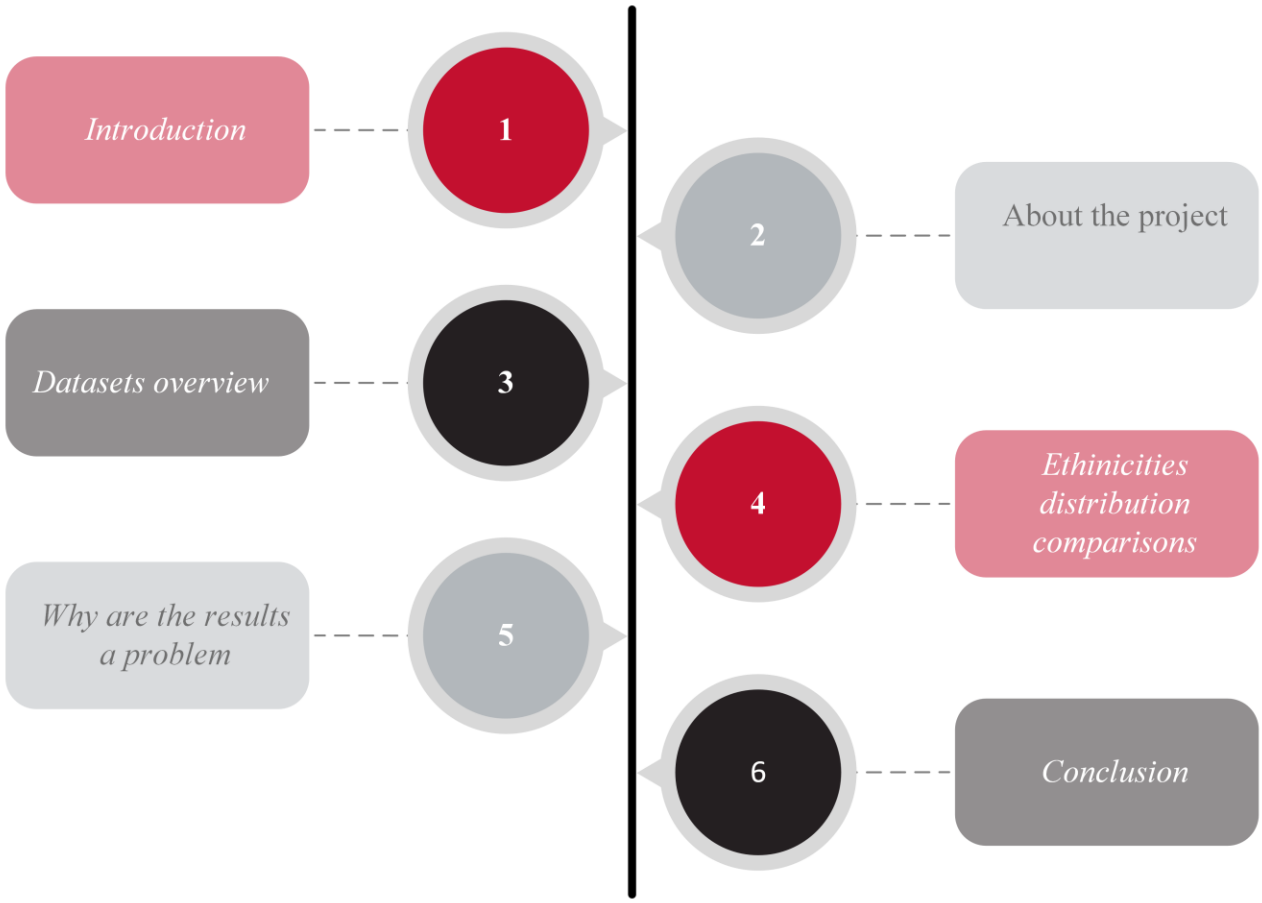
As machine learning algorithms and science progress, it is important to update commonly used datasets. It is important to spread the importance that variables such as ethnicity and gender play in machine learning algorithms within health applications [8][10]. Thus, we believe that it would be of relevance to compile all datasets available in one centralized website where the data can also be analyzed to characterize their demographic breakdown and compare them and ensure adequate inclusion of various demographics..

7. Web Dashboard

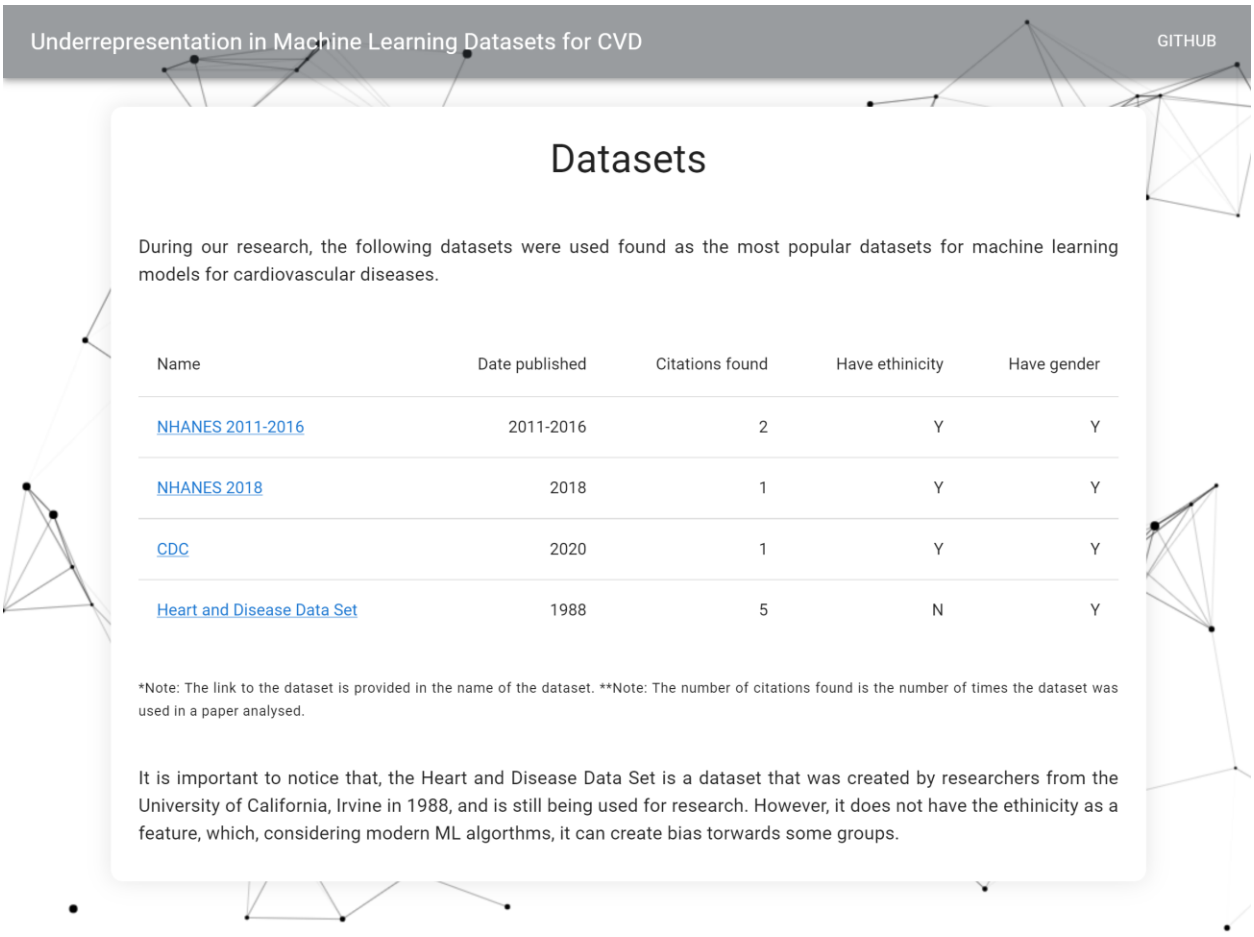
To best display the research within this paper, we created a web-based dashboard. This web platform displays the researched metrics and visuals in an interactive user-friendly medium. The simplistic nature of the platform serves to declutter the research and allow the user an effortless data interpreting experience. An additional goal of the web dashboard is to allow future research groups to easily append further findings as a pseudo-living document as the project develops. The dashboard has both LAN and mobile capabilities.

7.1 Dashboard Implementation Overview

The website has a “show while scroll” architecture. This means that contents are displayed while the user scrolls through the page. In addition to that, content subjects are also divided into components, therefore, each scroll reviews a different content with a different section of the research. Figure 7.1 describes the content flow of the application as the user scrolls, while figures 7.2 and 7.3 portrait examples of a content component on the website.



(Figure 7.1 - website content flow)



(Figure 7.2 - Datasets component on the website)

(Figure 7.3 - Ethnicity distribution component on the website)

Content wise, the website primarily describes the collected data and the importance of a well-distributed dataset for classification machine learning algorithms. The first section explains the project in a paragraph summary. After that, the second section shows the datasets that were analyzed during the research in a table format. Following the datasets, ethnicity distribution over the datasets is established by making use of bar graphs, stacked bar graphs, and radar area charts. All the charts are interactive, and users can filter the data that is being displayed. Finally, accuracy comparisons from algorithms that use race as a feature and those that do not was added to state the importance of the subject. The last section is a conclusion paragraph that summarizes the findings produce in this research.

7.2 Technologies and Infrastructure

The website was developed using the following frameworks, dependencies, languages, and platforms:

1. GitHub
2. Node.js
3. React
4. Material UI
5. ChartJS

A GitHub repository has been set to for collaboration purposes. Node was used specifically for NPM, the node package manager, to build the application at run-time. React, the open-source JavaScript framework is at the core of the development of the dashboard. React was used to build the interactive user interface, chosen specifically for its streamlined developing abilities. Material UI was the main user interface framework applied. ChartJS is a data visualization library that helps developers to build interactive graphs.

Its structure follows a component per view-high architecture. Every section of the website is an independent React component that fills the entire view-high of the page. This makes the code repairability easier and decrease the difficultness to understand it. Figure 7.4 shows a code snippet where the component implements the cascading style sheets property used to fill the entire user's screen.

```
import React from 'react'
import '../App.css';
import { Typography, Grid } from '@mui/material'

function Conclusion() {

    return (
        <div className='small-section' data-aos='zoom-in' data-aos-
delay='300'>// content here
        </div>
    )
}

export default Conclusion
```

(Figure 7.4 - Example of a react component used on the website)

7.2 Evaluating the Usability of the Website

During the development of the website, five users were interviewed to collect feedback regarding the website, and if the content was satisfying its purposes. Questions regarding usability, user interface, and content were asked. Table 7.5 describes the questions that were asked. The entire evaluation and questions were in accordance with the Nielsen Norman Group usability testing guidelines, which is the standard guideline in the market [27]. The subjects were chosen from a group of university students that volunteered to take the usability testing.

Question
Q.1 After you open the page, what comes to your mind about what you have to do?
Q.2 What do you think about the content and what would you change?
Q.3 What is something that you do not understand while interacting with the data?

(Table 7.5 - Questions that were asked during the usability test)

Useful feedback was collected during the tests. After evaluating separately, changes were made into the website. Table 7.6 describes some of the comments collected from different users.

User	**Feedback**
User 1 response to question 2	“The only part that I would change is what is project about. Tell what you did, and that the website is just a summary of the paper”
User 2 response to question 3	“What would be the findings of having the census to compare?”

(Table 7.6 - Key user testing responses)

6. Conclusion

Underrepresentation continues to be a large and long-standing issue in society. In this paper, we investigated the impact of demographic bias in datasets used for cardiovascular research studies. We conducted a survey of recently published works and gained access to industry-standard datasets to compare the diversity in data used for cardiovascular disease research with national averages. Although three of the 11 datasets analyzed as part of the project included demographic information, the remaining seven did not. The prevalence of datasets without demographic breakdowns suggests a worrying trend - that current health research could come to potentially inaccurate conclusions based on the use of biased data.

The impact of demographic bias for cardiovascular disease research specifically is especially significant; it is well-established that many CVDs impact demographic groups differently. Thus, studies which use datasets with unclear or biased representation could come to inaccurate conclusions, which might exclude entire population segments from receiving adequate care. While more research is needed to quantify the exact impact of underrepresentation or misrepresentation in recent CVD studies, it is clear that dataset transparency is necessary to ensure equitable access to and benefits from cardiovascular research leveraging machine learning.

8. Bibliography

[1]C. M. Gijsberts *et al.*, “Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events,” *PLOS ONE*, vol. 10, no. 7, p. e0132321, Jul. 2015, doi: 10.1371/journal.pone.0132321.

[2] M. I. Alhusseini *et al.*, “Machine Learning to Classify Intracardiac Electrical Patterns During Atrial Fibrillation,” *Circulation: Arrhythmia and Electrophysiology*, vol. 13, no. 8, Aug. 2020, doi : 10.1161/circep.119.008160.

[3]S. Yusuf *et al.*, “Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study,” *The Lancet*, vol. 395, no. 10226, pp. 795–808, Mar. 2020, doi: 10.1016/s0140-6736(19)32008-2.

[4]P. Bizopoulos and D. Koutsouris, “Deep Learning in Cardiology,” *IEEE Reviews in Biomedical Engineering*, vol. 12, pp. 168–193, 2019, doi: 10.1109/rbme.2018.2885714.

[5]S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?,” *PLOS ONE*, vol. 12, no. 4, p. e0174944, Apr. 2017, doi: 10.1371/journal.pone.0174944.

[6]K. Singh, A. L. Beam, and B. K. Nallamothu, “Machine Learning in Clinical Journals,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 13, no. 10, Oct. 2020, doi: 10.1161/circoutcomes.120.007491.

[7]J. Coe and M. Atay, “Evaluating Impact of Race in Facial Recognition across Machine Learning and Deep Learning Algorithms,” *Computers*, vol. 10, no. 9, p. 113, Sep. 2021, doi: 10.3390/computers10090113.

[8]A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12592–12594, Jun. 2020, doi: 10.1073/pnas.1919012117.

[9]C. Tesche and V. Brandt, “Calling for a New Framingham,” *JACC: Cardiovascular Imaging*, vol. 14, no. 3, pp. 626–628, Mar. 2021, doi: 10.1016/j.jcmg.2020.12.027.

[10]P. Li, Y. Hu, and Z.-P. Liu, “Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods,” *Biomedical Signal Processing and Control*, vol. 66, p. 102474, Apr. 2021, doi: 10.1016/j.bspc.2021.102474.

[11]G. Parthiban and S. K. Srivatsa, “Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients,” *International Journal of Applied Information Systems*, vol. 3, no. 7, pp. 25–30, Aug. 2012, doi: 10.5120/ijais12-450593.

[12]L. Yahaya, N. David Oye, and E. Joshua Garba, “A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques,” *American Journal of Artificial Intelligence*, vol. 4, no. 1, p. 20, 2020, doi: 10.11648/j.ajai.20200401.12.

[13]M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson, “Ethical limitations of algorithmic fairness solutions in health care machine learning,” *The Lancet Digital Health*, vol. 2, no. 5, pp. e221–e223, May 2020, doi: 10.1016/s2589-7500(20)30065-0.

[14]A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “UCI Machine Learning Repository: Heart Disease Data Set,” *Uci.edu*, 2019. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[15] L. C. Brewer and L. A. Cooper, “Race, Discrimination, and Cardiovascular Disease,” *AMA Journal of Ethics*, vol. 16, no. 6, pp. 455–460, May 2013, doi: 10.1001/virtualmentor.2014.16.6.stas2-1406..

[16]United States Census Bureau, “U.S. Census Bureau QuickFacts: United States,” *www.census.gov*, 2020. https://www.census.gov/quickfacts/fact/table/US/PST045221

[17]CDC, “National Health and Nutrition Examination Survey,” *wwwn.cdc.gov*, Feb. 2020. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm

[18]I.-K. Yeo, “A new family of power transformations to improve normality or symmetry,” *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[19]Shameer, Khader, et al. “Machine Learning in Cardiovascular Medicine: Are We There Yet?” Heart, vol. 104, no. 14, 2018, pp. 1156–1164., https://doi.org/10.1136/heartjnl-2017-311198.

[20]Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (DIS)contents: A survey of dataset development and use in Machine Learning Research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

[21] Ambale-Venkatesh, Bharath, et al. “Cardiovascular Event Prediction by Machine Learning.” *Circulation Research*, vol. 121, no. 9, 2017, pp. 1092–1101., https://doi.org/10.1161/circresaha.117.311312

[22]Eggermont, Jeroen, et al. “Genetic Programming for Data Classification.” *Proceedings of the 2004 ACM Symposium on Applied Computing - SAC '04*, 2004, https://doi.org/10.1145/967900.968104.

[23]Johansson, U., Lofstrom, T., & Niklasson, L. (2007). The importance of diversity in neural network ensembles - an empirical investigation. *2007 International Joint Conference on Neural Networks*. https://doi.org/10.1109/ijcnn.2007.4371035

[24] Wikimedia Foundation. (2022, October 31). National Health and Nutrition Examination Survey. Wikipedia. Retrieved December 14, 2022, from https://en.wikipedia.org/wiki/National_Health_and_Nutrition_Examination_Survey

[25] Microsoft Power BI - Visualization. (n.d.). Retrieved December 15, 2022, from https://app.powerbi.com/view?r=eyJrIjoiN2M1MGE4MmEtNWViZi00NDhjLWI5OGUtOTFmYmMzOGY1OGM3IiwidCI6IjQ3ZmYyZWI3LWFmYjMtNDJjOS1iNGUxLTc4NGRhMTlhOGNhNyJ9

[26] Segar, M. W., Jaeger, B. C., Patel, K. V., Nambi, V., Ndumele, C. E., Correa, A., Butler, J., Chandra, A., Ayers, C., Rao, S., Lewis, A. A., Raffield, L. M., Rodriguez, C. J., Michos, E. D., Ballantyne, C. M., Hall, M. E., Mentz, R. J., de Lemos, J. A., & Pandey, A. (2021). Development and validation of machine learning–based race-specific models to predict 10-year risk of heart failure: A multicohort analysis. Circulation, 143(24), 2370–2383. https://doi.org/10.1161/circulationaha.120.053134

[27] World Leaders in Research-Based User Experience. (n.d.). Usability testing 101. Nielsen Norman Group. Retrieved December 15, 2022, from https://www.nngroup.com/articles/usability-testing-101/.