Plan Overview

A Data Management Plan created using DMPonline

Title: Mitigating bias in healthcare for underrepresented groups with multimorbidity, through AI-based synthetic data generation

Creator: Nazia Nafis

Principal Investigator: Dr Venet Osmani

Project Administrator: Dr M Villa-Uriol, Dr Alvaro Martinez-Perez

Affiliation: The University of Sheffield

Funder: Engineering and Physical Sciences Research Council (EPSRC)

Template: EPSRC Data Management Plan

Project abstract:

The rapid rise of multimorbidity and increasing health inequality are two urgent and interrelated public health challenges facing the world today. Multimorbidity refers to the coexistence of two or more health conditions in an individual. It can include not only physical and mental health conditions but also issues such as learning disabilities, sensory impairments, and substance abuse. Multiple studies have shown that patients from certain socioeconomic backgrounds and ethnicities are at a greater risk of multimorbidity. In this context, healthcare systems that have traditionally only been designed to address individual diseases (as opposed to their combinations) have lately begun relying on artificial intelligence (AI) and big data analytics to manage complex situations. However, these algorithms and the datasets they are trained on, often themselves contain hidden biases against underrepresented sections of society, thereby exacerbating existing inequalities in healthcare and doing more harm than good.

This research aims at mitigating bias in healthcare for underrepresented groups with multimorbidity. Specifically, we target bias in popular health datasets and propose using synthetic data generation to de-bias them, so that any algorithms and machine learning models built on top of them are representative of the real-world population. Synthetic data generation is the use of generative AI to create artificial data that mimics real data. We plan on using the MIMIC III dataset and tailored generative adversarial networks (GANs) and Diffusion-based models for this purpose. We also aim towards analysing existing evaluation methods and develop guidelines for a new and comprehensive evaluation method to assess the quality of the synthetically generated data that fully captures the bias-mitigation aspect of the synthetically generated data.

ID: 145972

Start date: 01-10-2023

End date: 30-04-2027

Last modified: 05-06-2024

Grant number / URL:
https://drive.google.com/file/d/1xjSg7jEz62i5YdPvTF0P5H9zYhbJbQuV/view?usp=sharing

Copyright information:

Mitigating bias in healthcare for underrepresented groups with multimorbidity, through AI-based synthetic data generation

Data Collection

What data will you collect or create?

This work involves working with **secondary data** as input, therefore I will not be collecting any new data.
We are working specifically with the very popular **MIMIC III** (**M**edical **I**nformation **M**art for **I**ntensive **C**are III) dataset. It is a large, **anonymised**, **de-identified** health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012.
The official citation of data is as follows:
Johnson, A., Pollard, T., and Mark, R. (2016) 'MIMIC-III Clinical Database' (version 1.4),*PhysioNet*. Available at: https://doi.org/10.13026/C2XW26.
The data produced as the output of this work will resemble the MIMIC dataset in terms of its overall structure, but it will be smaller in dimensions. It will be **completely synthetically generated** (and therefore, not referring to any person). It will be stored as comma-separated value (CSV) files.

How will the data be collected or created?

MIMIC III is **publicly available** but access requires researchers to put in a formal request via the following process:

- The researcher needs to complete a recognized course in protecting human research participants, which includes Health Insurance Portability and Accountability Act (HIPAA) requirements.
- The researcher needs to become a credentialed user on the PhysioNet website.
- The researcher needs to sign a data use agreement, which outlines appropriate data usage and security standards and forbids efforts to identify individual patients.

Once access is obtained, the data is available to the researcher to download, in the form of comma-separated value (CSV) files.

I have completed the aforementioned steps and obtained access to the data

Documentation and Metadata

What documentation and metadata will accompany the data?

The research paper introducing the MIMIC III dataset, titled "MIMIC-III, a freely accessible critical care database" was published in **Nature** Scientific Data open access journal in May 2016. The dataset is provided to researchers (upon confirmed access) as a collection of comma-separated value (CSV) files. Scripts to help with importing the data into database systems including PostgreSQL, MySQL, and MonetDB are also provided to those who require it. A public code repository also exists regarding the same, at https://github.com/MIT-LCP/mimic-code.
The data produced as the output of this work will be stored as comma-separated value (CSV) files. It will resemble the MIMIC dataset in terms of its overall structure, but it will be smaller in dimensions. It will be named appropriately with **version control** so that data generated as a result of different synthesis algorithms does not get mixed up. I will also maintain an overall **README for project-level documentation**, including details on the Python codes and data.

Ethics and Legal Compliance

How will you manage any ethical issues?

I do not expect ethical issues by virtue of using this secondary data itself. MIMIC-III contains **anonymised** health-related data. The anonymisation process includes **deidentification**, **date shifting**, and **format conversion** of data. It has been made accessible to

researchers internationally under a data use agreement. The **open** nature of the data allows clinical studies to be reproduced and improved in ways that would not otherwise be possible. I do not intend to share this data with anyone.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

I am working with secondary data. MIMIC III is credentialed under the PhysioNet Credentialed Health Data License, version 1.5.0, and is the **copyright (c) of MIT Laboratory for Computational Physiology** 2024. The license can be viewed at this link. Researchers who obtain access to it are restricted from sharing the data with anyone else.

Storage and Backup

How will the data be stored and backed up during the research?

MIMIC data will be stored on the **research storage (X: drive)** of the University of Sheffield. It provides up to 10TB of shared storage per research group for researchers to store the project data during the life of a project. Research storage on the University networked filestore (X: drive) is **backed up automatically** between two geographically separated data centres

How will you manage access and security?

The research storage (X: drive) of the University of Sheffield is **secure** storage space and data will be **accessible both on and off campus** through the university's Virtual Private Network (VPN) to get the files via a web browser.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

I would prefer to **delete MIMIC III data** from the X: drive at the end of my project. The data produced as the output of this work, along with the algorithms used to produce it, **will be retained**. To the best of our knowledge, the data provider of MIMIC does not limit the storage of any data that has been synthetically generated using MIMIC, as this data is completely AI-generated and does not resemble the records of any real person.

What is the long-term preservation plan for the dataset?

The **long-term preservation plan** for the data produced as the output of this work is to place it in **ORDA**, the University of Sheffield's research data repository. ORDA is ideal for non-sensitive data that can be made publicly available online, and we believe our output data falls in this category. Our code and data could be of potential long-term interest to other researchers as well, so ORDA would be the storage location of choice.

Data Sharing

How will you share the data?

I will not be sharing MIMIC data with anyone However, it should be noted that anyone who requires access to MIMIC data may obtain it since it is **publicly available** and accessible to researchers after putting in a formal request.
The data produced as the output of this work would be placed in ORDA, the University of Sheffield's research data repository. Our code and data could be of potential long-term interest to other researchers as well, so ORDA would be the storage location of choice. A **data availability statement** will be included in my thesis and other publications arising out of this work about how the data can be accessed. This will include the **DOI** of the data that we will put in ORDA.

Are any restrictions on data sharing required?

MIMIC data is copyright (c) of MIT Laboratory for Computational Physiology 2024. Researchers who obtain access to it are restricted from sharing the data with anyone else. **My project plan does not include sharing MIMIC data with anyone else**
To the best of our knowledge, the data provider of MIMIC does not limit the storage of any data that has been synthetically generated using MIMIC, as this data is completely AI-generated and does not resemble the records of any real person. Our code and data could be of potential long-term interest to other researchers as well. Therefore, the data produced as the output of this work would be placed in **ORDA**, the University of Sheffield's research data repository.

Responsibilities and Resources

Who will be responsible for data management?

I will be responsible for the management of data, with the support of my supervisor. We have detailed our plans to place all relevant data and code in **ORDA**, the University of Sheffield's research data repository.

What resources will you require to deliver your plan?

No further resources should be necessary.