Team

GenHealth

Video Link
https://www.youtube.com/watch?v=G6YzAoNFVvY

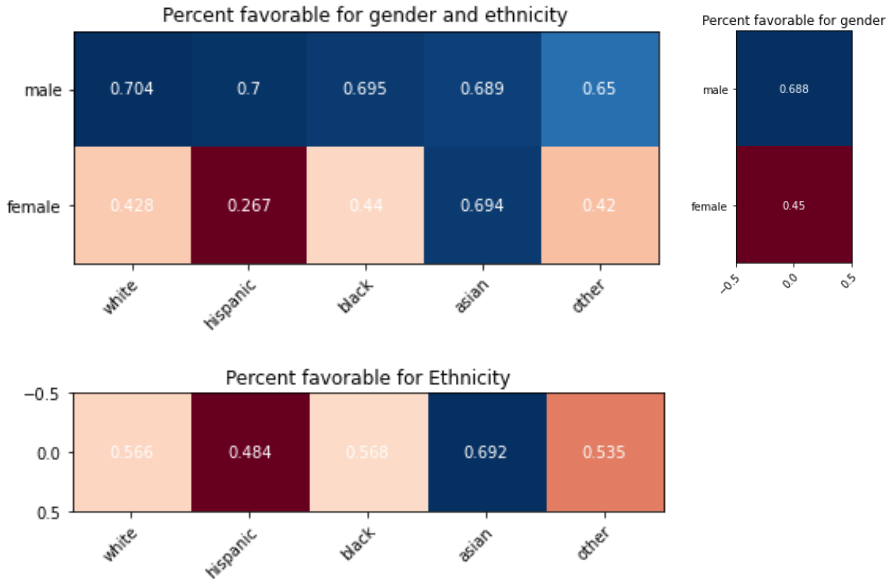
Team members
Ricky Sahu (rickysahu@gmail.com)
Ethan Siegel (09esiegel@gmail.com)
Eric Marriot (marriottew@gmail.com)

Contact
rickysahu@gmail.com
865-414-5994 (please text if no response)

Abstract
(200 words, does not count toward page limit)

GitHub code (less than 1 page)

https://github.com/genhealth/bias_challenge



Methodology Overview (less than 1 page)

Our team's approach to measuring and mitigating bias was informed by commonly used mathematical and statistical approaches, our experience developing machine-learning algorithms, and our experience managing, measuring and transforming data to the FHIR standard for 40M patients at 1upHealth.

To measure bias, our team approached the problem by considering the principles of **equal opportunity**, **equal odds and differential validity**. This approach allowed us to most effectively measure **retrospective and latent bias** present in the output of a given model. If our tool is used to measure a model's output over time as it is updated with newer data, we would also be able to measure the change in bias over time.

We utilized demographic data in conjunction with the provided binary outcome data to measure bias across the intersection of two group categories: sex/gender and race. While we used binary outcome data to produce these plots, continuous outcome data could also be used. We calculated the following metrics:

- Equal odds difference
- Equal odds ratio
- Demographic parity difference
- Demographic_parity ratio

Our solution supports output from all classification models including classical linear models, svm tree and XGboost type models, neural networks, and sequence to sequence generative AI models. We tested our bias measurement tool with multiple models which included biased, poor performance, excellent performance, and random guess output. Through these tests and by using our own generative AI model we were able to test and develop the above bias measurement criteria.

To mitigate bias, we created a classifier using XGBoost in combination with a threshold optimizer. The classifier is trained to predict a binary outcome based on a column defined in the input dataset. The model requires a number of arguments to be supplied in order to define the protected classes and reference classes.

The model also uses a threshold optimizer from Fairlearn in order to adjust the sample weights to minimize the equalized odds between the protected and reference classes. In other words, it is trying to match true positive and false positive prediction rates across all classes as defined in the input.

Value Proposition (less than 1 page)

Our model most effectively identifies retrospective and latent bias, meaning the likelihood for a predicted favorable outcome or classification to differ between population groups. It utilizes simple and easily understandable statistical methods to derive relative outcome percentages across population groups. It also outputs plot visualizations and a report displaying the differences. We believe that using simple methodologies can foster a greater trust in the outcome of any model as it ensures that results can be easily explained and therefore trusted.

The metrics used to identify bias could be binary outcome data (e.g. favorable vs unfavorable outcomes), or continuous data (e.g. percent likelihood of hospital readmission) in combination with demographic data. In our code, we utilize sex/gender and ethnicity, but other protected class groupings such as age, disability status or veteran status could easily be added.

The bias identified is both social and predictive. Based on the inputs, the tool will be able to identify the equalized opportunity and differential validity deltas between different population groups.

Finally, we've tested this model on the new technology of generative healthcare AI that our team has built. We believe generative AI will be a large part of the future of healthcare AI as will FHIR sourced data. Given that those models are trained on claims and clinical FHIR data we have demonstrated that our disparity measurement tool can be used to calculate the bias on models that will compose a large part of the future market. To implement this test we "prompted" our generative API with patient level event histories to predict certain outcomes. The favorable and unfavorable outcomes were encoded as correct or incorrect predictions for various conditions, procedures, medications, etc. We then populated the same input data structure and ran it through our bias detection tool proving that we can use it on the cutting edge of machine intelligence.

Healthcare Scenario (1 page)

The tool would be able to identify large discrepancies in outcomes/recommendations made across various population groups. Because the tool is highly flexible, it can be used for a variety of use cases including:

- Population cohort outcome analysis
- Clinical trial recruitment
- Drug response analysis
- Predictive imaging or diagnostic algorithms
- Hiring and recruitment
- Diagnostic bias in health records

While the tool can detect these large differences, the tool alone would not be sufficient for determining root cause or remediation methods. In healthcare it is especially difficult to understand whether an output from a machine-learning model is producing an accurate prediction based on biology, or is perpetuating a societal inequality reflected in the training data, or some uncomfortable combination of both factors.

While these philosophical issues are difficult to answer, practically, the use of this or other similar bias identification tools could be used to provide fast feedback on algorithms used in a clinical, laboratory or administrative context.

For example, if an algorithm was developed which identified retinopathy using a novel machine-learning image classification algorithm, the outcomes from the algorithm could be fed into our bias identification tool to determine whether there was a large discrepancy in recommendations made across protected groups. While a large discrepancy would not be sufficient on its own to prove bias, it would be a “yellow-flag” to prompt further investigation into the accuracy of the algorithm for the identified protected group. This responsibility to investigate and/or remediate would fall on a cross-functional team consisting of both the technological team responsible for developing the algorithm, as well as the clinical team utilizing the algorithm in daily practice.

In practice, this could be done very easily by creating weekly data exports of recommendations made to the technology team which would include basic demographic data such as age, ethnicity and sex/gender.

Creating continuous workflows like the one described above would be able to ensure that different protected groups receive equitable care, especially when the care is being recommended using a machine-learning based approach.

Operational Requirements (less than 1 page)

The tool we've developed can be run locally on a personal laptop, or in any cloud services vendor (AWS, Google Cloud, or Azure etc) in any operating system which runs Python. The amount of processing power required depends on the size of the input data, but generally speaking, the requirements are minimal. 2-4GB of RAM and 1-2 vCPU's on commodity hardware would be sufficient to run the tool on modest to large data. GPU's are not required, although they would accelerate the performance of the mitigation model. The tool requires the user to have a Python3 installation with a number of other library requirements listed in the requirements.txt file in the Github repo.

There are no proprietary dependencies or third-party vendors required to run the tool. The only requirement is that the input data must conform to the expected CSV formats described in the README.

Per the instructions, the underlying architecture is extremely simple, consisting of two standalone Python files.

This tool can be widely and easily deployed by following the instructions in the Github README. We acknowledge that the tool will be distributed under the BSD 3 license. Finally we plan on hosting a managed API endpoint of this bias detection tool where a CSV can be uploaded for a model's output and our tool can respond with the various bias measurement metrics without any requirement for the executing party to run this system themselves.

Sustainability Plan (1 page)

While from a technical perspective it would be very simple to deploy the bias detection tool in a clinical or administrative setting, in practice there would be many stakeholder groups involved in the ongoing maintenance and administration of the tool in order to promote successful adoption. The responsibilities of each group are described below:

Accessibility
Ultimately we plan on building a company around our Generative Healthcare AI tools at GenHealth where it will be imperative to employ tools for bias detection and mitigation. As a commercial entity we expect we can not only fuel the further development of AI in a responsible manner, but also expand the use of responsible and less biased AI more broadly. These AI tools will be made available both through APIs and as fully fledged applications. Therefore individuals and organizations across the industry would be able to integrate the service into their solutions and have access to AI which has been thoroughly tested for bias and corrected. Finally we will enable users to train and fine tune the model using their data set. Our team plans on iterating on models quite frequently where we will both publish performance as well as bias validation data sets, allowing customers to test bias on the model's output for their own data using the open source version of the bias measurement script published as part of this challenge. This open source version will both be available in github for individuals or organizations to run on their own and as a managed API that can be invoked without the requestor running the bias measurement service themselves. We expect the managed API to greatly increase accessibility of bias measurement in the industry.

Technologists: This group refers to the team responsible for developing and maintaining the technical components of the bias detection tool. We expect part of this team to be the GenHealth team which is building and implementing this model and bias detection services. However we plan to extend our model to a team of technologists that are outside of our organization who can bring their own data and bias measurement algorithms to the tool. This team would likely consist of a machine-learning engineer or informaticist, a clinician and a product manager. They would be responsible for guiding the development of the algorithm, soliciting feedback from the other stakeholders, and implementing any feedback for improvement.

Clinicians: This group refers to the doctors, nurses, physician's assistants and other clinical personnel who would be the "users" of the model that the bias tool is measuring. They would be responsible for viewing the output of the tool, and providing regular feedback to both the scientists/developers of the original model being measured, as well as the bias detection tool technology team.

Administrators: This group refers to the administrative personnel responsible for the successful implementation and use of the bias detection tool.

While the bias detection algorithm itself is fairly simple to run and utilize, the future upkeep costs of the tool are likely to be primarily related to the team responsible for running, deploying and continuously monitoring the tool with data provided by the underlying model being measured.

Generalizability Plan (1 page)

The tool we've developed uses Python and commonly used libraries and statistical approaches familiar to most data scientists and statisticians. Besides the technical dependencies, we only require input files formatted according to the schemas described in the prompt in order to function. These limited requirements maximize the potential for the tool to be used in a variety of environments for a wide variety of use cases.

Our tool can be used by data scientists, informaticists, software engineers, data engineers and clinical staff who have taken an introductory class on Python and/or Data Science. It can easily be run on an ongoing basis to continuously monitor bias and discrepancies present in an underlying classification algorithm at a point in time or over time. It can be applied to broader types of clinical decisions such as predictions, diagnosis and treatment recommendations as these types of algorithms are usually classification based algorithms. A SOP is provided in the Github README to ensure it can be easily implemented at any location which has access to a minimal server with an internet connection. The tool can be implemented in as-is without additional support

We have demonstrated that the measure disparity module can improve the fairness and trust of ML based algorithms in healthcare settings as it is able to clearly display the differences in recommendations and outcomes across various protected groups. The tool could easily be adopted for use in any clinical setting including cardiology, oncology, obstetrics or any other discipline which utilizes a machine-learning classifier in a clinical context.

Given that the code for calculations are published, other users of the bias measurement tool can fork and modify the logic to apply to their variants of data input and additional bias measurement scores they may want to consider.

In our implementation requirements below, we suggest an API as an additional mechanism to the open source solution to further generalize access to our bias measurement tools.

Implementation Requirements (1-2 pages)

We propose two methods for implementing our bias measurement tool.

1) The open source solution which can be cloned, modified, and executed in any environment including on prem systems
2) A public API which can accept files and execute the bias measurement as a managed service in the cloud without any expertise or effort required to manage or deploy the solution.

1) The open source solution which can be cloned, modified, and executed in any environment including on prem systems

Implementation Team

In order to implement a bias detection tool into a clinical or administrative environment successfully, a multi-disciplinary team consisting of a machine-learning engineer/informaticist, a clinician and a product manager would be required. They would need to be able to deploy the tool in a healthcare environment, and integrate the tool with a regular data feed provided by the underlying model being measured. The input data to the bias detection tool would need to contain the recommendations made by the underlying model being measured, as well as basic demographic data about the patients including sex/gender, age and ethnicity.

Measurement

The results of the model would need to be shared with the clinical personnel operating at the facility on a regular basis. A successful implementation would be one where a model's "bias" as measured by the tool would be monitored on a daily or weekly basis, and any large discrepancies in recommendations or outcomes identified by the tool would be flagged as requiring further investigation. These investigations would require manual work on the part of the product manager and/or clinical staff to identify if the discrepancy identified is symptomatic of a bias in the underlying model that needs correction, or not. In this case with self hosted deployments, however, measurement of various models and their performance with respect to bias will be contained in these individually deployed environments.

Deployment

In this option, users of the bias measurement tool are responsible for deploying, maintaining, and using these services on their own accord. Most organizations may want to produce a service that executes the bias measurement tool in response to some model's output. This would require the organization to further develop and maintain the system which is why we propose an additional option of a managed service which does not require any deployment or maintainance.

2) A public API which can accept files and execute the bias measurement as a managed service in the cloud without any expertise or effort required to manage or deploy the solution.

Implementation Team

In this case, the GenHealth team will produce a publicly available endpoint which can accept input from models and respond with the output of the bias measurement tool. Our team of data

scientists, physicians, and engineers will continue to develop and publish improved versions of the bias measurement tool for external teams that must test and validate their models. External teams will only be responsible for producing the output for their models in the format required by the API and POSTing that data as a file to the endpoint. As a response, our API will produce the various scores for the protected classes. Ultimately this is a more optimal implementation from a systems perspective. Rather than multiple individuals at multiple organizations needing to understand and maintain a bias measurement service, only one organization (GenHealth) must do so and can do that with a more specialized and focused team.

Measurement

In the managed API case, we have an opportunity to publish (at least in an anonymized manner) bias measurement scores for various models. We can produce both distributions and percentiles for those models. Because our API service receives those inputs from multiple organizations in a single managed environment we can collate those bias measurement scores and model performance. We plan to publish those measurements as a public graph denoting the number and percent of models that achieve certain measure scores. Users both in the self hosted and managed use cases can publicly view those distributions to better understand where they sit with respect to bias in the industry. Furthermore the public and policy makers may gain additional insight into the performance of these models in the wild. We will have to be careful in how we project these measurements as they could be falsified or inaccurate for non production ready models and could ask organizations to provide some additional context for the models that they run through the bias measurement tool.

Deployment

GenHealth will deploy and maintain this API and service in our own cloud environment. We will monitor uptime and manage the scalability of the service without the need for users to do so. We plan to deploy this as a cloud function which can scale up or down on demand. The bias measurement service will be wrapped in a Flask API layer with an API key that allows developers to invoke the API on demand and connect it with their ML Ops pipeline. Those score will be cataloged and stored so that these organizations can reference their historical model performance on these bias measures and so that they and the public can see how their own models perform against the sea of other AI services with respect to bias mitigation.

Lessons Learned (less than 1 page)

The tool we've built can positively identify large discrepancies in recommendations or outcomes produced by a machine-learning tool in a general healthcare context. It can be used in a clinical, research, government, or laboratory setting. We believe that its utilization of a flexible, simple and powerful approach, our bias measurement service will help build equity and trust in machine intelligence algorithms.

Standardized measurement though a public data set

Although we have made great progress in composing a bias measurement tool, The challenge as posed, is limited in scope and the complexities of real world use cases. A large part of bias in machine learning algorithms stems from the data that it is trained upon. Without models that can be publicly tested on a consistent data set, it is difficult to produce a standardized bias measurement. As we develop GenHealth and our own Generative Healthcare AI models, we expect to produce a sample public data set on anonymized data similar to MNIST or ImageNet so that researchers can train, test, and validate both the performance of their models and the bias of their models on a level playing field. A standardized data set will help level the playing field to train models and can be further used as a consistent data set upon which bias is measured for various protected classes. We plan on creating a de-identified FHIR data set to be used for this purpose.

Alternative approaches

This challenge did not allow for us to measure bias by using protected classes as inputs into the model. Instead it assumed the outputs of the model will be generated without modification of its input features. Increasingly, as we see MI models trained on ethnicity, gender, and other demographic information, it is possible to alter the input gender from male to female, for example, and see the output of the model change. In healthcare, we find that biology and bias are difficult to distinguish, so altering those inputs may drastically change the predicted care paths. In our research we saw alternatives where coupling the variable inputs with cost data can offer more holistic ways to measure equity. Those measures are more economic models that try to identify whether certain protected classes get more or less resources (or dollars spent) to aid their care. With newly available cost public cost data from price transparency regulations, this becomes an easier solution. We plan to explore cost based measures to identify bias across classes.

Conclusion

We find that bias detection and mitigation in machine intelligence throughout healthcare will foundationally impact the acceptance and adoption of these models and ultimately equitable care for individuals. From our experience the proliferation of FHIR data offers unique methods of 1) training and measuring bias for these models and 2) measuring bias in the source data from which they arise. Furthermore generative AI and its introduction into healthcare will become more important. Our team and tools demonstrated that we can measure bias in machine intelligence built on both traditional data sets for statistical machine learning models and proprietary generative AI trained on FHIR data. Building and deploying these AIs responsibility from the onset will be critical to ensure broader adoption while protecting the population from biased treatment.