CareData: Deriving a Representative Dataset with Benchmarks for Machine Learning in Healthcare

CS4991 Capstone Report, 2024

Albert Huang
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
kfa7fg@virginia.edu

ABSTRACT
Machine learning engineers that produce tools in the medical field to be used for patient care seek representational data and methods for identifying biased models to mitigate discriminatory tools. To address this issue, I propose efforts to create a universal dataset for Americans that effectively captures unbiased healthcare data along with benchmarks that denote equitable performance from tools. To create this dataset, I suggest gathering data from underrepresented groups in a transparent manner with augmentation and preprocessing to increase their quality. A benchmark can then be constructed with this dataset and a fine-tuned large language model to quantify the effect of bias on model performance. If the dataset and benchmarks are constructed, I expect to see improvements in both diagnosis and treatment determination by these tools on underrepresented groups. If successful, techniques used on this dataset could be applied to other fields that have unrepresentative data for machine learning.

1. INTRODUCTION

Healthcare prices have skyrocketed from $146 to $10,739 on average per person over the last 60 years and it has been estimated that a quarter of this spending is wasted (Khanna, et al., 2022). To combat the high costs of healthcare and its inefficient usage, companies turn towards leveraging machine learning (ML) models to create more efficient and effective systems. The possibilities for these artificially intelligent tools are endless: analyzing patient data to detect those at risk for illnesses, determining personalized treatment options, and streamlining healthcare processes. These tools need to be reliable and useful for all demographics by providing fair and precise decisions for users.

In the current age of big data, it is now possible to supply ML models with vast amounts of information generated by the healthcare industry to produce tools. It is essential for this data to represent the populace it intends to serve. A model learns the relations in the data it is trained on, which it utilizes for making predictions. Developers can attempt to rework a model's architecture or further align a model to data but will not be able to overcome the inherent property that data is a limiting factor on model performance.

In healthcare, where decisions have direct, life-altering impacts, it is crucial for the model to use balanced and representative data free from bias to ensure fair and accurate outcomes. High-quality dataset creation is both time and effort intensive, however, and developers with little resources are forced to buy datasets. Datasets are a lucrative product, and the commercialization of this commodity has led to fierce competition among different procurers who may cut corners (Alberto, et al., 2023). Taking shortcuts compromises the quality of the data, leading to skews that can reflect healthcare barriers for certain

demographics. Left unchecked, the skewed data and resulting biases will cause the model to treat different demographics in different ways, leading to a flawed tool that perpetuates racism.

2. RELATED WORKS

In 2019, a landmark study found that a commercial ML tool that identified high-risk patients to appropriately allocate resources was discriminatory towards black patients. When algorithmic bias was corrected by replacing flagged healthy white patients with less-healthy black patients until the level of health was relatively the same, the fraction of black patients marked as high-risk tripled from 17.7% to 46.5%. These black patients were also substantially less healthy than white patients in health markers like diabetes, high blood pressure, renal failure, cholesterol, and anemia (Obermeyer, et al., 2019). The disconnect between patients needing healthcare and receiving the healthcare shows both the prevalence and severity of racial bias embedded in existing datasets and tools.

Current dataset developers have adopted a *laissez-faire* attitude towards collecting data, believing that a larger dataset results in more data for a model to train on and consequently better model performance (Paullada, et al., 2021). These developers embrace a mentality of "anything goes" and ingest all the data they can find. To commercialize the data before competitors, they rush to label the dataset as quickly as possible, using numerous annotators, human or machine, to generate ground-truth labels for data points. This negligent approach to data collection produces massive datasets that cannot be effectively analyzed to ensure they contain only valuable and accurate information.

Researchers have attempted to address biases in datasets and still make use of skewed data. One approach is to either enhance the data or model design to benefit at-risk groups. In Draghi, et al. (2024), synthetic data is generated to supplement difficult-to-predict samples in the dataset. While the generations helped address bias in the synthetic dataset, the data results were not representative of a real data distribution. Another approach is to reduce healthcare inequalities by targeting factors that cause individuals or groups to be exempt, like barriers for capturing and digitizing relevant health data (Arora, et al., 2023). However, limited resources make it particularly difficult to overcome these systemic barriers in the areas where they are rooted the deepest.

3. PROPOSAL DESIGN

This report aims to use a data-centric approach to develop a universal dataset along with benchmarks, allowing developers to confidently develop and test models that can make healthcare decisions without being influenced by biases.

3.1 Problem Formulation

Obermeyer, et al. (2019) attributed the risk-prediction tool's discriminatory performance to be based on the model developer's assumption that a patient's future healthcare needs were representative of a patient's future health care needs. This premise led to the tool disregarding poor patients that faced barriers for receiving healthcare and resultantly spent less on healthcare.

Problem formulation is a common dilemma that arises in areas involving the use of data. The tasks for which developers are trying to provide a solution are often amorphous and do not have set indicators or features to make predictions on. It is then up to developers to engineer how the data will get manipulated and used by the model.

By choosing to use healthcare costs as a predictor for healthcare needs, developers based the foundation of their tool on a flawed relationship. Choosing the label a model will predict is the most important decision when

producing a tool and it is not always intuitive. It takes great thought to choose labels that do not perpetuate biases due to the separation between the abstract optimal prediction and labels that represent structural inequality.

To identify the target health indicators I wish to address with my dataset, I will conduct a literature review on similar healthcare issues experiencing racial bias. I will target instances of race correction in medicine that have concrete indicators for health metrics, which will provide an objective basis to compare biased predictions to. A few possibilities include the American Heart Association's Heart Failure Risk Score, which gave higher risk of death to non-black patients, and a study that found black patients were given scores for better kidney function, which resulted in delayed treatment (Tsai, et al., 2021).

3.2 Data Collection and Synthesis

Once the core features to be predicted by the dataset are identified, I will begin gathering and consolidating the data into a format suitable for ML models. Following methods outlined by Towse, et al. (2021) for creating high-quality "open" data, I will collect data from healthcare providers that serve diverse areas in a transparent manner. This dataset will be hosted on an accessible site, welcoming insight from experts in healthcare from different demographic backgrounds. The focus for finding data will be to increase data diversity beyond race, including age, socioeconomic status and disabilities. The data will encompass all known metrics that can be used to determine the target health indicators.

Current methods of interchanging the labelers of a dataset lead to the correctness of the data becoming distorted by inconsistent judgment. I will address this by outlining a general objective label system for each task that volunteer researchers will follow. If there is any uncertainty about what a label should be, annotators will be encouraged to discuss with peers to reach a consensus about the ground-truth label.

To further target the factors that cause individuals or groups to be exempt from healthcare, I will spearhead workshops that raise awareness and educate healthcare practitioners on stigmas that certain groups may have towards seeing doctors. The dissolution of the belief that doctors are discriminatory will lead to larger numbers of marginalized groups receiving healthcare and, consequently, more data on the struggles they face. I will also utilize methods proposed by Draghi, et al. (2024) to augment the dataset by generating synthetic data to target demographics with little presence in current healthcare data. To address the issue of an artificial data distribution, the synthetic data will be generated on real patients to protect sensitive information. By addressing privacy concerns, valuable data from underrepresented groups will be available to add to the dataset.

3.3 Preprocessing and Data Analysis

Once the data has been collected, the dataset will be preprocessed to further increase the quality of the data. Common preprocessing techniques like scaling the data to a bounded distribution to assist ML techniques sensitive to feature magnitude and binning continuous data into categories to address skews in data will be applied. I will also use techniques of feature engineering, like feature selection and feature extraction to create a more robust dataset. Feature selection finds the most useful features, while extraction creates new features by combining existing ones.

An important stage of preprocessing is thorough data analysis that removes both outliers and bad data. Hastily collected data frequently contains irrelevant information that can hurt the model's generalizations. ML models are black-box and the opaque nature of models is why researchers are only able to observe conclusions models draw without understanding why those decisions are being

made. By both discarding harmful data and using feature selection, I will create a pure dataset where misconstrued features that perpetuate discrimination are omitted.

3.4 Benchmark Creation

To help developers gauge the degree to which existing ML tools are affected by bias, I will create benchmarks and metrics to measure the potential influence of bias on a model's performance. I will explore two possible approaches: evaluating the current datasets on which developers are training their tools and evaluating the ML models themselves on the dataset proposed above.

To evaluate current datasets for inherent biases, I will develop a diversity metric called the "Inclusive Score." Using human annotators, a collection consisting of the universal dataset created in this project, well-regarded open-source datasets, and commercial datasets will be given a numeric grade for inclusivity on a range from 1 to 10. This Inclusive Score will be based on a series of guidelines that have high indication of if a dataset may perpetuate racial bias. An existing large language model will then be trained and aligned on the datasets and their respective Inclusive Scores. Developers can then use this fine-tuned model to generate Inclusive Scores for datasets, offering a comparison of commercial dataset scores to benchmark scores of open-source datasets that are considered diverse and bias-free.

Evaluating the extent to which ML models are swayed by bias follows a similar procedure. Using prompt design, an existing large language model can compare an ML tool's predictions to the ground-truth labels and assign a numeric grade for the magnitude to which bias has affected the performance. I will then create benchmark scores for current unbiased tools for performances of future ML tools to be compared against as a measure for how they align with objective health predictions.

4. ANTICIPATED RESULTS

After the dataset and benchmarks are created, applicable ML tools that predict similar health indicators to those in the dataset can be scored using the fine-tuned models developed for benchmarking. These scores, when compared to benchmarks, help developers gauge the degree to which their ML models are being influenced by bias. The difference in performance may also reveal actionable insights on how to further improve the tools to serve demographics better. This project's aim is for improvements in both diagnosis and risk determination by tools on underrepresented groups. This project will also facilitate the development of more inclusive and accurate ML tools.

5. CONCLUSION

To foster fair and accurate ML tools, I propose developing a dataset and benchmarks to evaluate whether a model is learning discriminatory patterns and making equitable predictions across all demographics. This project addresses the need for universality in healthcare and provides developers with a resource to tackle unique healthcare challenges faced by diverse populations. CareData will promote inclusivity through more accessible and affordable healthcare for everyone and will be the basis for future research exploring algorithmic bias.

6. FUTURE WORK

To determine the feasibility of the proposed project, I will explore accessible healthcare datasets and experiment with creating a toy dataset. The toy dataset will be used to simulate the effect of constructing and preprocessing the dataset along with the effectiveness of benchmarking methods on a smaller scale. If the dataset and benchmark prove helpful when fabricated, it may be useful to continue expanding the dataset to cover other predictable health indicators to be applicable to more ML tools. The methods

outlined in this proposal might also be used to address similar issues, like gender data gap.

REFERENCES

Alberto, I. R. I., Alberto, N. R. I., Ghosh, A. K., Jain, B., Jayakumar, S., Martinez-Martin, N., McCague, N., Moukheiber, D., Moukheiber, L., Moukheiber, M., Moukheiber, S., Yaghy, A., Zhang, A., & Celi, L. A. (2023). The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet. Digital Health*, *5*(5), e288–e294. https://doi.org/10.1016/S2589-7500(23)00025-0

Arora, A., Alderman, J. E., Palmer, J., Ganapathi, S., Laws, E., McCradden, M. D., Oakden-Rayner, L., Pfohl, S. R., Ghassemi, M., McKay, F., Treanor, D., Rostamzadeh, N., Mateen, B., Gath, J., Adebajo, A. O., Kuku, S., Matin, R., Heller, K., Sapey, E., & Sebire, N. J. (2023). The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, *29*(11), 1–10. https://doi.org/10.1038/s41591-023-02608-w

Draghi, B., Wang, Z., Myles, P., & Tucker, A. (2024). Identifying and handling data bias within primary healthcare data using synthetic data generators. *Heliyon*, *10*(2). https://doi.org/10.1016/j.heliyon.2024.e24164

Khanna, N. N., Maindarkar, M. A., Viswanathan, V., Fernandes, J. F. E., Paul, S., Bhagawati, M., Ahluwalia, P., Ruzsa, Z., Sharma, A., Kolluri, R., Singh, I. M., Laird, J. R., Fatemi, M., Alizad, A., Saba, L., Agarwal, V., Sharma, A., Teji, J. S., Al-Maini, M., & Rathore, V. (2022). Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. *Healthcare*, *10*(12), 2493. https://doi.org/10.3390/healthcare10122493

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

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. ScienceDirect. https://doi.org/10.1016/j.patter.2021.100336

Towse, A. S., Ellis, D. A., & Towse, J. N. (2021). Making data meaningful: guidelines for good quality open data. The Journal of Social Psychology, 161(4), 395–402. https://doi.org/10.1080/00224545.2021.1938811

Tsai, J. W., Cerdeña, J. P., Goedel, W. C., Asch, W. S., Grubbs, V., Mendu, M. L., & Kaufman, J. S. (2021). Evaluating the Impact and Rationale of Race-Specific Estimations of Kidney Function: Estimations from U.S. NHANES, 2015-2018. EClinicalMedicine, 42, 101197. https://doi.org/10.1016/j.eclinm.2021.101197