Title
Enabling Synthetic Data Usage for Medical Research

Author
Fornaca, Charlie Ann

Publication Date
2022

Peer reviewed|Thesis/dissertation

Enabling Synthetic Data Usage for Medical Research

by

CHARLIE ANN FORNACA

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Vladimir Filkov, Chair

Nicholas Anderson

Xin Liu

Commitee in Charge
Spring 2022

Enabling Synthetic Data Usage for Medical Research

Charlie Ann Fornaca

Acknowledgments

Special thanks to the DataLab at the University of California, Davis in the Peter J. Shields Library. Thank you to the CITRIS, CTSC, and Department of Public Health Sciences in collaboration with DataLab for funding my research with the awarded Translational Health Data Science Fellowship. Thank you to Dr. Vladimir Filkov and Dr. Nicholas Anderson for you guidance through my research. Thank you to all of my human research participants and my use case teams.

And special thanks to Gavin Plesko. Without your love and support at home, I wouldn't have been able to achieve my academic goals.

To Roswell

I dedicate this culmination of my efforts you. I'll miss you always and will see you again at the Rainbow Bridge.

Abstract

Enabling Synthetic Data Usage for Medical Research

by

Charlie Ann Fornaca

Master of Science in Computer Science

University of California, Davis

Professor Vladimir Filkov, Chair

Acquiring data can be a major hurdle to any data science problem. Sometimes there isn't enough data or, as is particularly the case for healthcare data, it may contain sensitive information such as personal identifiers that should not be shared. By generating synthetic health data, researchers aim to overcome obstacles of data access and privacy concerns and thereby allow for quicker and broader use of data by the research community. Through this thesis I have surveyed the current state of synthetic data usage in medical research, recorded the thoughts, experiences, and opinions of synthetic data use in medical research from interviewing medical researchers, selected synthetic data generation tools, assessed the accessibility, usability, and efficacy of the selected data generation tool with the help of two different use case groups, experimented with creative ways to use the chosen synthetic data tool, and used my experiences to write resources for current and future researchers who need assistance getting started with synthetic data generation through the UC Davis DataLab.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Synthetic Data

Artificial intelligence, machine learning, and data science has advanced to a point in which we can give an algorithm some data and have it return a new dataset that is both similar and different from the original data. These datasets maintain the statistical properties and distributions of the original data, but can also hide what exactly the original data looked like. In medicine this means that the rows in a synthetic dataset do not correspond to identifiable patients or individuals in the original dataset which is paramount in protecting patient privacy.

Both statistical simulation and computational derivation can be used to generate synthetic data. In just a few moments, entire populations of patients can be created that statistically simulate human physiology and disease states in a multi-dimensional many-featured space. Methods for generating synthetic patient data all follow a similar scheme: provide

private electronic medical record (EMR) data samples, choose and fit a model, and generate new synthetic EMR samples from the chosen model.

1.2 Contributions

This thesis is the culmination of over twelve months of research efforts. I began my research with completing a comprehensive survey on the current state of synthetic data through the literature to form the background and theory behind my other contributions. To date, there hasn't been a survey of this scale for synthetic data generation in the clinical realm from an academic standpoint and especially not that that hasn't had an underlying purpose to promote a particular paper, program, or algorithm. This survey is contained in chapter 2 of this thesis and is included as the theoretical foundation and background for my research and experiments.

Because of this background, I was able to round up a sum of tools for synthetic data generation and freely available medical datasets to examine and report on in detail. Once again, this is likely the largest comparison. Most papers have been able to compare only a few synthetic data tools and methods – particularly ones including a method or algorithm also created by the writing team.

I also uncovered previously unknown thoughts and experiences of medical researchers on the topic of synthetic data use. Though there has been plenty of publications advocating for why medical researchers should use synthetic data, I could not find any papers that delved into *how* medical researchers use or understand synthetic data generation tools and methods.

In addition to this novel information, I conducted new ways to use an open source syn-

thetic data generation tool to satisfy a use case. I also generated synthetic data for an additional use case and reported on the efficacy.

In summary, my research contributions are:

- Surveyed the current state of synthetic data usage in medical research in order to contextualize my research.
- Used the aforementioned literature survey to develop interview questions in order to record the thoughts, experiences, and opinions of synthetic data use in medical research from actual medical researchers.
- Used the findings from the interviews to guide the selection of a couple of synthetic data generation tools.
- Assessed the accessibility, usability, and efficacy of the selected data generation tool with the help of two different use case groups overseen by some of the medical researchers I connected with.
- Experimented with creative ways to use the chosen synthetic data tool and produced promising use case data.
- Used my experiences to write resources for current and future researchers who need assistance getting started with synthetic data generation through the UC Davis DataLab.

1.3 Overview of Findings

The following is a brief overview of the various findings I have uncovered during my research for this thesis.

The literature that currently exists has high hopes for synthetic data use in medical research and yet very few of the medical researchers that I interviewed were familiar with or had used synthetic data in their work. I also found that most of the researchers were willing to use and willing to trust synthetic data where it is properly cited in publications. The researchers that were interested in using synthetic data were at a loss for how to get started with generating their own data. There are hundreds of proprietary synthetic data tools and much fewer open source tools.

Using an open source synthetic data tool, such as SDV, is an accessible experience. SDV in particular can be manipulated in ways to allow for creative use with promising results for balancing unbalanced real datasets. The efficacy of the synthetic data generated was "good enough" for use in research by the use case teams I worked with.

1.4 Purpose of Research

My research aims to answer the question of how we can serve clinical researchers at UC Davis by enabling them to use synthetic data generation tools. To do this, I sought to understand why researchers use or elect not to use synthetic data, how well researchers might know about synthetic data availability and use cases in medical research. Additionally, I sought to determine attitudes, stigmas, and experiences with using synthetic data for research and published studies and gauge researchers' familiarity with synthetic data tools,

concepts, and practices. Once I understood this, I studied insights into how these tools could help and then be improved to meet the needs of researchers as well as how these tools and training on how to use them can be made more accessible to the people who conduct research in clinical areas of interest at UC Davis.

The purpose of my research can be summarized as:

1. Uncover novel information about the relationship between clinical researchers and synthetic data use.

2. Find out to what extent, if any, are researchers familiar with and trust synthetic data.

3. Use the findings of the previous items to pinpoint a synthetic data generation tool or library that meets the needs of researchers at UC Davis.

4. Understand how the proposed solution to share works, how accessible it is, and how reliable the results are.

5. Deliver synthetic data resources to the UC Davis DataLab for current and future researchers to use.

DataLab at the University of California, Davis

The DataLab at the University of California, Davis, strives to increase UC Davis's research impact with expertise based on data-driven projects and collaborations, support the next generation of data-capable researchers and students by hosting workshops, and foster and coordinate data-enabled researchers and university efforts. DataLab facilitates research in data science and applied data science in a myriad of domains. Through DataLab, UC

Davis researchers at any stage in their career can receive training, advice, and collaboration services.

A Toolkit for DataLab

By conducting this research, I will be able to develop a research "toolkit" on generating synthetic data for DataLab. DataLab toolkits are self-service reference documents that help researchers learn more about a tool or method regardless of their skill level [6].

These toolkits are also open source projects that depend on the community to contribute to. The research in my thesis won't just serve current clinical researchers at UC Davis, but future researchers from any domain who are interested in using synthetic data to supplement their projects will also be able to use the toolkit I develop.

On October 22nd, 2021, I gave a demonstration through DataLab to the greater UC Davis research community about how to use two of the synthetic data generation tools that I had been comparing. From this presentation and demonstration, I have produced an outline of the toolkit for DataLab:

1. Background information
    a) What is synthetic data?
    b) How is synthetic data is created?
    c) Why use synthetic data in health research?
2. Tools for synthetic data
    a) Synthea

b) Synthetic Data Vault (SDV)

c) A brief listing of other tools

3. Synthea Tutorial

a) Set-up and getting started

b) Running Synthea

c) Changing parameters

4. SDV Tutorial

a) Set-up and getting started

b) Running SDV

c) Changing parameters

The development of the toolkit is already underway and I plan to have it finished by June of 2022.

1.5 Context: The Clinical Data Access Process

Obtaining data at UC Davis can be a lengthy process for physicians hoping to conduct research. The following section illustrates the process of gaining access to clinical data for research.

Clinicians can only see the data of patients with an alignment to their specialty or patients who have been referred to them. UC Davis clinicians who are doing research have different data access rights than UC Davis non-clinician researchers [2].

All physicians have access to the electronic medical record (EMR) system and many also have access to data storage environments such as Epic Clarity and Epic Caboodle as well as data exploration tools like Epic SlicerDicer [12] [21]. Even with a tool like Epic SlicerDicer which is self-service [2] and the allocated access, data is still limited as physicians can view the data in a simple form but cannot run extractions. Ultimately, data access will depend on both the privileges of the researcher requesting the data, what tools support the data access, and whether that researcher even has the expertise to make use of those tools [2].

Simple form A physician can query the Epic SlicerDicer tool to see whether a patient population exists. The tool can only provide a de-identified aggregate data dashboard [2] about the population, but cannot provide data about an individual patient contained in the population [21].

Extraction Taking data from the EMR, Epic Clarity, or Epic Caboodle. Data cannot be extracted without going through a specific data request process through UC Davis's Clinical and Translational Science Center (CTSC) [2] [21].

After defining a query, the physician can make a request to the UC Davis Clinical and Translational Science Center (CTSC) to acquire either identifiable or de-identified data. This process requires IRB approval due to its nature of being human subject research [2] even if only de-identified data is requested and if the research may be considered exploratory and no Protected Health Information (PHI) is provided. With this data, the physician can explore what exactly was extracted from EMR to fulfill their query. What is extracted may not always reflect what is the true state in the EMR [2]. The CTSC will work with the physician to understand what questions the physician [21].

This data that is extracted is almost never [2] organized in tidy columns. Additionally, there may be a lack of transparency of when, if any, cleaning and evaluation stages occurred. The data cannot be assumed to be anything without prior insight into what data transformations have been previously applied [2]. Physicians and researchers might have to extract data from verbose doctor's notes stored in the EMR. By the time the researcher can use the data, it will have had to be cleaned up a bit. At this point, the realization that there isn't enough data may occur [21].

Instead of going through the CTSC to acquire data, a physician could contact the clinical research coordinators and research staff. This path will require the physician or principal investigator to acquire IRB approval to collect patient information directly from the EMR. This process, called a "direct patient extraction" [21], which is a research data request that involves building a query in SQL or similar query language [2], is performed through querying for status reports over a specified time period (daily, weekly, monthly) [21] and is merged into one of three routes depending on the research intended. These three routes are Clinical, Clinical Research, or Research. The Clinical route is for quality assurance and patient care studies if there is no IRB. With an IRB, quality assurance and patient care studies merge into Clinical Research. If there is no ability granted to look at the EHR data, the extraction is merged into Research [2].

After the data is extracted and merged, it needs to be delivered to a specified secure location. This is especially true if the data contains any PHI. From there, the data continues to be monitored and managed under the requirements set by the IRB [2].

At this point, research associates will look at patient records as they pass through the specific clinics and identify candidates that match the inclusion criteria. This will only

happen under the case of "prep to research". "Prep to research" is a specific term under The Health Insurance Portability and Accountability Act of 1996 (HIPAA) that disallows the researcher to view the data and instead a separate "honest" entity who is reviewing the charts of the patients extracted from the EMR to confirm the correctness of extraction [2].

The research staff can then approach the patients and enroll them into the study. The patients' information is collected (manually, most of the time) from the patients through clinical systems, workflows, and labs depending on the needs of the research as defined in the IRB protocol [2]. The patient is tracked over the previously established time period mandated by the studying. Changes are monitored and researchers identify exclusion criteria as they occur [21]. This might mean monitoring patients for adverse events within a prospective study. In a retroactive study, this would be developing a phenotype iterative based on repeated interactions with the EHR and data sources [2].

Depending on the skill of the principal investigator and the clinical research coordinator or the clinical research office overseeing the associated department, the data collection could potentially be manual, manual and automatic, or rarely completely automatic. Manual data collection is when data is collected on paper, stored in physical binders, or digitally in local Excel files and free text notes. Manual and automatic data collection is often collected in certain parts by hand, usually as consent documents with physical patient signatures. Other information may be collected automatically from the EMR. Completely automatic data collection does not happen often and for various reasons [21].

Upon reflection of this whole process, it becomes clear to see how accessible synthetic data generation tools would benefit researchers. Synthetic patient data can be generated at a work or research personal computer with just a few inputs and a couple clicks of the mouse.

It makes sense to begin the process of acquiring the actual data and in the interim using synthetic data to begin pre-processing data and building models. Synthetic data, including data that mimics human subjects or patient health records, requires no IRB [26].

Chapter 2

Background & Theory

2.1 Overview

In this chapter, I explore the research that has already been completed regarding synthetic data generation methods, tools, and evaluation in the domain of healthcare. I will also expand on the reasons why medical researchers may be encouraged to use synthetic data. .

I determined a number of common themes surrounding the use of synthetic data in healthcare research found in the current literature. These themes are by no means mutually exclusive of each other. For example, privacy can be an issue in and of itself but it can also be a concern in how we evaluate a synthetic data generating method.

The following themes derived from the literature are as follows:

- Why researchers use synthetic data
- Methods for generating synthetic data
- Validating synthetic data results

- Privacy concerns
- Access and ownership of real and synthetic data
- Challenges and limitations of synthetic data

2.2 Why Researchers Use Synthetic Data

There is no shortage of reasons as to why the greater research community and specifically medical research could be using synthetic data. In fact, synthetic data has been studied for over thirty years now [34] and the arguments for the usage continue to strengthen as more advances in the efficacy and reliability of synthetic data are made. Synthetic data has been a key part in research, development, and education across a number of domains.

In medicine, synthetic data usage already has a robust resume of applications. Synthetic data helps train assistant medical AI's such as ones that help pathologists make decisions when diagnosing tumors. It has also been used to tune and train clinical decision support tools to help eliminate bias in diagnosing skin cancer [3].Secondary uses for synthetic data can include education, training, software testing, and machine learning and statistical model development. Transfer learning applied from synthetic data to real data also improves machine learning algorithms [10].

Cost, patient privacy, confidentiality limits conducting trials and studies using real patients and their data [24]. Far from retrospective studies, synthetic data finds use in clinical trials. Synthetic data can be used to mimic a control group of patients receiving active therapy in early phase clinical trials [8]. Synthetic data can also be used for large trials

that consider novel targeted therapies involving genomic gadgets. With these kinds of trials having a large enough randomized clinical trial is impractical [8]. A concrete example of how synthetic data can benefit medical research can be illustrated through drug responsiveness trials. Though randomized clinical trials are regarded as the gold standard to evaluate drug effectiveness, the phase 3 clinical trials are notably expensive. Decision-makers can be provided evidence for validly conducted studies at an accelerated rate by using synthetic data derivatives. Thus the costs of these trials can be lowered. Trials are not often designed or have enough power to evaluate how effective a treatment is comparatively. Using synthetic data, studies about assessing real-world treatment effectiveness and patient outcomes can be implemented [8].

Synthetic data can pave the way for reducing bias in medical research. Generating synthetic data facilitates finding live-saving insights that we aren't currently able to see for entire populations and select demographics [27]. Bias in data collection can occur in unexpected places when it comes to medical data. For example, second-hand data is more readily available than the data that needs to be collected for a specific analysis or question in a well-designed clinical trial. This ratio of readily available general data to specific case data is unbalanced. Data that is collected in a hospital contains the data of more severe patients that have already been diagnosed with the specific disease. Ultimately, in available data, there are far fewer patients who do not have the diagnosis or disease of the target study than the number of patients specifically targeted with the diagnosis [34].

A consumer revolution in healthcare is dependent on analysis of cost data. Synthetic data can solve this issue because of how financial outcomes can be incorporated into synthetic data generation [29]. Financial data in healthcare such as total claims, claims amounts,

negotiated rates, and billing codes are often proprietary and difficult to obtain for economic and social improvement studies [29]. Additionally, financial healthcare data extracted from a hospital system's very own electronic medical health record database can often lack associated clinical data [29]. Generating synthetic financial healthcare data with complete synthetic patient records hardly exists in the real world. Having this kind of data with real statistical distributions can dissolve the isolation between different provider groups and lead to better health and financial outcomes for patients [29]. Having this particular kind of synthetic data at the ready allows for open source community members with different skills and backgrounds to develop solutions to enhance the value of care [29].

There are many systems of interest that fall short of good quality data. Data is scarce due to lack of collection tools or there is a limited rate of data acquisition [13]. Additionally, depending on the data capture process, the recordings of clinical observations can vary greatly and lack consistency. Determining if real clinical effects were truly represented and observed becomes a challenge [24]. Large and representative datasets are required for researchers to develop, refine, and improve treatment guidelines [8]. Synthetic data can be used in rare disease studies to augment existing data [8] where there is simply just not enough data to conduct a proper study.

Data can be increasingly difficult to share across organizations especially when it contains sensitive patient information. For this reason, medical data is not broadly available to the larger research community due to privacy concerns [10]. Even large research and medical institutions might lack the infrastructure and support to increase a study size or share data at scale internally between hospital systems [8]. One might think that "simply" de-identifying clinical data should be enough to prepare the data for sharing, however determining the

efficacy of de-identification methods on real data have been largely inconclusive especially for large datasets [10].

Publicized medical datasets are difficult for outside researchers to gain access to because of patient confidentiality. Ensuring proper usage when a researcher can access a database is a lengthy process with strict legal requirements. Because of these restrictions and time until research results, translational benefits to patient care are severely delayed [34]. Time is a paramount factor when it comes to performing medical research. This has been especially important during the global COVID-19 pandemic where rapid development, testing, and deployment has been of the essence. The benefit that synthetic data has in healthcare is pivotal to our success in advancing medical research and practice. Some say that it is our ethical responsibility to find ways to use all technology and scientific advancements available to us to improve healthcare [27]. Several factors that often are the culprits of taking up time in research can be eliminated with the availability of good synthetic data or the necessary tools to generate data as needed. In addition to the general limitations (privacy, regulations, laws, security, data ownership) that are associated with accessing electronic medical record (EMR) data, the approval process of the local institutional review board (IRB) can delay research.By using synthetic data that contains no human subjects, research can commence while approvals are underway. IRB approval, while important and valid to structured, ethical research, can also create difficulties in collaboration and sharing data. Additionally, research grants can be earlier applied for if preliminary data can be extracted or generated and analyzed before beginning the IRB application process [26].

Unfortunately, in healthcare research there is no ImageNet or MNIST equivalent of safely collected medical data [13] [27] and the datasets that do exist for use in research are often

limited by the dearth of data submission and collection [28]. Representative samples of human populations are necessary to improve and develop machine learning models [28] and if there isn't enough data, machine learning and predictive models cannot be built with confidence. Being able to summon a synthetic dataset allows healthcare researchers to test the algorithms they develop. These algorithms and models can be anything such as diagnosis and treatment recommendation systems as well as future event prediction systems [28]. Oftentimes, developing and validating a machine learning method for certain tasks doesn't require real data at all and a synthetic dataset would work instead [27].

2.3 Methods for Generating Synthetic Data

There are several models, techniques, and approaches for generating synthetic data. These approaches break down further in generating healthcare data into specific methods. For example, one method has been to produce virtual patients with completely fabricated individual medical histories.

Goncalves *et al.* identify and evaluate three methods of data-driven synthetic data generation approaches including probabilistic models, classification-based imputation models, and generative adversarial neural networks (GANs) [10]. The importance of using data-driven methods such as probabilistic models, classification-based imputation models, and generative adversarial neural networks (GANs) for generating synthetic data is that it doesn't require a subject matter expert to curate the data [10]. These data-driven methods work by using generative models that have been trained on observed data [10].

Independent Marginals

A simple baseline method for generating synthetic data is to sample from independent marginals (IM). This is completed by sampling the empirical marginal distributions of each feature in a dataset. Though this particular approach is efficient and the estimations can be performed in parallel, IM is not able to capture the statistical dependencies found amongst the features of the dataset [10].

Iterative Proportional Fitting & Updating

Swarup and Marathe use two methods to generate a population of agents with realistic demographic attributes: Iterative Proportional Fitting (IPF) algorithms and Iterative Proportional Updating (IPU) algorithms [30].

The IPF method is achieved by feeding the algorithm marginal distributions over demographics from a sample of household census records. IPF maintains the dependence structure of sub-samples within the greater sample by matching these dependencies to the whole population's marginal totals. Columns and rows of data are adjusted incrementally to match the given proportion.Once a satisfactory joint distribution has been established, it is repeatedly sampled from and matched to demographic records to create a synthetic population [30].

Similarly, using IPU can specifically generate a population of individuals and households with realistic demographic attributes whereas IPF was only applicable to the household level. Another folly of the IPF method is that discrepancies in the distributions at the personal level within generated households can occur. To avoid this in IPU the sampling weights of the household records are adjusted in such a way that the distributions over individuals in

the synthetic population more greatly resembles the results of individual-level IPF [30].

To augment these synthetic populations further, logistic regression and direct lookup in probability tables is used to model the evolution of a population with life events such as aging, mortality, birth, marriage and union formation and dissolution, and migration [30].

Imputation

Imputation based approaches for generating synthetic data conduct statistical analysis with a focus reducing the risk of disclosure of sensitive data. Multiple imputation is performed and then sensitive data is actually treated like missing data. Randomly sampled imputed values are released in the place of the sensitive data. To achieve this, both linear and nonlinear models can be used such as generalized linear regression and random forest or neural networks respectively. Even though imputation methods are fully probabilistic, they may not always generate a model that estimates the full joint probability of the population that was sampled. However, any statistical modeling method that learns a joint probability distribution is still able to fully generate synthetic data. Multiple imputation based methods are a popular choice for generating synthetic data from sensitive original datasets. Using multiple imputation is quick and can easily handle continuous and categorical features. However, it isn't always certain that using a multiple imputation method estimates the joint distribution of the data despite being probabilistic [10].

Bayesian Networks

Using a bayesian network creates probabilistic graphical models of nodes representing a random feature. The edges between the feature nodes represent probabilistic dependencies amongst the features. These graph structures and conditional probabilities are inferred from the real data in order to create synthetic data. This is done by first learning a directed acyclic graph from the data.This graph contains all the pairwise independent or dependence conditionals across the features and estimates the maximum likelihood for conditional probability tables (CDP) for each feature. Using a Bayesian network to generate synthetic data scales well with the dimensionality of the dataset that it is generating from. Additionally, it is computationally efficient. However, the full joint distribution involved in the network is too generate and simplifies the assumption on the structure. This will cause the resulting synthetic data to fail at representing higher-order dependencies found in the original data [10].

Gaussian Methods

Gaussian methods for generating synthetic data use a lower dimensional continuous latent space and nonlinear transformations to map points in the latent space to probabilities for generating the categorical values. Latent space is the embedding of a set of items within a set collection of points where items that resemble each other more closely are more closely positioned to each other. These models assume that each patient record has some continuous latent low-dimensional representation. The Gaussian method doesn't use fully conjugate models, but allows for techniques to vary. The Gaussian model doesn't model dependence

across features but it can capture the dependence across patients. In addition, the shared low-dimensional latent space can capture the dependence across variables or features implicitly. Using a Gaussian model has better latent non-linear mappings than using a Bayesian model. These mappings can represent complicated full joint distributions. Additionally, clustering and data visualization is facilitated when using a Gaussian model because of the inferred low-dimensional latent space. Some of the drawbacks to using a Gaussian model include how the non conjugacy of the model complicates inference. Additional Bayesian inference method is required to overcome this drawback. The inference of Gaussian models also doesn't scale well regarding the data size [10].

Even though there are bottlenecks in the computational runtime of Gaussian synthetic data generation methods, I do not think that they should be considered. The amount of time saved by using synthetic data is already potentially eons of time saved versus waiting for human data use IRB approval to come through.

Generative Adversarial Networks

The class of deep neural networks (DNN) known as Generative Adversarial Networks (GANS) are used for completing unsupervised learning tasks by creative two jointly-trained neural networks. One of these networks generates the synthetic data that mimics the real data and the other network tries to discriminate and judge the synthetic data from the real data. GANs tend to be better suited for generating image data and other high-dimensional continuous datasets. GANs are more flexible than Bayesian and Gaussian networks as they do not require strict probabilistic model assumptions. Additionally, GANs work well with mixed categorical and continuous data types.

The tuning of a GAN model, however, is an arduous process that requires great understanding of the hyper-parameters. GANs tend to have stability issues associated with the min-max optimization problem and are also notoriously difficult to train [10]. Though they perform better than other generative methods such as variational autoencoders (VAE), GANs are typically known for not being used well for learning distributions of discrete variables [5]. The efficacy of using a GAN or other autoencoding method to generate non-image synthetic data is a hot debate with mixed results from validation experiments across the board. There is also much concern over whether synthetic data generated by using an autoencoder is considered free of privacy risk [26].

Generative models don't always exist in a vacuum. There have been many research efforts to include GANs and other similar models as a supplementary method. Some other synthetic data generation tools use a generative model to build statistical similarities. By using a generative model, assumptions about the specific distributions of the original data are required. This is often difficult because these shapes can be very complex or nonparametric [9].

Process-driven Methods

Alternatively, process-driven synthetic data generation methods such as numerical simulations, Monte Carlo simulations, agent-based modeling and discrete-event simulations derive the synthetic data from physical underlying processes represented by computational or mathematical models. These methods tend to require subject matter expertise to shepherd the curation of synthetic data [10].

Tools for Generating Synthetic Data

There are several out-of-the-box synthetic data generation tools available as proprietary resources or as open source endeavors. The following is a brief overview of some of the tools specifically mentioned in the literature I reviewed.

There are several open-source solutions for generating synthetic data such as synthpop and SimPop for R and DataSynthesizer for Python [10].

The research conducted by Anat Reiner Benaim *et al.* compares several medical research results based on synthetic data to their real data counterparts. In addition to their own validation efforts, they open their publication by identifying several synthetic data generation tools that were used to provide synthetic data for various medical studies [26].

Synthea is an open-source tool for generating synthetic patients complete with electronic health care records. The goal of the Synthea project is to be able to provide readily available synthetic electronic health records that can be used in industry and innovation, as well as for research and educational purposes. In addition, Synthea's synthetic electronic health records are free of legal privacy, security, and intellectual property restrictions [35]. Having general population data isn't enough for some research. Synthea can additionally generate data based on models of disease progression and the standards that correspond to treatment of those diseases [35]. Using public datasets and health statistics, the Java-based synthetic data generation tool, Synthea, can imitate the outcomes and progressions for many clinical conditions. However, because Synthea creates data based on clinical guidelines and expertise, it may be too "ideal" to be used in place of real data [26]. Additionally, Synthea links synthetic patient records to financial records [29].

The Observational Medical Dataset Simulator (OSIM). OSIM. OSIM uses observations considering features such as time, gender, and age and then generates data based on diseases and drugs using the probability distributions from real data. However, OSIM isn't able to reflect more complex relationships due to how restrictive the format is [26]. The Observational Medical Outcomes Partnership (OMOP) developed one of the first simulated data programs using an empiric approach. Unlike previous models, OMOP's Observational Medical Dataset Simulator (OSIM) modeled the characteristics of the data itself instead of the biological processes captured in the data [24]. Later, OSIM2 was an Oracle SQL stored procedure that was freely available through the OMOP website. Patients are generated using an individual Monte Carlo approach selecting values from a module [24]. OSIM2 feasibility testing was limited to comparing data characteristics and distributions [24].

As previously explained, Generative Adversarial Networks (GANs) have also been used to generate synthetic patient data. With large datasets, an autoencoder can learn a representation and then generate a new representation of the data that mimics the original. The research of Choi has led to the development of medGAN which had generally impressive results for both generated binary variables as well as count variables. medGAN is able to generate high-dimensional, multi-label discrete variables that can represent the events found in electronic medical records. However, it can only generate features of counts and binary variables. It is unable to take into account the longitudinal nature that accompanies observed medical events [26]. The medGAN method uses minibatch averaging to challenge the issue of overfitting for a few samples [5]. To validate the medGAN method, comparing the synthetic data results to real data results was demonstrated by reporting distribution statistics and classification performance. Additionally, a medical expert reviewed the results

[5].

Another GAN-based synthetic medical data generation tool, MDClone, seems to consistently perform well in studies that pit the original data against synthetic data generated with MDClone [9]. MDClone is a tool that directly queries an EMR based on what is of interest to the researcher. It then generates a synthetic dataset based on the freshly-fetched underlying queried data. To achieve this, the algorithm uses a covariance measure to generate all the variables together. It does not assume the underlying distributions and Anat Reiner Benaim *et al.* claims that it allows for the discovery of relationships that are not previously known before loading the data [26].

MDClone works by querying the actual data and then generating an obfuscated synthesis [26]. This is implemented by using an algorithm that is multivariate and generates the variables all together using a measure of covariance [26]. To support patient privacy, values of populations that may be grouped and considered identifiably unique, are censored and the algorithm continues to derive statistical characteristics from the data thus giving the synthetic dataset similar properties [26]. Though I highly value the research done by Anat Reiner Benaim *et al.* and it is important to criticize developing methods of medical data synthesis, I am skeptical about the claims made in their comparative study. In their study, they claim that MDClone, a proprietary software, has been used in their institute's information technology platform since 2017. Though they may be familiar with MDClone, I am inclined to reject the praise they bestow upon the tool in their comparison.

Finally, a handful of synthetic patient models will build groups of patients. These groups are then used to create individual patients. MDClone uses this technique to generate entirely new synthetic patients. Because of the extra "jump" in generating these synthetic patients,

original patient privacy is confidently assured [9].

Virtual Patient Models

Kartoun has proposed a method for creating repositories of virtual patients called electronic medical records bots or EMRBots. These bots are generated from the given configuration of population-level and patient-level characteristics. Kartoun makes it clear that this methodology should be used for training, education, assisting in hackathons, and developing computational methods. Kartoun advises that these EMRBots are not ideal for use in studying or predicting outcomes for real patients [18].

A unique feature of the virtual patient model described by R. Shamsuddin, B. M. Maweu, M. Li, and B. Prabhakaran [28] is that this particular method can include synthetic time-series data. Time-series data has proven to be a challenge to generate for medical research purposes and is often cited as a limitation to what we can emulate with synthetic data. R. Shamsuddin, B. M. Maweu, M. Li, and B. Prabhakaran were able to achieve this by implementing a genetic algorithm.

Other Methods of Data Synthesis

Other approaches to population synthesis reweighting methods like combinatorial optimization and generated regression weighting (GREGWT). For example, combinatorial optimization estimates a micro-population by stochastically reweighting the given micro sample data. It then randomly allocates individual points of data to each feature and iteratively replaces the data based on how it improves the fit [30].

Tucker introduces a method of generating synthetic data based on probabilistic graphical models [34]. One simple approach to generating synthetic data is to add noise to existing real data. Usually this will not be enough to protect patient privacy in the case of medical data. This can be slightly improved by using a distribution such as a Laplace mechanism [34].

Another method is using generative models of data that can capture relationships between data features. Sometimes these relationships must be hardcoded. Other times, these relationships can be inferred using Bayesian networks and neural networks [34]. After a Bayesian or neural network can identify the relationships present in an existing dataset, a GAN could be a possible solution for generating synthetic data. It is said that GANs can be used to create a more robust and less biased dataset than one generated on the real data alone [34].

Researchers commonly use Synthetic Minority Oversampling Technique (SMOTE) to resample data in machine learning when working with unbalanced samples. The synthetic data points generated using SMOTE are then used to supplement the existing data [34].

Synthetic Databases

Synthetic data databases are also available for researchers to utilize. For example, Gretel is an open-source synthetic data library that can generate electronic health records while maintaining the privacy of the patients in which the data was derived from [36].

2.4 Validating Synthetic Data Results

To test the robustness of synthetic data, traditional statistics, machine learning approaches, and spatial representations of the data can be used [9]. Because different methods of data generation have different evaluation metrics, comparing data generation methods is difficult [10]. There is a dearth of discussion and agreed-upon metrics for synthetic data validation [10].

What will be considered "good" synthetic data has the potential to vary greatly. Aspects of data that might be satisfiable when developed for a time series study, for example, might not suffice for a static study [27]. Additionally, requirements for the evaluation of synthetic data will depend on what the intended usage of the dataset was. For each need in healthcare, such as predictions, survival analysis, clinical trials, causal inference, and decision-making, specific types of synthetic data, performance metrics, and evaluation methods will also be required [27].

We need to validate synthetic data in order to ensure that biases, overfitting, and high variance can be discovered and accounted for [34]. To be effective at imitating electronic medical records, the synthetic data generated should reflect both linear and nonlinear relationships between features. In addition, the data should also consider the temporal arrangement of medical events [26]. The fidelity at the individual sample level should make sense for generated data. For example, a cisgender male patient shouldn't have gynecological records generated [10]. If a trained machine learning model performs well on synthetic data, then it is indicative that the synthetic data is similar to the real data [17].

Some may agree that effectiveness of synthetic data can be measured by how closely the

generated dataset resembles and reflects the original data. In M. van der Schaar and N. Maxfield's article, it is stated that synthetic data be compared "in terms of the joint distribution of features," which takes into account the multidimensionality of datasets especially in medical research [27]. Marginal and joint distributions of features should also make sense bringing fidelity at the population level [10].

Though this isn't a priority in my thesis research, privacy preservation should also be a metric of concern [10].

Common Metrics for Validation

The simplest approach for validating synthetic data is to compare the distributions of the columns across the original and the resulting synthetic dataset. The distributions for all datatypes should match [36].

Kullback-Leiber (KL) divergence doesn't measure dependencies among features. It can successfully measure the probability mass functions (PMG) for each given feature. When both synthetic data and real data distributions are identical, the KL divergence is zero. The higher the value, the more divergence is observed between datasets [10]. Because KL divergence or relative entropy measures how one probability distribution is different from a second, features that have a high KL divergence may not be ideal for synthetic data. This could be due to randomness or other limiting factors [36].

Insights and relationships must also be maintained across the features of a generated dataset. By measuring the correlation using a method, such as Pearson's correlation coefficient, between two values can be quantitatively expressed [36]. Pairwise correlation difference (PCD) measures correlation between features. The smaller a PCD value is, the more similar

the synthetic data is to the real data. PCD specifically measures the difference in terms of the Frobenius norm of the Pearson correlation matrices [10].

Measuring log-clustering metrics demonstrates the similarity of the underlying latent structures of the clustering in the synthetic and real datasets. Cluster analysis is performed on a merged synthetic-real dataset. Large values of the log-cluster metrics indicate differences in the distribution of the synthetic and real data [10].

Support coverage metrics measure how coverage of the synthetic data features support the real data. The ratio of the cardinalities of a feature's number of levels is considered. This metric can catch if a synthetic dataset is not representing less frequent categories [10].

Cross-classification metrics capture how a synthetic dataset represents the statistical dependence structures that exist in the original data. Using this technique measures dependence using the predictions generated for one variable based on variables using a classifier. There are several classification methods that can be used for this metric [10].

Comparing the synthetic data and the ground truth data in machine learning classification tasks and sensitivity analysis can give insight into the efficacy of the generated synthetic data [34]. Essentially, a comparison between any two algorithms that were used on synthetic data should reflect the comparisons of the same two algorithms on real data [17]. The metric is improved when more algorithms are included [17].

Alternative Validation Methods

Tucker's study uses chi-squared, KS, and KLD tests between the real data samples and the synthetic data samples [34].

Instead of evaluating the synthetic dataset as a whole, Alaa *et al.* proposes auditing

each individual sample to test their quality. If a single sample doesn't appear authentic according to their evaluation metrics, the sample is discarded from the dataset. Thus the entire remaining dataset is improved [1].

Alaa *et al.* propose a three-part approach to validating the effectiveness of synthetic data: fidelity, diversity, and generalization. Fidelity is the quality of a model's synthetic samples and can be measured with α-Precision which is the fraction of synthetic samples that resemble the "most typical" α real samples. Diversity is defined as the extent to which these samples cover the full variability of the real samples. Diversity is measured using β-Recall which is the fraction of real samples covered by the most typical β synthetic samples. Lastly, generalization is the extent to which a model overfits, thus copying, the original training data. This measure is quantified by an authenticity metric implemented with a hypothesis test for data copying based on the observed proximity of synthetic samples to real ones in the embedded feature space [1]. Alaa *et al.* claims that using α-Precision and β-Recall works better than using typical precision and recall techniques to evaluation synthetic data because the actual probability densities of both distributions are taken into consideration [1].

Autoencoding methods such as GANs cannot be properly evaluated using likelihood metrics. This is because likelihood metrics including log metrics do not scale well in highly dimensional spaces. Additionally, points of model failure are blended into one number that does not give much context as to where the model is failing [1].

In previous research, time-series data has been particularly challenging to generate [28]. However, we see methods now for generating synthetic time-series data [27]. A comparative analysis of time between two points (such as time between similar conditions, time between

sequential conditions, number of days with conditions, time between first and last conditions of the same type, time between similar drug starts, time between consecutive drug starts, and number of days with drug starts) can help evaluate the longitudinal fortitude of synthetic data. Feasibility for lengths of time can be analyzed using correlation coefficients ($R\hat{2}$) [24].

In R. Shamsuddin, B. M. Maweu, M. Li, and B. Prabhakaran's paper, the effectiveness of using synthetic data in their machine learning algorithms is done by using a comparative analysis of predictive outcomes. Models such as support vector machine (SVM), naive Bayes, and "bagging" were used to measure how successful the synthetic data reflected the relationship between the labels and features [28].

The van der Schaar Lab has defined an approach to measuring synthetic data quality called Synthetic Ranking Agreement (SRA) [17]. This method enables researchers to choose the best algorithms to try on real data after comparing the performance of trained and untrained machine learning algorithms on synthetic data. This allows algorithms to be passed to the real data tenant if they were developed separately from the final tenant. The method of SRA entails comparing a smaller set of algorithms over time in order to develop the machine learning algorithm. Curiously, the creators claim that to score highly using SRA, it does not need the synthetic data to be distributed in the same way the real data would be [27] [17]. Additionally, the van der Schaar lab has tested the SRA method with another metric, Synthetic data and Testing on Real data (TSTR), and has concluded that SRA has stronger privacy guarantees than TSTR [17].

2.5 Privacy Concerns

Using synthetic data can mitigate the risk of invading patient privacy when forming research hypotheses and estimating analyses [26].

There are no standards or universally accepted definitions that are quantifiable for data "identifiability" [27]. Officially, there are no tangible requirements set in stone for privacy regulatory efforts. Neither the General Data Protection Regulation (GDPR) of the European Union nor the United States' Health Insurance Portability and Accountability Act of 1996 (HIPAA) is able to provide the proper definitions, safeguards, or reassurances for data privacy [27].

There is risk even in de-identified datasets. The possibility of linking de-identified patient data to other datasets (such as social media data) can open up new risks for patient identification [34]. Removing identifiable features, "perturbing" them by adding noise, or grouping variables into broader categories to ensure that there is at least more than one individual in each category are all current approaches for obscuring or de-identifying patient data [34].

Simply de-identifying data does not completely eliminate the risk of privacy concerns. Residual patterns can still be distinguished using features such as diagnoses, lab tests, visits across healthcare providers, and genomic variants [5]. Though generating data can create less-risky, workable synthetic patient data, data synthesis is not itself a method for anonymization or de-identification [8]. Data anonymization techniques in generating synthetic data often include aggregation, subsampling, and adding noise . Aggregation generalizes certain features of a dataset by associating a higher category to some of them. To achieve the target population size, subsampling is derived from a larger population. These tech-

niques, however, generate some skepticism in the quality of the resulting synthetic dataset [9].

I will not be focusing on evaluating privacy metrics in the scope of my master's thesis, but it makes for an excellent topic for further research. Some researchers use the term "differential privacy" as a metric to measure how well synthetic data obfuscates or departs from the real data in which it was generated from [27]. For quantifying privacy itself, the closeness of individual synthetic data to real patient data can be scored by using outlier statistics and distance metrics [34].

2.6 Access and Ownership

Data guardians, not data users, are the entities that set the terms for providing data for distribution or research [27]. Data shareability essentially solves the issue of being able to reproduce research [27] as private medical datasets cannot be shared with third parties wishing to verify models [26]. Lack of data ownership in the hands of the patient makes receiving care and resolving financial matters for care a lot harder [29].

Sharing data between research and hospital systems is essential for developing cross-institutional and generalizable insights in medical research [8]. Comparing and contrasting new patient data with other hospitals and health organizations on a local and global scale can greatly improve the process and speed of treating patients [36]. Being able to combine different datasets from across institutions also helps to create a more comprehensive view of the proverbial patient [8].

Sharing data and synthetic data derivatives across institutions can shorten the idea-

to-insight time from years down to hours [8]. This increases efficiency and lowers costs for research and development [8]. This can be done much more easily than sharing real data across institutions. Synthetic data isn't subject to the same regulatory and ethical impediments for sharing, securing, storing, and transferring [26]. Sharing synthetic data enables out-of-organization researchers to access data similar to the real data without the risk or regulations that come with sharing real data.

Synthetic data can be generated and distributed across domains, not necessarily just healthcare. This allows for many researchers to build models and algorithms to use on the real data once returned to the original data holder [17].

Patient files can vary across systems and even within the same system. For example, multiple patients may have the same type of appointment in a hospital, then have the same type of lab work done in the same building. However, the data recorded in the files may vary. This variance is wasteful, harmful to patients, and reduces the speed of care access to patients [29].

Having a third-party platform to create synthetic data installed on an institution's data storage systems can reduce data ownership concerns.The data created from a proprietary software can often be combined and shared across institutional boundaries. Sharing synthetic data lessens legal and ethical barriers traditionally encountered in sharing real patient data [8].

2.7 Challenges and Limitations for Synthetic Data

No agreement has been formally established on how a synthetic dataset should be generated. Researchers note that defining domain and model neutral evaluation for synthetic data generation models is both important but has yet to be universally agreed-upon [1]. As early as 2018 there hasn't been any approval from the Food Drug Administration in the United States on guidance of using synthetic datasets for studies [8]. To complicate matters further, we do not yet know if drug responsiveness studies are better predicted using fully synthetic versus even a partially synthetic dataset [8].

Most of the research on the validity of synthetic data has been focused on structured data. More research on the validity of synthetic medical imaging and natural language processing (NLP) data is called for [26].

As a recurring theme: bad data generates bad data or as is heard frequently in the realm of data, "garbage in, garbage out". This is also true when generating synthetic data from ground truth data. Certain general trends of a dataset can be replicated in the synthetic dataset [8]. If the data being fed to the synthetic generation model is flawed, biased, or otherwise captures a trend, these respective properties will likewise be reflected in the resulting synthetic datasets.

Even highly-structured data can also be incomplete and imperfect. Results from synthesizing data are drawn directly from the original data [26]. Missing values can cause a limitation to generating synthetic data. Because imputation of missing values can vary from researcher to researcher, inconsistencies may arise in how the data is generated either by researcher input or algorithmic decision [26]. Missing data, which is quite commonly found

in medical datasets, can have a myriad of effects on how synthetic data from the dataset is generated. During the modeling process, missing data must be accounted for. Missing data may have value in itself or be a part of a greater pattern of missing data. When put to the test, Tucker found that distributions are generally closer to the original data when missing data is preserved and included in the model [34].

Low sample size, high sparsity, high dimensionality, and highly irregular distributions can affect the resulting generated synthetic data and also how it can be interpreted compared to the real data [9]. Small populations can be a challenge for synthetic data generation because they can limit the quality of statistical characteristics especially for high-dimensional multivariate distributions and outliers. In addition, a small population used as seed data can create a selection bias if there are safeguards for protecting identifiable features leading back to a specific patient in the original data [26].

Another challenge is getting the synthetically generated data to be "different enough." This is a challenge because, yet again, of the lack of any particular standards to hold synthetic (and as we've seen previously, the original data) to. The data shouldn't be so similar as to risk a breach in personal information and privacy. The aforementioned SRA technique of the van der Schaar Lab supposedly overcomes privacy issues because of a characteristic of the method where the distribution of the synthetic data doesn't have to match the actual data that it will be implemented on [27] [17].

Individual synthetic data generation methods and tools also come with their own challenges. Such as with OSIM2, some synthetic data methods cannot precisely approximate the clustered nature of encounter-based data. Models may assume that the relationships of certain features remain stable over time when in reality that may not be the case [24]. MDClone

also struggled with converting longitudinal data into a format that facilitates synthetic data generation [9].

Medical research presents plenty of unique challenges for synthetic data. The temporal nature of many types of health data can impact the production of realistic synthetic data [34].The delicacy of patient privacy exacerbates some of these challenges. Eliminating the issue of patient data re-identification via the use of synthetic data comes with the trade-off of potentially relying on domain-specific knowledge bases and curating the generated data manually [10].

The main challenge of generating high-fidelity synthetic patient data is preserving relationships, distributions, predictive capabilities, and patients' privacy [34]. Patient records are usually high-dimensional datasets and have complicated distributions [27]. This is further complicated by small numbers of people who have rare diseases or may be an outlier [27]. This makes it more of a challenge to represent the complexity of a realistic patient without duplicating a specific individual's data [27].

The "black box" problem is an additional challenge faced when generating synthetic data for medical research. The black box problem occurs when the relationships between features are not explicitly identified and biases arise from this. Unwanted correlations are not easily identifiable either. Some suggested approaches to handle the black box problem include using a probabilistic graphical model and tree-based models [34].

Multivariate categorical data in high dimensions with a dependence on the structure of the data is common in electronic health records. Though Bayesian networks could be used to smoothly handle this kind of data, Gaussian and Dirichlet mixture models are more flexible non-parametric models that don't necessarily require the aforementioned dependence

structure [10].

2.8 Propositions

An analysis of existing synthetic generation tools and methods reveal that some may be an accessible option for clinical researchers at UC Davis. Furthermore, an analysis of experiences, thoughts, and opinions on synthetic data use in healthcare research reveals how DataLab can support researchers at UC Davis. By experimenting on a set of use cases with some of these tools, we can affirm whether they are viable solutions to offer our medical researchers as they come to DataLab for assistance with generating synthetic data.

2.9 Scope

The scope of this study covers what a researcher comting to the UC Davis DataLab would expect to benefit from.

Open source promotes an advancement of common, community-driven computing. Most open source tools and repositories are free to use and contribute code to. Because of this, using an open source tool would be cost-effective and time-effective as most proprietary tools do not list their pricing information and require that you request a demonstration or consultation session with a non-engineer representative before moving forward.

A Python library solution would be ideal since my familiarity is with Python and the typical modern data science stack (Numpy, Pandas, and Scipy). A library that is open source and has a robust community of developers and users is optimal.

Most importantly, the tool has to have good efficacy and meet the researcher's needs. The tool will have to be flexible enough to cover the wide array of data types and relationships found in medical research.

Chapter 3

Methods

3.1 Introduction

I used three methods to carry out my research. First, I researched the tools that are available for synthetic data generation. I interviewed several clinical researchers at the University of California, Davis. Finally, I worked with two different research teams and applied a selected synthetic data generation library that meets their needs. We conferred on what metrics should be used and evaluated the resulting synthetic data.

3.2 Synthetic Data Tools Research Design

Using the literature that I reviewed, I searched for tools that can be potentially used for my use case experimentation, written into DataLab educational toolkits, and eventually adopted for use by our medical researchers. I also included some open medical datasets available on the internet that were also mentioned in the literature.

To compare the synthetic data generation tools, I maintained a Google Sheets table containing the name, a link to the tool's documentation, how the tool is procured, data requirements, whether the tool is open source or a proprietary service, any associated publications, how the tool handles data privacy, a brief description and assessment, and any additional notes.

It was sometimes challenging to track down the tools that were previously discussed in the literature. I found that some of the names of tools may have changed as well as the ownership of the tool. Some tools that were open source were purchased by companies and no longer accessible for me to test out in my research. I encountered a lot of dead links when following the links included in the bibliographies of the papers I read.

I chose this table approach to evaluate the tools because it allowed me to reduce my list to just a few viable options. This also gave me the context of tools that were general purpose or for purposes completely unrelated to medical research.

3.3 Interview Design

Interviewing researchers at UC Davis allows me to collect qualitative data in order to build a framework surrounding my research objectives to:

- Understand using synthetic data or electing not to use synthetic data from the perspective of researchers.
- Understand the context of specific situations and actions in how synthetic data fits in a researcher's personal schema.

- Understand how well researchers might know about synthetic data availability and use cases in medical research.
- Determine attitudes, stigmas, and experiences with using synthetic data for research and published studies.
- Determine researchers' familiarity with synthetic data tools, concepts, and practices.
- Determine insights into how these tools could be improved to meet the needs of researchers.
- Determine insights into how these tools and training on how to use them can be made more accessible to researchers.

I developed a series of interview questions to gain insight into the above objectives. The questions would be used to explore researchers' experience level with synthetic data.These semi-structured interview questions are as follows:

1. Once the data is retrieved, cleaned, and prepared for analysis, do you have enough of it?
2. Does the amount of available data ever discourage certain studies?
3. What options do the researchers have if there is not enough data for a study?
4. Have you heard about synthetic data use in research?
5. Have you ever used synthetic data?
6. How did you generate your synthetic data?

7. Were there any challenges to generating your synthetic data?
8. Is there anything you know now that you wished you would have known when you began using synthetic data?
9. Why did you use synthetic data?
10. Did you generate synthetic data from data you already had?
11. Is pure synthetic data enough to use in a study?
12. Do you use a hybrid of synthetic and actual data?
13. What validation do you need for synthetic data to be useful?
14. If you feel that synthetic data is not ideal, what would make it "good enough"?
15. In your opinion, what are the pros and cons of using synthetic data?
16. Do you have any cultural reasons not to use synthetic data (i.e. stigma)?
17. Do other researchers frown upon using synthetic data?
18. Can you publish your research if you use synthetic data?
19. If you knew that there are cited studies that have used synthetic data would you be more inclined to use it?
20. Could we have gained something (during COVID) by using synthetic data when there wasn't enough in time?
21. Clinical vs. Research – would you use synthetic data in one but not the other?

After developing my interview questions, I trained and applied for IRB approval for human subject research. Following the approval (IRB number: 1751771-1), I networked to find UC Davis medical researchers to be interviewed.

To perform the interviews, meetings were conducted over Zoom, a secure video and voice conferencing software.This software was necessary because the research took place during the COVID-19 pandemic.

Tracking down willing medical researchers that work with large amounts of data who actually have time to sit on a Zoom call for thirty minutes to an hour was a challenge. Most of my sample of research participants were referred to me by each other. The sample is not very representative of the entire population of medical researchers at UC Davis nor is it very large.

Going into this part of my research, I expected a great deal of skepticism of synthetic data use in medical research based off of what I learned in the literature review. I also hypothesized that researchers would be moderately familiar with synthetic data tools.

Sample

I interviewed eight different researchers associated with the University of California, Davis, in the School of Medicine. For the researchers that I interviewed, the specialities and areas of research are outlined in the included table.

Subject	Specialty	Title(s)	Date of Interview
A	Vascular neurological emergencies	Assistant Clinical Professor, MD	July 8th, 2021
B	Pulmonology, Critical care, Internal medicine	Assistant Professor of Clinical Medicine, MD, MAS	July 9th, 2021
C	Critical care	Clinician, Clinical Researcher, MD	July 26th 2021
D	Point of care, novel technologies	Clinical Nurse Scientist, PhD, RN	July 30th, 2021
E	Quality improvement, Critical care	Assistant Adjunct Professor, PhD, MBA, MSN	November 1st, 2021
F	Clinical data	Senior Data Engineer, Data Scientist, MS	November 2nd, 2021
G	Psychiatry, Behavioral Medicine	Leadership role, MD	November 9th 2021
H	Vascular and Radiology Interventional Radiology	Assistant Professor, MD, PhD	January 3rd 2022

Table 3.1: Subject list of researchers that were interviewed.

Interview Methodological Assumptions & Limitations

There are a few notable assumptions and limitations to my interview subject list. First, all researchers interviewed are associated with the UC Davis School of Medicine. This means that their experiences and the scope of the interviews are going to be limited to this particular school in this particular university. Though some of the researchers work with data across the different University of California campuses and with some start-up companies in the neighboring San Francisco Bay Area, the experiences of the researchers are assumed to be

linked geographically to California.

Second, as mentioned previously, this is only a small sample of clinical researchers. UC Davis Health & the School of Medicine has hundreds of clinical researchers from a myriad of specialties. I was only able to connect with eight busy researchers. Some of the researchers in this group were actually recommended by each other. The close connection between this particular group of researchers could be of concern if opinions shared between researchers are based on familiarity or proximity.

3.4 Use Case Design

The purpose of experimenting with two different use cases is to investigate the practical approaches to synthetic data generation for two approaches:

1. How fully generated data based off of parameters performs for the use case.

2. How generated data to augment existing datasets performs for the use case.

Synthetic Data Vault (SDV)

I chose the Python library, Synthetic Data Vault, or SDV to generate synthetic datasets for my use case teams. SDV was easy to install into my Python Anaconda environment, has fairly good documentation, is open source with a lively community accessible by Slack, a communication chat platform commonly used by technology teams and workplaces. SDV was originally developed by the Massachusetts Institute of Technology (MIT) Data to AI Lab in 2016.

SDV has four models available for single tabular data:

GaussianCopula (Sometimes just referred to as "the Gaussian model" throughout my thesis) Copulas are functions that describe a joint distribution of multiple random variables by analyzing the dependencies between their marginal distributions [20].

TVAE Tabular variational autoencoder (TVAE) is an autoencoding method for mixed-type tabular data generation. TVAE differs from existing VAE models by the alteration of the loss function [37].

CTAN Conditional Tabular Generative Adversarial Network (CTGAN) is a GAN method that models tabular data distributions and rows from the original data's distribution. CTGAN uses mode-specific normalization [37].

CopulaGAN CopulaGAN is a modification of the CTGAN model. It uses a cumulative distribution function (CDF) based transformation that the GaussianCopulas apply to facilitate learning the data by the underlying CTGAN model [20].

SDV also has an extensive testing suite that offers many metrics to gauge the efficacy of the synthetic data that has been generated. Below is a small sampling of some of the metrics that were most notable and potentially useful to use in the evaluation of the use case data along with some commentary and concerns surrounding the evaluation approaches.

Some of the choice statistical approaches included:

Kolmogorov-Smirnov test This metric uses the two-sample Kolmogorov-Smirnov (KS) test to compare distributions of columns with continuous values using the empirical cumulative distribution function (CDF). The output for each column is 1 minus the

KS test statistic. This score indicates the maximum distance between the expected CDF and the observed CDF values [20].

For the KS test, the value returned is 1-D where D is the KS test statistic. This statistic is a measure of the maximum distance between the two distributions. Since this works on probabilities and is an absolute value it must be between 0 and 1. D will equal 0 if the two distributions are identical. Therefore, for output from the KS test, numbers close to 1 are viable values.

Chi-Squared test This metric uses the Chi-Squared test to compare the distributions of two discrete columns. The output for each column is the CSTest p-value, which indicates the probability of the two columns having been sampled from the same distribution [20].

When using SDV's CS test, p-values are returned and the CS test is testing the null hypothesis that the distributions across categories is the same between the original and synthetic datasets. This may be a bit dubious due to sometimes with very large sample sizes the null hypothesis can be rejected for small, clinically non-significant deviations while for small sample sizes the null hypothesis can easily fail to be rejected due to lack of power. This could cause clinically meaningful deviations to be missed. However, SDV's CS test is still simple to calculate and is a familiar statistic.

The CS test as it is provided in SDV's evaluation can be used to assess distributional similarity but should be interpreted within the context of the observed frequencies.

Both the KS test and CS test assess whether the distributions are similar between the synthetic and real data.KS tests are applicable to numeric data while the CS test is for

categorical variables. According to the SDV documentation, if the test is provided with two data sets it will automatically only apply the KS test to numeric variables and CS test to categorical variables [20]. The documentation does not specify how the test will determine whether your features are categorical or numeric. We can assume that data types that are a float or integer, for example, are assumed to be numeric.

Likelihood metrics attempt to fit a probabilistic model to the real data and later on evaluate the likelihood of the synthetic data on it. Some notable likelihood metrics offered by SDV are:

Bayesian Network This metric fits a Bayesian network to the real data and then evaluates the average likelihood of the rows from the synthetic data on it [20]. It was determined that the log likelihood metric below would be a better likelihood metric than the regular Bayesian network metric offered by SDV.

Log Bayesian Network This metric fits a Bayesian network to the real data and then evaluates the average log likelihood of the rows from the synthetic data on it [20].

Gaussian Mixture This metric fits multiple Gaussian Mixture (GM) models to the real data and then evaluates the average log likelihood of the synthetic data. GM models are a sort of k-means clustering that assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [25].

If these models can capture the between-variable correlation structure of the data or if they are building independent models for each variable, then the offered likelihood metrics will be both viable and valuable for evaluation the synthetic data/ However, there is not

enough information in the documentation nor in any of the documentation's citations with additional info that would help to assess this.

Detection metrics measure how hard it is to distinguish the synthetic data from the real data by using a machine learning model. The metrics mix the original data and the generated synthetic data together with flags indicating whether the data is real or synthetic. It then cross-validates a machine learning model that will attempt to predict the flag. The output of the metrics will be the 1 minus the area under the average receiver operating characteristic curve (ROC) across all the cross-validation splits [20].

Logistic Regression Classification A detection metric based on a Logistic regression classifier from scikit-learn, a free machine learning library in Python.

SVC Classification A detection metric based on a Support Vector Classifier (SVC) also from scikit-learn.

A limitation and concern of the detection metrics that SDV offers is the uncertainty of what the default settings are. The documentation for the evaluation does not go into further detail. For example, are they fitting a straight logistic regression or are they using a penalized approach such as ridge, LASSO, or elastic net? If SDV is using a straight logistic method, is it using a stepwise procedure based on p-values? If so, how does it handle high correlation among variables? If using a penalized approach, which approach are they using? Furthermore, which cross-validation approach is being used to select penalty parameters? Is there a training and testing split or is it using the full data?

Finally, SDV also offers machine learning efficacy metrics. These metrics are evaluated by trying to solve a machine learning problem and can only be used on datasets that contain

a target column that can be predicted using the rest of the data [20].

Other Python Libraries and Tools

In addition to using the SDV library to generate synthetic data, I used several other libraries commonly used in the Python data science stack. I performed my coding with Jupyter Notebooks [19] in an Anaconda Python 3 environment.

Pandas Pandas is an open source community-supported data science and analytics library for Python. Pandas gives access to the DataFrame object which is a flexible table-like data container. Pandas allows for easy reading and writing of data from or into a variety of different file types. It also makes reshaping, filling, computing data easy and perfect for this research [33].

NumPy Though used a bit less frequently in this endeavor, NumPy is an open source library that enables numerical computing with Python. It allows use of the NumPy array data type [11].

Matplotlib Matplotlib is a Python visualization library that facilitates rapid plotting of data within the Jupyter Notebook [14].

Plotly Plotly is a free and open source Python graphing library [15]. The interactive capabilities of using Plotly visualizations facilitated exploration of the data.

Liver Oncology

The liver oncology data comes from the UC Davis DataPath, a standardized warehouse maintained by UC Davis. It is a de-identified dataset with features including:

- An identifier, gender, race, ethnicity, age, status of cancer, status of diabetes, HIV status, hypertension status, and obesity status
- Several kidney measurements such as blood urea nitrogen and creatinine
- Information about the patients' electrolyte levels including platelet count, sodium, and potassium
- Liver metrics such as Alkaline phosphatase (ALP), Alanine aminotransferase (ALT), Aspartame amino transferase (AST), and bilirubin
- One fasting lipid profile, cholesterol
- Liver scarring such as fibrosis-4 (fib-4) metrics
- Finally, a column indicating whether or not the patient developed cancer

The goal of this use case is to create a balanced outcome dataset based on the demographics represented in the original data. Structuring the division of the data generation and essentially experimenting with how the tool can be used is the method for achieving this. The team acknowledges the limitation of losing precision but in turn gaining an increased N-value for certain population demographics. The demographics represented could ideally be any feature column. In this study, race, ethnicity, and gender will be the primary focus of how the data is to be balanced.

To balance the data by structuring how the SDV tool is used, I use the GaussianCopula model. As discussed in the literature, GAN methods would not be the ideal method to use since the data is tabular and not, for example, composed of images. The general procedure I created for achieving a balanced dataset is as follows:

Algorithm 1 Balancing demographic data using SDV

Create an empty list for the ending DataFrame

for each demographic in a list of possibilities for that demographic **do**

Get all cells within the DataFrame where the current demographic is a match

Create a new GaussianCopula model

Fit the model to the data based on the current demographic

Sample from the model ▷ Sample size = number of rows / number of possibilities

Append the sample above to the empty list from outside the loop

end for

Concatenate each of the DataFrame samples within the list of DataFrames

The liver oncology team generally evaluates their data by considering recall, precision, sensitivity, specificity, accuracy, F-measure, and measuring Cohen's kappa. Though the efficacy outcome is less important than whether a balanced population can be achieved, in the experimentation, I evaluate how similar the synthetic data is compared to the original data in addition to whether the target outcome of whether the patient developed cancer or not is captured accurately in the synthetic data.

Postoperative Respiratory Failure (PRF)

Postoperative Respiratory Failure (PRF) is a pulmonary complication that can happen after a patient goes through surgery. Clinical events seem to be a better indicator to recognize early signs of PRF versus the actual blood gas exchange measurement that characterizes this complication [4]. This kind of research is hindered due to the rareness of the complication. The patient population is not great enough to generalize findings by clinical models.

That being said, the goals of this use case are to be able to augment rare case data and see how faithful each model is to the original data and to find out what parameters and changes within using the model can be made to get better results. The original dataset contains only 828 rows and the research team requests to see how the synthetic data changes at double the amount of data.

From the literature, we already know that GANs do not perform as well when replicating tabular, non-image data. However, SDV's own CTGAN claims to perform just fine when generating tabular synthetic data. To put this to the test, for the PRF use case team, all four SDV models are used to generate synthetic data to augment the original sample size. The best-performing model out of the four will be further modified and the results will be evaluated once again.

In addition to exploring whether SDV can meet the needs of our researchers, doing experiments demonstrates how amenable the software is to changes to get us into what is acceptable for those needs. There are several parameters that can be altered when using the SDV models. For example, in the GaussianCopula model, customization can include setting transforms, setting bounds, specifying rounding for numerical columns, exploring probability

distributions, setting distributions for individual variables, and conditional sampling [20].

The dataset in this use case is a small de-identified table of patient data that includes metrics such as:

- An indication of whether the data is case or control data.
- Qualitative demographic information such as sex, combined race & ethnicity and quantitative information about the patient's age, weight in kilograms, and height in centimeters.
- Primary payer insurance information. These values included Medicare, private insurance (including military), Medicaid, and "other".
- The patient's ASA Physical Status Classification [7]. This system helps assess and communicate a patient's pre-anesthesia medical comorbidities. The classification values found in the dataset were ASA I, II, III, IV, V, and the lack thereof indicated by a `nan` ("not a number").

  ASA I A normal healthy patient

  ASA II A patient with mild systemic disease

  ASA III A patient with severe systemic disease

  ASA IV A patient with severe systemic disease that is a constant threat to life

  ASA V A moribund patient who is not expected to survive without the operation

- Alcohol use and smoking indication.

- Functional status indicating if the patient is independent or partially or totally dependent at home prior to surgery.
- A column each with a `String` indicator of whether or not the patient has: asthma, kidney disease, chronic obstructive pulmonary disease (COPD), cardiac disease, dementia, diabetes, dysphagia, dyspnea, gastroesophageal reflux disease (GERD), heart failure, hypertension, impaired sensorium, liver disease, neurologic disease, sleep apnea, and recent unexplained weight loss.
- Measures of hemoglobin, creatinine, and albumin.

Efficacy of the synthetic data for each model is measured with KS testing, mean squared difference of the correlations, summary statistics with input from the team's domain expert, and pairwise correlation. I also use SDV's logistic detection metric which provides a normalized score (from [0,1]). This is a custom score that matches how the other metrics demonstrate a "better" score with a higher (closer to 1) score.

Chapter 4

Findings

4.1 Tool Comparison Findings

I found that were numerous synthetic data tools including proprietary platforms, open source toolkits, and downloadable datasets. Strangely, a handful of tools disappeared or became defunct from the time I started my research in the end of 2020 to the time of writing this thesis in 2022. This chapter contains a few tables with the results of my exploration into what tools exist for the synthetic data generation needs of researchers.

Table 4.1 contains the tools and datasets that I examined in detail. This table contains the name of the tool, how the tool is procured, any data requirements needed to use the tool, whether the tool is open source or not and which license it contains, any associated publications used in the literature review, privacy information if provided, and a number that corresponds to the numbered notes that follows the table.

Name of Tool	Procurement	Data Requirements	Open Source?	Associated Publications	Privacy	Notes
Medicare Coverage Database	Download via website	Dataset	Limited free use	Commonly cited & published	N/A	1
SynPUF	Download via website	Dataset	Limited free use	Some citations	N/A	2
Medidata (formerly Accelys)	Contact sales team	N/A	No	[32]	Not responsible for customer privacy	3
Accelario	Contact sales team	Upload pre-existing schema	No	None	Offers safeguards	4
Eunomia	The Comprehensive R Archive Network (CRAN)	Dataset runs in R instance	Yes Apache License 2.0	N/A	N/A	5
MDClone	Contact sales team	Unsure	No	[26][9]	Unsure	6
Synthea	Download & run with Java	None	Yes Apache License 2.0	N/A	[35][26][29]	7
SDGym	PyPi	Benchmarking framework	Yes MIT License	N/A	N/A	8
Synthetic Data Vault (SDV)	PyPi	Input data required	Yes MIT License	[37]	User dependent	9

Table 4.1: Initial set of tools and datasets examined.

1. **Medicare Coverage Database** Contains Medicare claims data. Use is limited to use in Medicare, Medicaid or other programs administered by the Centers for Medicare and Medicaid Services (CMS) [22].

2. **Medicare Claims Synthetic Public Use Files (SynPUFs)** SynPUF data may be used to develop software applications, train researchers, and support safe data mining operations. The data structure of the Medicare SynPUFs is very similar to the CMS Limited Data Sets, but with a smaller number of variables [23].

3. **Biovia & Medidata** (Formerly known as Accelys) A product sold by the company, Dassault Systemes. Life sciences database and platform solution. Offers synthetic data for clinical trials via Synthetic Control Arm tool [31].

4. **Accelario** Can be integrated with Oracle. Does not disclose method of generating synthetic data. Couldn't find any publications specifically that used the synthetic data generation tools.

5. **Eunomia** No longer exists in CRAN. Originally an R package for testing and demonstration. Contained a general CDM package. Part of the HADES product of The Observational Health Data Sciences and Informatics (OHDSI).

6. **MDClone** Not much information is given on their website. They do however keep a list that is easy to access of publications using data from MDClone.

7. **Synthea** Generates populations of synthetic patients with extensive health history. Extremely easy to set-up and use with basic Java knowledge.

8. **SDGym** A benchmarking framework for synthetic data. Not a synthetic data generation tool itself. By the same laboratory that SDV came from.

9. **Synthetic Data Vault (SDV)** Generates synthetic data based off of four models. Completely open source. Thriving and accessible community on Slack. Can be used with tabular, multi-table, and timeseries data. Quite flexible. Decently documented. Domain-agnostic.

In comparing the tools, I downloaded and tested out some of the open source ones available including SDV and Synthea. SDV was accesisble enough that I found it could be used in further experimentation (see use cases). Instead of established a use case for Synthea, I presented a demonstration through the UC Davis DataLab on how to install and use the software.

Start-up culture seems to have also affected the niche of synthetic data generation. I found a myriad of mostly-proprietary software companies offering these synthetic data solutions. The trend of not-disclosing how the synthetic data generation tool works is apparent on the websites of most of these solutions.

The following tables are a directory of the aforementioned solutions. Table 4.2 contains tools that generates structured synthetic data whereas table 4.3 contains tools that generate unstructured synthetic data. These solutions will be discussed further later on in this thesis.

Structured synthetic data platforms & tools			
Sogeti ADA	Diveplane	Datomize	Geminai
Facteus	Generatrix	Gretel	Hazy
Instill AI	Kymera Labs	KerusCloud	MDClone
Mirray.AI	Mostly.AI	Octopize	Oscillate.AI
Pionic.AI	Informatica	Sarus	Statice
Syndata	Syntegra	Synthesized	Syntheticus
Synthetig	Syntho	Tonic	YData
Veil.AI	BizData	Curiousity	Synth
ExactData	GenRocket	iData	
The Synthetic Data Generator	Test Data Manager		
Replica Analytics			

Table 4.2: Additional structured synthetic data platforms & tools that have yet to be examined.

Unstructured synthetic data platforms & tools			
Alreverie	Anyverse	Autonom AI	Bifrost
Cvedia	Coohom Cloud	Datagen	DeepVision Data
edgecase.ai	Lexset	mindtech	neurolabs
Oneview	Parallel Domain	Neuromation	Reinvent Systems
Rendered.AI	Scale	Sky Engine	Simerse
Synthetaic	SD	Synthetik	Synthesis AI
Vypno	Yuva AI	ZumoLabs	

Table 4.3: Additional unstructured synthetic data platforms & tools that have yet to be examined.

	A	B	C	D	E	F	G	H
Have you previously heard about "synthetic data"?	No	No	Yes	No	No	Yes	No	Yes
Have you ever used synthetic data?	No	No	No	No	No	Yes	No	Yes
Have you ever used "imputed" or "simulated" data?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Do you often find yourself without enough data for your study?	Yes	Yes	Yes	Yes	Yes	Yes	Sometimes	Yes
Do you think research using synthetic data can be published?	Yes	Yes	Yes	Yes	Yes	Unsure	Unsure	Yes

Table 4.4: A sampling of interview findings

4.2 Interview Findings

Further discussion of the interviews can be found in chapter 5 of this thesis. The following findings from the interview are more of a presentation of the results as a survey.

As seen in table 4.4, out of eight researchers, three had previously heard about synthetic data generation. One of these three researchers had only heard about synthetic data before from a recent presentation hosted by the UC Davis DataLab. Only two out of the eight researchers said that they have worked with synthetic data before. However, all eight researchers recalled using "imputed" or "simulated" data in their research. All but one of the eight researchers recall often finding that they do not have enough data to embark on a study right away. The one researcher that did not say yes, instead replied that they find they do not have enough data on a case-by-case basis ("sometimes"). Six out of eight researchers believe that studies using synthetic data can be published in research journals. Two out of

Data Input	Column to Balance	SDV Model	Demographic Result	Outcome Result
Original data	Race	Gaussian	Perfectly balanced	Lost fidelity, dropped category
Original data	Outcome	Gaussian	Lost fidelity, dropped categories	Perfectly balanced
Original data	Outcome	TVAE	Lost fidelity but kept categories	Perfectly balanced
Synthetic data (result from (previous row)	Race	TVAE	Perfectly balanced	Nearly balanced

Table 4.5: Results of synthetic liver oncology datasets

eight researchers were unsure if studies using synthetic data could be published in this way.

4.3 Use Case Findings

Liver Oncology: Balanced Data

The original liver oncology dataset has 4,272 rows of data. Out of these 4,272 patients, 173 patients (4.0%) developed cancer and 4,099 patients (96.0%) did not develop cancer (see right bar in figure 4.1). For the racial composition of the original dataset as illustrated in the right bar of figure 4.2, 67 (1.6%) patients were identified as American Indian or Alaska Native, 294 (6.9 %) were Asian, 292 (6.8%) were Black or African American, only 36 (0.8%) were Native Hawaiian or other Pacific Islander, 2,745 (64.3%) were White, 220 (5.1%) patients were of unknown race, and 618 (14.5%) were denoted as "other".

After using the GaussianCopula model to generate a new dataset of the same size based

off of the original data, the new data was examined to see if the statistics were maintained.

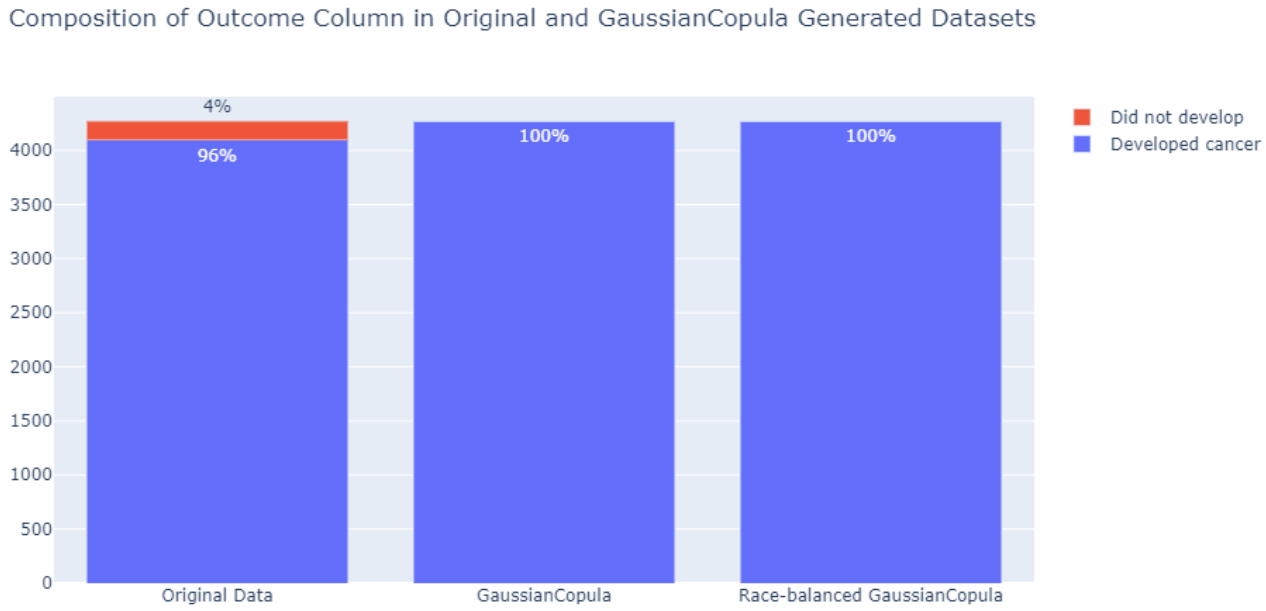


Figure 4.1: **Left:** Patient outcomes in the original data. **Middle:** Balance of patient outcomes in resulting GaussianCopula model synthetic dataset. **Right:** While balancing the race composition, the GaussianCopula model did not pick up on the original compositon of outcomes. **Note** how the Gaussian model did not pick up on the original composition of the patient outcomes in the original dataset.

For patient outcomes of whether the patient did or did not develop cancer, the original statistics were not maintained. As pictured in the bottom of figure 4.1, the model was not able to mimic the split between 4.0% of patients who developed cancer and the 96.0% of patients who did not. Instead, the model produced a dataset where 100% of the patients did not develop cancer.

However, when it came to the racial statistics of the new dataset, the original composition matched perfectly with 67 (1.6%) patients were identified as American Indian or Alaska Native, 294 (6.9 %) were Asian, 292 (6.8%) were Black or African American, only 36 (0.8%) were Native Hawaiian or other Pacific Islander, 2,745 (64.3%) were White, 220 (5.1%) pa-

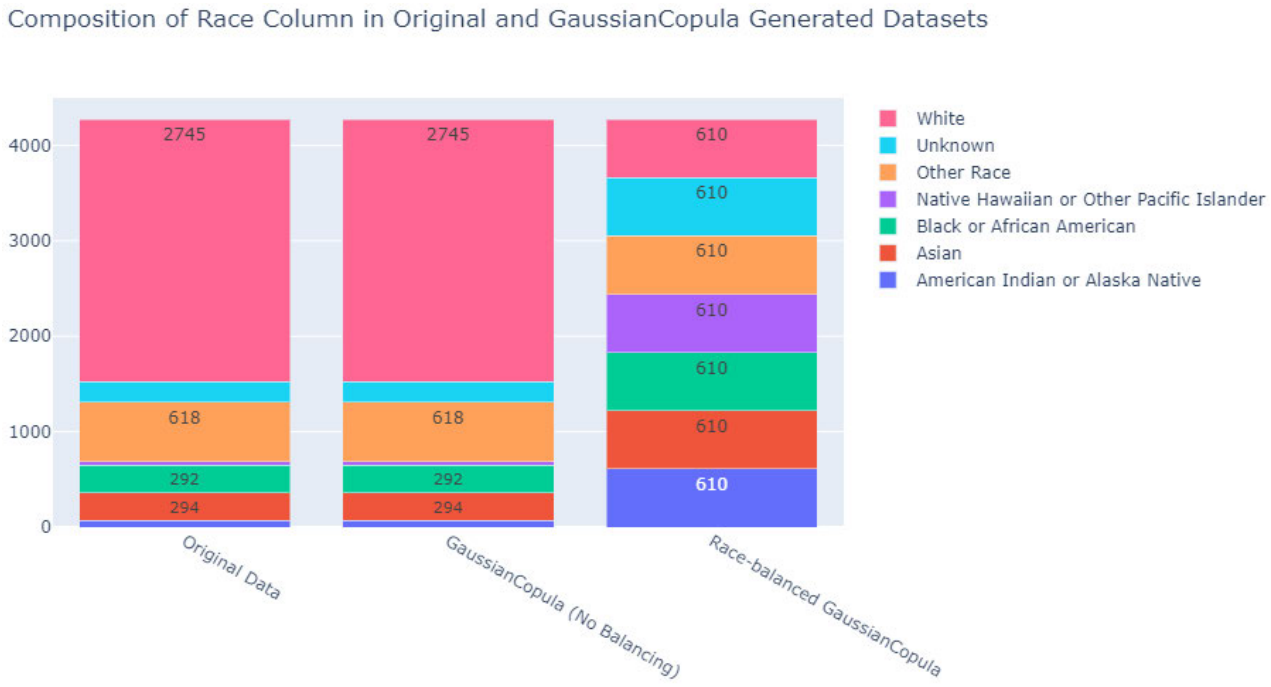


Figure 4.2: **Left:** Composition of patient race demographics in original dataset. **Middle:** Composition of patient race demographics in the GaussianCopula model synthetic dataset with no balancing applied. **Note** how the Gaussian model was able to mimic the composition exactly in the resulting synthetic dataset. **Right:** Race-balanced synthetic dataset using the GaussianCopula model.

tients were of unknown race, and 618 (14.5%) were denoted as "other" shown in the middle bar of figure 4.2

Using the algorithm outlined in chapter 3 and choosing race as the category to balance, the resulting dataset (see right bar of figure 4.2) produced seven equally balanced groups of 610 patients with each group contributing to 14.3% of the new dataset. But just as the model produced a dataset previously in which 100% of the generated patients were classified as "Did not develop" cancer, this dataset unfortunately also disregarded the original statistics of the outcome column (see right bar of figure 4.1).

When generating a synthetic dataset using the GaussianCopula model with the goal of

balancing the outcome column so that patients who did develop cancer and patients who did not develop cancer constitute an even 50% split each (such as in the right bar of figure 4.3), the composition of the racial representation of the dataset became more unbalanced than the original data (see bottom figure 4.4).

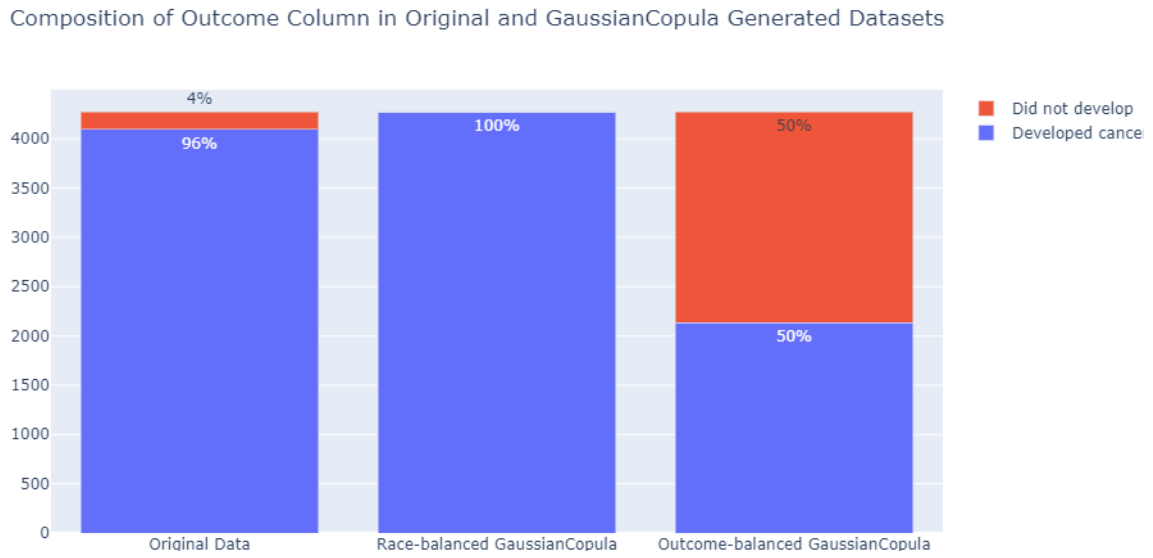


Figure 4.3: **Left:** Original data composition of outcome column. **Middle:** Composition of the race-balanced GaussianCopula model outcome column. **Right:** Outcome-balanced outcome column using the GaussianCopula model. **Note:** how the race feature dropped two categories (Other Race and Unknown) as well as the overall change in composition.

Significantly different results were achieved by using the TVAE model from SDV instead of the GaussianCopula model. When using the algorithm in conjunction with the TVAE model to balance by patient outcome (50% developed cancer, 50% did not develop cancer), all seven races that were in the original data were represented in the resulting balanced synthetic dataset. Note that when using the GaussianCopula model to balance outcome data, the GaussianCopula model was only able to represent five (White at 72.7%, Native Hawaiian or Other Pacific Islander at 0.3%, Black or African American at 13.2%, Asian

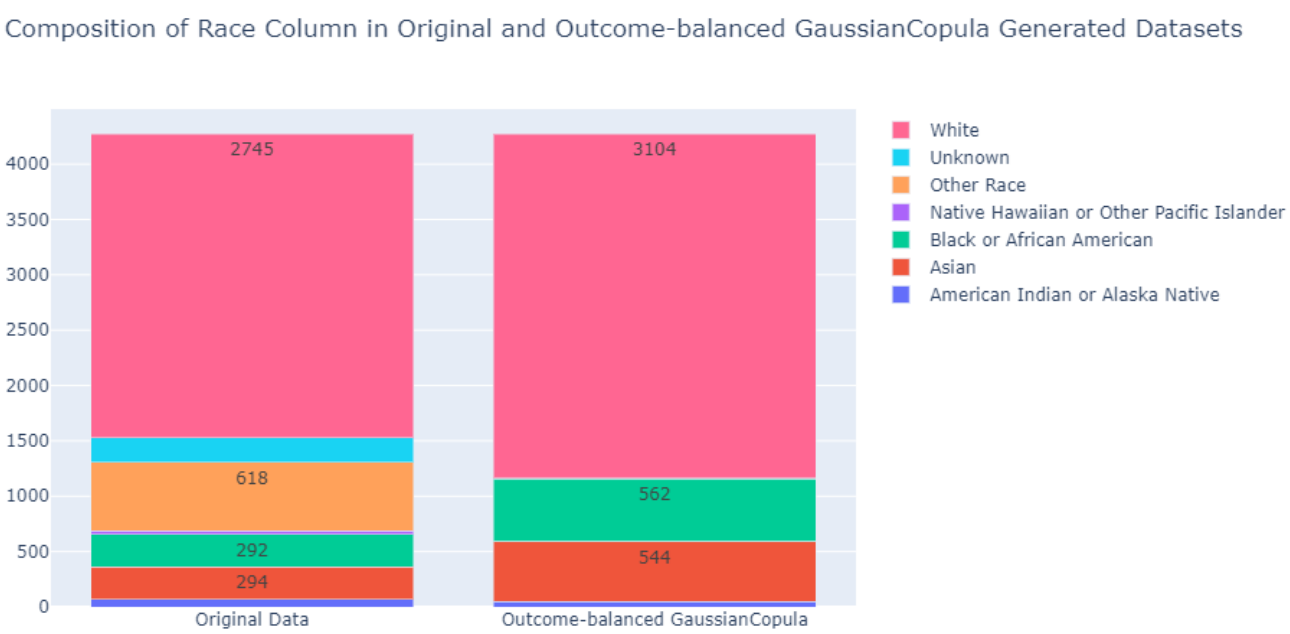
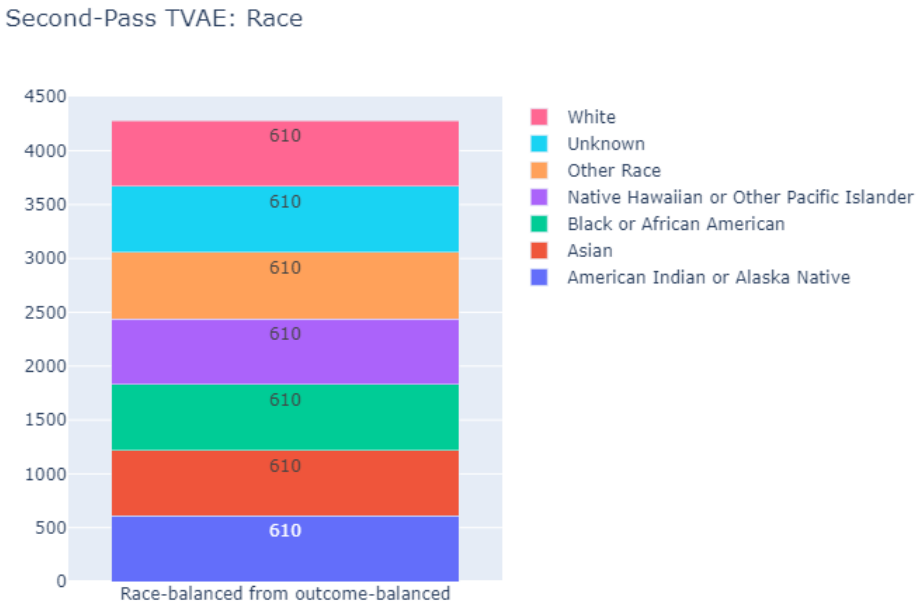


Figure 4.4: Note how the race feature dropped two categories (Other Race and Unknown) as well as the overall change in composition.

at 12.7%, and American Indian or Alaska Native at 1.1%). Though the composition of all seven categories within the race column are changed to no longer resemble the original statistics, the fact that the TVAE model was still able to retain all seven categories is the key component for the following dataset result.

Using the aforementioned dataset (TVAE-created, balanced by patient outcome) and the balancing algorithm once again this time with race as the column to balance, a second synthetic dataset that is both demographically-balanced and a massive step in the right direction for balancing the patient outcome emerges (see figure 4.5). This dataset's patient outcome is comprised of 67.8% of synthetic patients not developing cancer and 32.2% of synthetic patients developing cancer. This is clearly no longer the 50/50 split as it was previously balanced to be, but it is significantly closer to a 50/50 split than what was



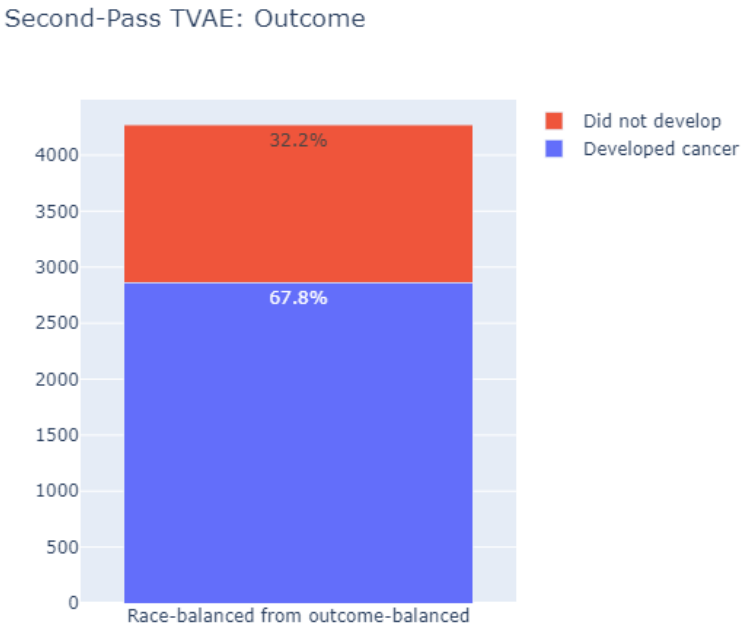


Figure 4.5: **Top:** Outcome composition of race-balanced dataset made from out-come balanced dataset using the TVAE model. Once again, the data is balanced to equally represent each category found in the race column of the dataset. **Bottom:** Race composition of race-balanced dataset made from out-come balanced dataset using the TVAE mode. In generating a race-balanced dataset, the TVAE model maintained a patient outcome column that is much closer to being fully balanced.

represented in the original dataset (96.0% did not develop cancer; 4.0% did develop cancer).

PRF: Model Performance

The following visualizations present the mean, interquartile range, minimum, maximum, and outliers found in the quantitative variables of the PRF dataset for each of the synthetic data models (GaussianCopula, TVAE, CTGAN, and CopulaGAN) and the original data. Figure 4.6 shows that the GaussianCopula model's output was the most similar out of the four models to the original data for the age of the patient population. The Gaussian model also performed well in comparison to the other models in generating heights for the synthetic patient population as seen in figure 4.9.

Figure 4.7 displays an unusual result. The original data contains outliers that none of the models were able to mimic, but all but the GaussianCopula model was able to generate. The creatinine summary statistics of figure 4.8 are also mixed. At a glance, it appears that the CopulaGAN produced the most similar results. Both hemoglobin (figure 4.10) and weight (figure 4.11) also produced mixed resulting datasets.

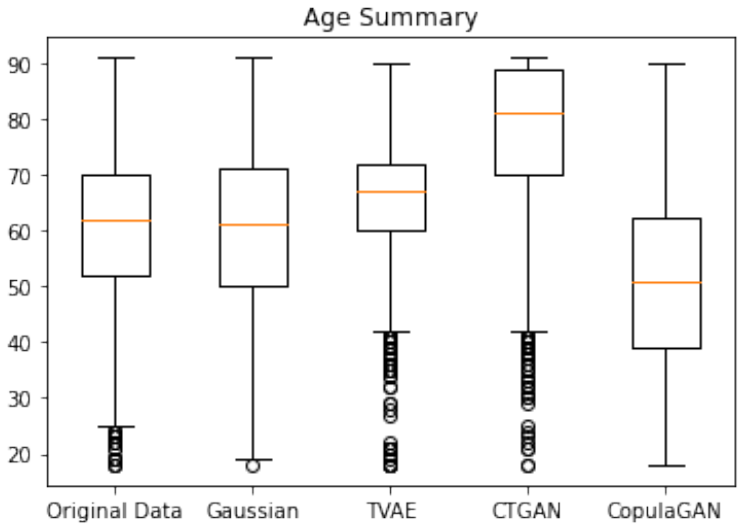


Figure 4.6: Summary statistics for the age column in the original and generated datasets.

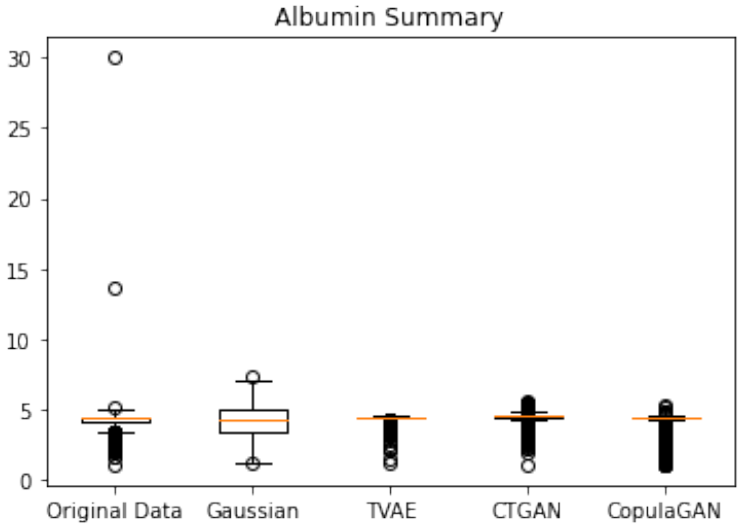


Figure 4.7: Summary statistics for the albumin column in the original and generated datasets.

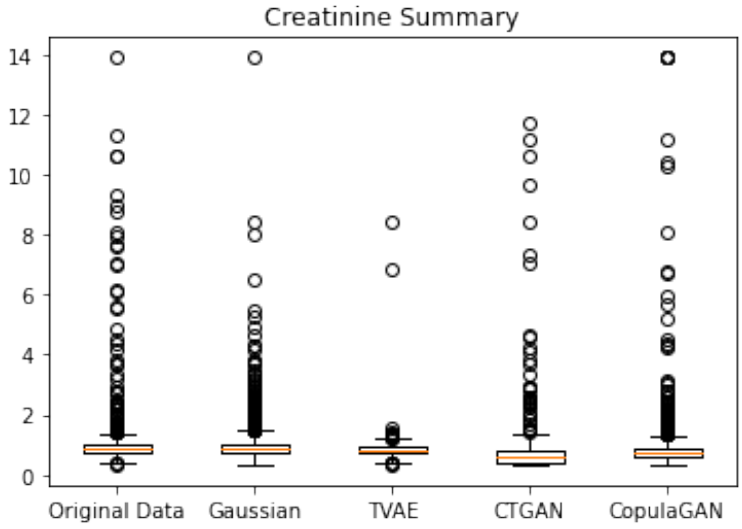


Figure 4.8: Summary statistics for the creatinine column in the original and generated datasets.

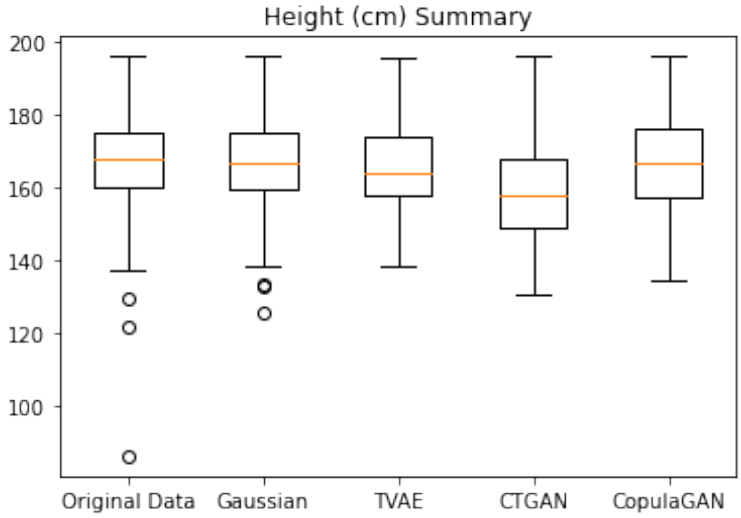


Figure 4.9: Summary statistics for the height column in the original and generated datasets.

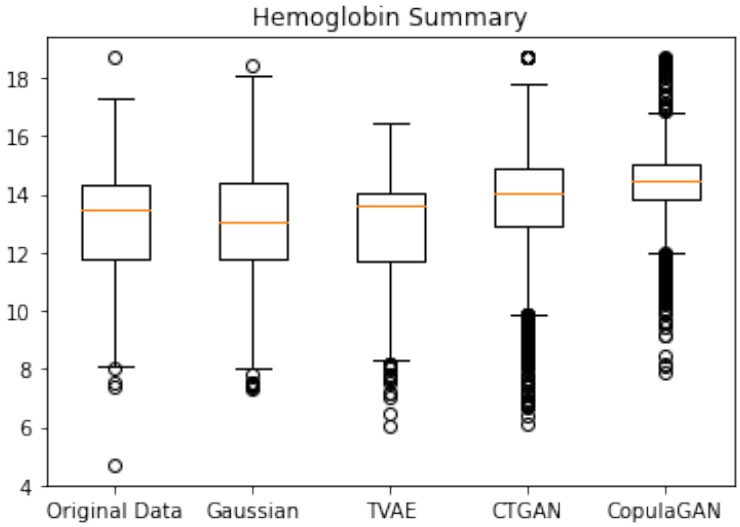


Figure 4.10: Summary statistics for the hemoglobin column in the original and generated datasets.

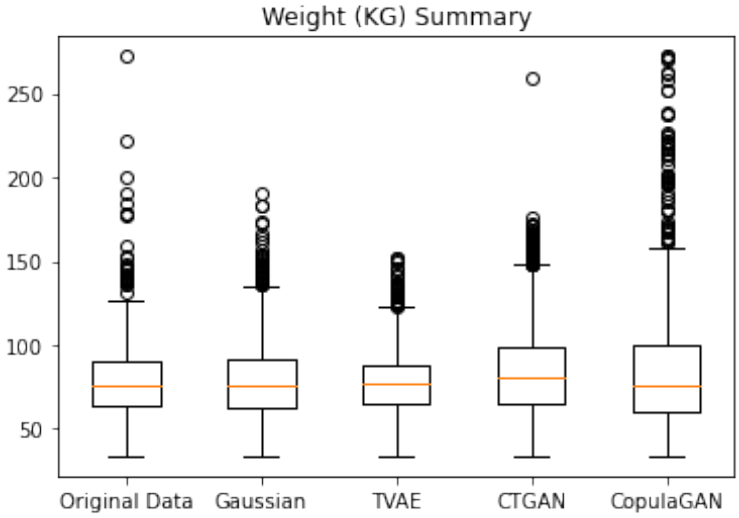


Figure 4.11: Summary statistics for the weight column in the original and generated datasets.

Figure 4.12: Age groups that make up the synthetic patient population compared to the original dataset.

In addition to the general spread of the summary statistics for the age group, the PRF team also found value in viewing how the models performed in generating patient ages based on significant groups. This is represented in figure 4.12 where it appears that the Gaussian model maintains the best mimicry when compared to the other models.

In comparing the original correlations (figure 4.13), there is a moderate (0.41) positive correlation between patient weight and height. There is a small positive correlation between hemoglobin measurement and the patient's height (0.22) and weight (0.20). Additionally there seems to be a slight (0.13) positive correlation between hemoglobin and ablumin. Finally, there is a small (-0.22) negative correlation between hemoglobin and creatinine. If the models perform well, we would expect to see these correlations show up in the resulting synthetic datasets.

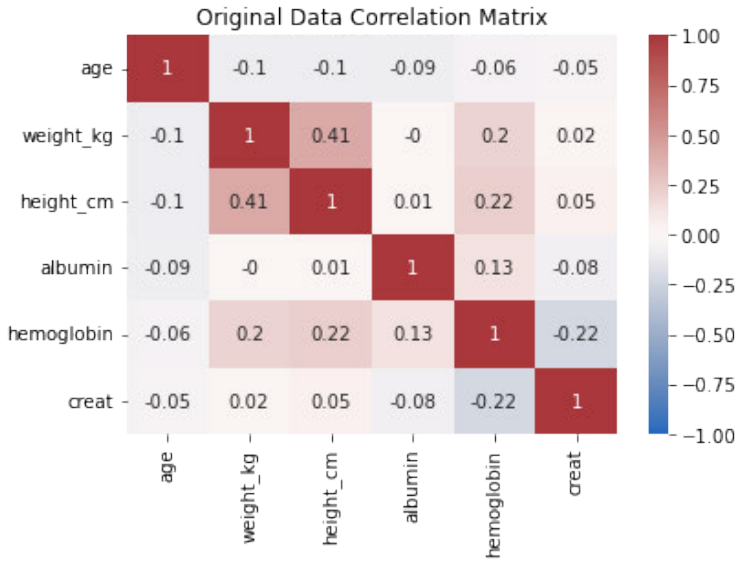


Figure 4.13: Correlation matrix of the original dataset.

The GaussianCopula model (figure 4.14) was able to mirror the positive correlations but also over-estimated the amounts that these variables were correlated. Some of these

over-estimations were greater than others. For example, the correlation between hemoglobin and albumin increases by 161% (from 0.13 to 0.34) which is over triple the amount from the ground truth dataset. In addition, the negative correlation between hemoglobin and creatinine was not significantly reflected in the GaussianCopula correlation matrix.

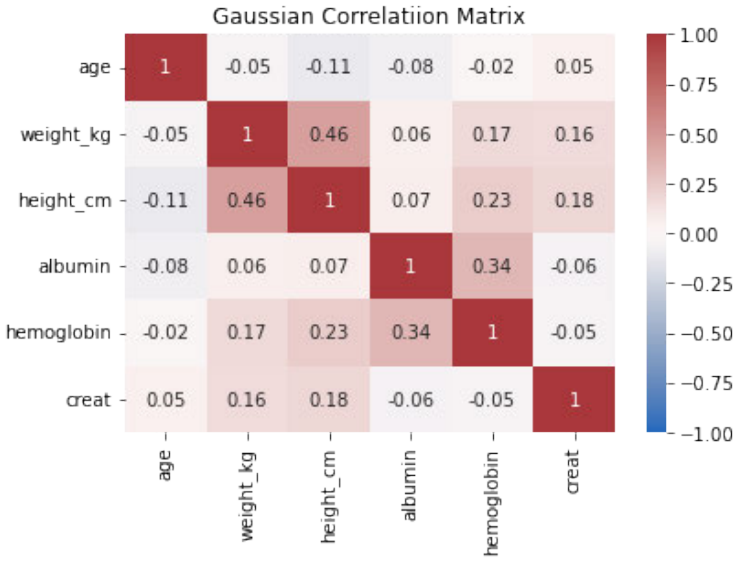


Figure 4.14: Correlation matrix of the data generated by the GaussianCopula model.

The TVAE correlation matrix (figure 4.15) shows some mixed results. Some of the postitive correlations were reflected in the TVAE dataset while others were dropped. The TVAE model overestimated a larger negative correlation where there was previously a smaller correlation but missed the negative correlation between hemoglobin and creatinine.

Both the GAN models, CTGAN (4.16) and CopulaGAN (figure 4.17), performed poorly upon observing their correlation matrices. CTGAN's data was barely correlated across the board ($\leq$ 0.11 in any direction) and the data that CopulaGAN generated was even less so ($\leq$ 0.06 in any direction).

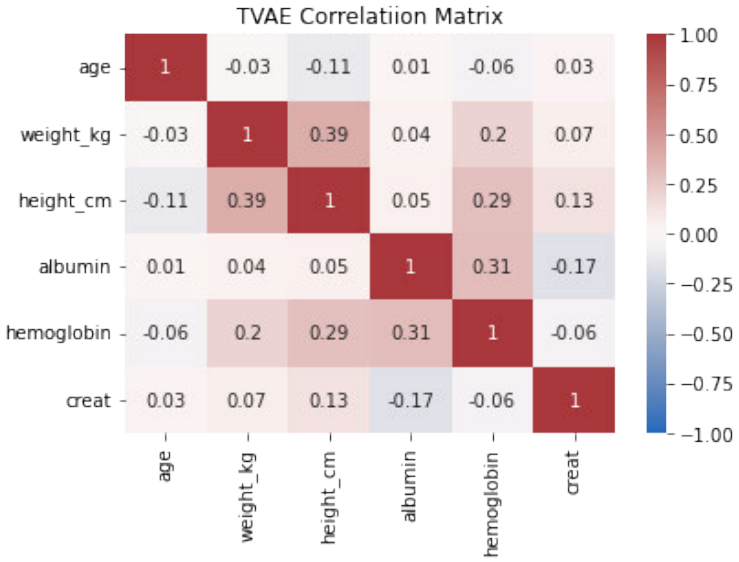


Figure 4.15: Correlation matrix of the data generated by the TVAE model.

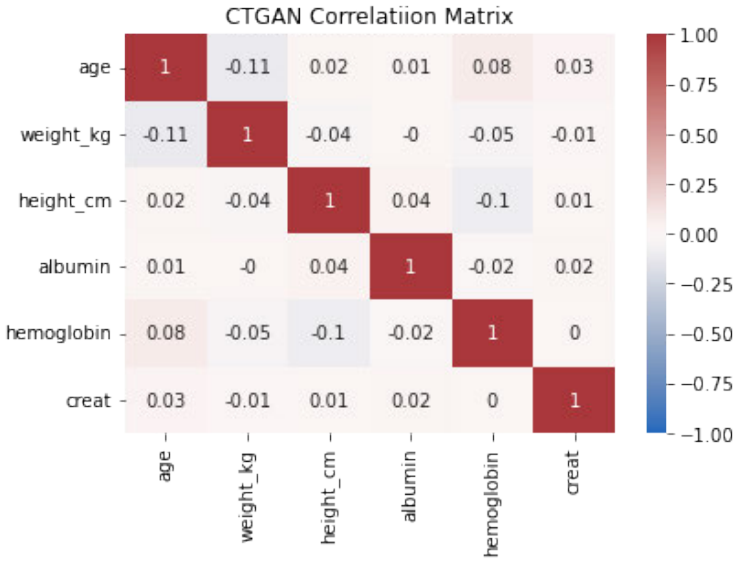


Figure 4.16: Correlation matrix of the data generated by the CTGAN model.

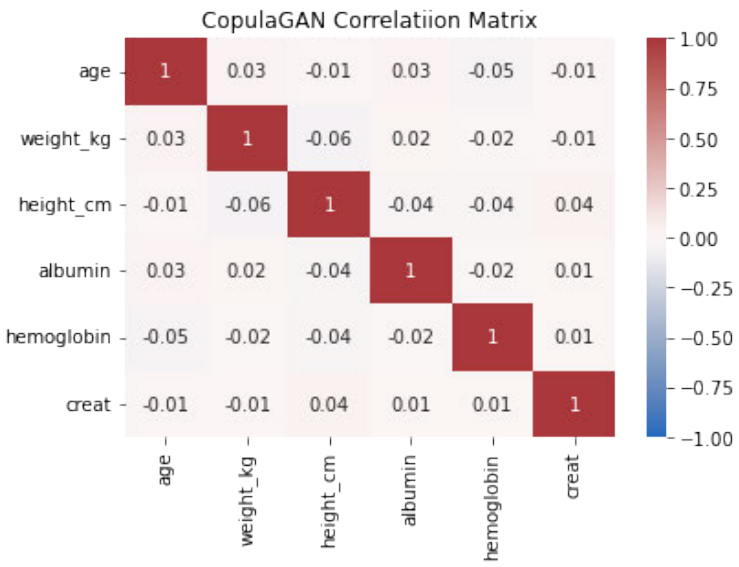


Figure 4.17: Correlation matrix of the data generated by the CopulaGAN model.

To determine meaning in the distances between the correlation of the original data and the better-performing datasets (GaussianCopula and TVAE), I calculated the squared differences. Because both of the GAN models produced data where hardly any correlation could be detected, I did not calculate the squared differences for the correlations of CTGAN and CopulaGAN. Between the GaussianCopula (figure 4.18) and the TVAE (figure 4.19) squared differences, the GaussianCopula produced more results closer to 0 which indicates that the difference between the correlations of the original and GaussianCopula datasets are more similar.

Samples of pairwise correlations of the different resulting synthetic datasets were mapped against the original data. I transform each synthetic dataset as well as the original into its correlation matrix. Doing so compares every possible pair of quantitative variables and summarizes the relationship as the correlation between each pair. The pairwise correlations in figure 4.20 show a familiar split in performance. The GaussianCopula pairwise correlation

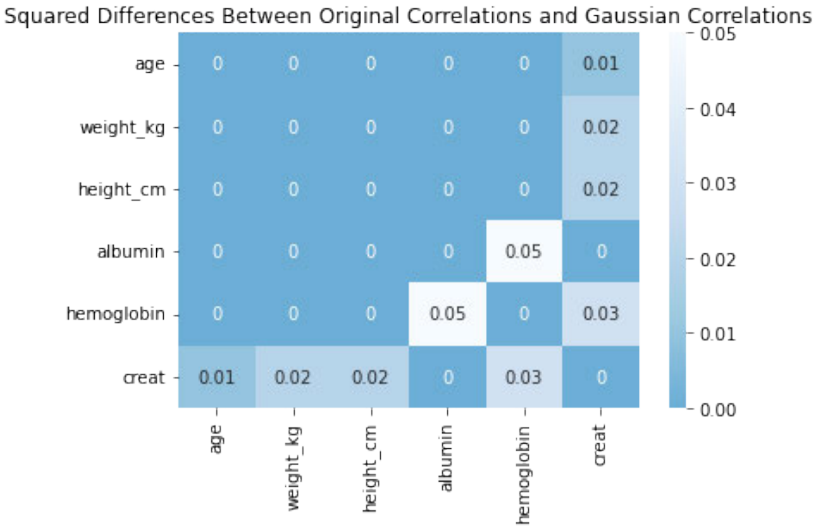


Figure 4.18: Squared differences between the original correlations and the GaussianCopula data correlations.

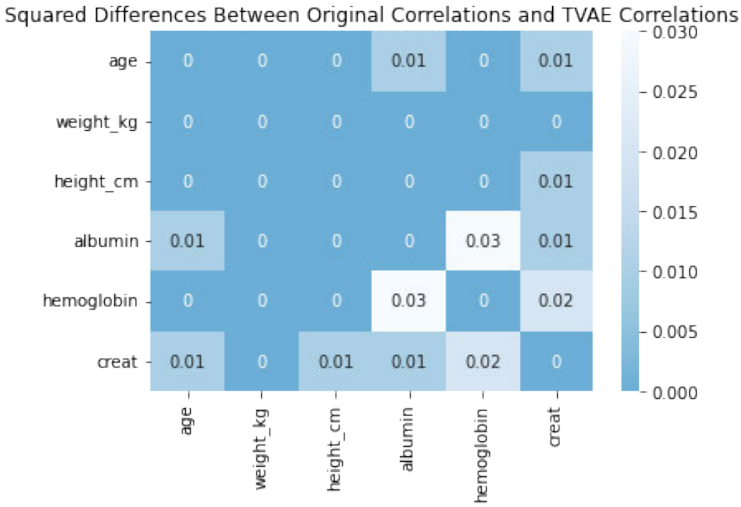
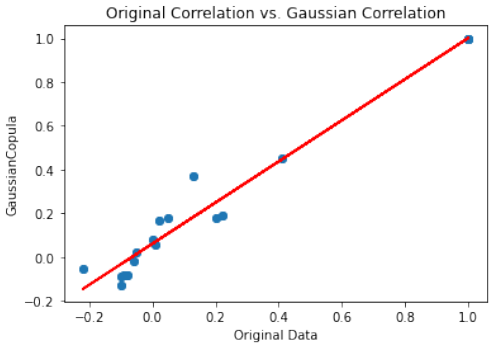
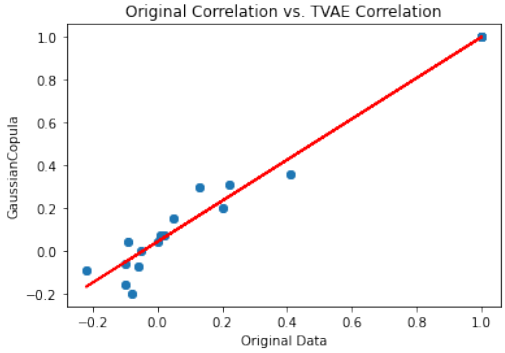


Figure 4.19: Squared differences between the original correlations and the TVAE data correlations.

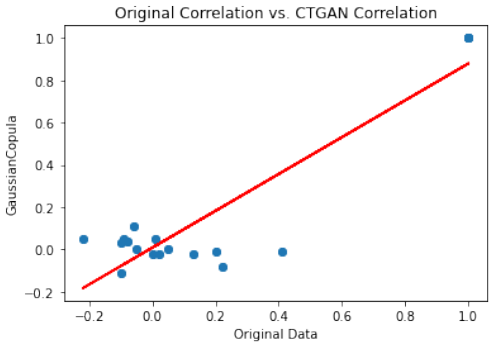
and the TVAE pairwise correlation both resemble one another while the CTGAN pairwise correlation and the CopulaGAN pairwise correlation resemble each other respectively.



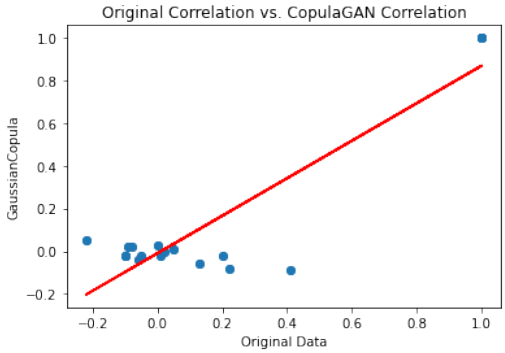
(a) GaussianCopula pairwise correlation



(b) TVAE pairwise correlation



(c) CTGAN pairwise correlation



(d) CopulaGAN pairwise correlation

Figure 4.20: Pairwise correlations of the synthetic datasets versus the original data correlations

I also used a chi-squared (CS) test to examine the association of the categorical features of the generated data to the categorical features of the original dataset. In addition to the CS test, I also calculated Cramer's V to measure the association between nominal variables (table 4.6). I was surprised at how high the CTGAN model's data scored (0.9423) in the chi-squared test compared to the rest of the datasets and especially compared to the CopulaGAN dataset's dismal score. Otherwise, the GaussianCopula model scored (0.8953) similarly to

Dataset	Chi-squared result	Cramer's V result
Gaussian	0.8953	0.0013
TVAE	0.8109	0.0013
CTGAN	0.9423	0.0014
CopulaGAN	0.6818	0.0012

Table 4.6: Categorical feature association scores compared to original dataset

the TVAE model (0.8109) which was to be expected after seeing so many other scoring similarities between the two models. The Cramer's V results followed similar suite to the chi-squared scores. Both the Gaussian and the TVAE Cramer's V scores resulted in 0.0013. CTGAN received the highest score at 0.0014 and CopulaGAN received the lowest at 0.0012.

After determining that the GaussianCopula model performs the best out of all four of the models, I started to experiment with the transforms that can be used with the model. By default, the GaussianCopula model uses one-hot encoding to interpret the variables of the columns. Instead of using one-hot encoding which would generate a column for each possible value within a feature, using a field encoder such as label encoding, for example, will replace values within a feature with a unique integer value. By changing how the features are interpreted, the fitting process of the model should improve [20].

The original dataset contains many columns in which a string value indicates "yes [condition]" and "no [condition]". In addition to using the field encoders provided by SDV, I also make a dataset where I manually change the string literals of "Yes/No" to boolean values of "True/False". I refer to the datasets generated with this data as boolean GaussianCopula models.

Because the field encoding applies to individual columns, on some models and resulting

datasets, I apply multiple encodings to columns in which they make sense to apply. For example, SDV's boolean encoding is applied to the "Yes/No" columns while SDV's categorical encoding is applied to the categorical columns within the original dataset.

GaussianCopula is the initial dataset generated from the original data without changing any parameters in the model function. No field encoding is applied.

Boolean GaussianCopula is the initial dataset with manual pre-processesing where the "Yes/No" features are changed to boolean object types before any encoding or modeling.

GaussianCopula with label encoding the initial GaussianCopula model with SDV's label encoding applied to categorical features.

GaussianCopula with categorical encoding is the initial GaussianCopula model with categorical encoding applied to categorical features.

Boolean GaussianCopula with categorical encoding this model uses categorical encoding on the pre-processed original data.

GaussianCopula with fuzzy categorical encoding in this model, categorical features receive fuzzy categorical encoding which adds gaussian noise.

Boolean GaussianCopula with fuzzy categorical encoding Boolean encoding is applied to the "Yes/No" features. The categorical features recieve fuzzy categorical encoding which adds gaussian noise.

Dataset	SDV KS-Test	SDV CS-Test	SDV Logistic Detection
GaussianCopula	0.88	0.89	0.22
Boolean GaussianCopula	0.87	0.19	0.0
GaussianCopula with label encoding	0.87	0.90	0.24
Boolean GaussianCopula with boolean encoding	0.87	0.2	0.0
GaussianCopula with categorical encoding	0.88	0.90	0.33
Boolean GaussianCopula with categorical encoding	0.87	0.19	0.00
GaussianCopula with fuzzy categorical encoding	0.87	0.82	0.34
Boolean GaussianCopula with fuzzy categorical encoding	0.87	0.23	0.0

Table 4.7: Results from applied various field encoders to the GaussianCopula model.

To begin with, the plain GaussianCopula data performed well with a KS-test score of 0.88, a CS-test score of 0.89, and a logistic detection score of 0.22 (see table 4.7). Applying either categorical encoding or label encoding improved the performance of the model. Applying categorical encoding to the GaussianCopula model did not change the KS-test score but improved the CS-test score by 0.10 points. It also improved the logistic detection score which rose from 0.22 to 0.33. Label encoding also slightly improved the CS-test and logistic detection scores but did not improve the KS-test score. Fuzzy categorical encoding slightly decreased the CS and KS-test scores but improved the logistic detection score by 0.12 points.

Pre-processing the data into boolean object types did not improve CS-test scores or logistic detection scores for the GaussianCopula model. It severely diminished KS-test scores were not affected by pre-processing the data. In fact, all resulting datasets that were generated from pre-processed data diminished greatly in their logistic detection scores when compared to the plain GuassianCopula model with default field encodings.

Chapter 5

Discussion

5.1 Discussion of Tool Research

Considerations for Choosing Readily Available Datasets

Although free-to-use datasets, such as the aforementioned Medicare Coverage Database, aren't classified as synthetic data in any means, I felt that they were important enough to include in the discussion of choosing a tool to generate, or rather to be provided with data for medical research. These datasets share similar features to generating one's own synthetic data as both are accessible and can tear down barriers to getting to real data. Both synthetic datasets and readily available datasets can achieve similar goals in facilitation the speed and development in research. Readily available datasets can also often be the foundation in which synthetic data is generated from until real data is available.

The GO FAIR Initiative is an organization that offers principles to evaluate data for how it can be found, accessibility, interoperability, and reusability (FAIR) [16]. The principles that

FAIR provides can be broken down based on further examinations of the dataset in question. These principles can be used as guidance for evaluating a potential research dataset.

The dataset should be easy to find. For example, the dataset could be reached via Google, in a repository, or perhaps through a publication citing the dataset without much trouble on the seeker's behalf. The dataset should additionally have a unique and persistent identifier such as a Digital Object Identifier (DOI) or a Uniform Resource Identifier (URI). The ease of finding and identifying a dataset should give a seeker confidence that the dataset found is indeed the one they originally set out to find. It is also important to consider how a dataset is versioned. As a dataset may be dynamic, versioning will give context into the snapshot of time in which the dataset is used for a particular bit of research. Metadata also plays a role in choosing a good dataset to work with. Examining the metadata should tell a researcher how the dataset was generated or collected. Additionally, it should be clear what kind and how many files are available within the offered data.

A benefit to using synthetic data is that it facilitates experiment repeatability. The same could be said for readily available datasets found online. The rights and restrictions of a dataset should be reviewed and finding how to cite the dataset will also aid in sharing research results for potential repetition.

Considerations for Choosing a Proprietary Synthetic Data Tool

I found that proprietary software platforms often do not disclose what methods are being used to generate synthetic data with the exception of MDClone. MDClone, however, does not go into detail on their customer website, but rather discusses the methods used to generate data in the Foraker [9] and Reiner Benaim [26] papers discussed in chapter 2.

With all proprietary software packages, an appointment or consultation must be set-up with a member of the sales team to explore whether the platform is a viable solution for the customer's (researcher's) needs. These tools are often integrated with bigger systems that one can assume a company or institution uses within their greater technology ecosystem. Though this kind of system likely exists for my institution, as an individual researcher hoping to fiddle with the tools, I wouldn't have been able to gain access and invite these companies into the system.

There are several considerations that came from the research into synthetic data generation tools. The following considerations will help guide researchers choosing a proprietary software for their synthetic data generation needs.

To begin, it is essential to consider the environment in which the researcher will be conducting their research in. Is the synthetic data generation tool in question browser-based or does it run on a standalone client on the researcher's computer? Furthermore, is the software machine agnostic – will it run regardless of the operating system that the researcher is using? In order to share research and enhance portability of models and data, it would also be important to determine if the software can run on multiple machines or operating systems.

Most proprietary solutions are going to require payment and licensing and thus it would make sense to consider the pricing plans and the licensing arrangements when. This is also a good time to consider how data and model ownership will be outlined with engaging a software company. What data does the company have privy to via the terms and conditions? Where is generated data stored? Where are the original data or data seeds stored? Who has access to generated data and data seeds?

Publication both in the past and the future also play an important role. Finding out what papers have been published, if any, using the tool in question should be paramount to considering the software. Who published those papers? Who funded the research? The standards for giving credit and citing the tool when used for research or product development should also be considered. Does the company of the proprietary tool receive compensation of some kind for innovation derived from generated data?

Data focused considerations include how much data is needed to get started with generating using the tool and whether or not a pre-existing schema needs to be uploaded. Consider how the underlying algorithms of the tool works as well. If a tool is built around GAN-based engineering, perhaps it would be better suited for image data but not for tabular data. Does it detect the feature type from a schema or bit of example data? Does it randomly generate data or does it generate data based on pre-existing relationships between features? Also consider whether the tool will work for the data that the research uses. For example, tabular data may be no issue to generate, but perhaps time series data is a weakness of the software.

It is crucial to understand how the developers were able to validate the efficacy of the synthetic data that their tool produces. This should also be transparently documented.

Finally, consider whether this data generation tool has been used for medical research or can be recommended as a viable tool for medical research. Some companies will only recommend that their software be used for model building, testing, or demonstrations. As previously mentioned, finding publications in which the tool in question was used would be particularly helpful.

5.2 Discussion of Researcher Interviews

Awareness of Synthetic Data

Synthetic elements of a study are commonplace. For example, there are already a lot of studies that routinely take place in a simulated environment or laboratory. These controlled studies often have very little noise and do not capture the essence of real life. Sometimes they might not even reflect real-world observations. Settings and conditions such as these provide researchers with "lab-grown data". As long as the purpose of the experiment is clear, this kind of data can suffice for proof-of-concept work.

Out of all the participants, only two were familiar with deep learning synthetic data generation techniques. One of these individuals is a data scientist who was working with the open source synthetic patient generation tool, Synthea. The other individual is a clinical researcher who has had experience using TensorFlow to create synthetic imaging data. This researcher also noted that augmenting and greeting synthetic image data is commonplace in medical imaging research.

Availability and Abundance of Data

Every researcher (clinical researchers and resident data scientists alike) all encountered times in their research where they did not have enough data. Some had to reframe their research question and goals to accommodate the lack of data. Some had to drop the study entirely.

Sometimes it isn't the nature of the data that influences its scarcity. The lack of available "manpower" to record the data after designing the study contributes to the scarcity. Having

an underpowered studies can hinder receiving funding from grants. For example, the National Institutes of Health (NIH) will not fund studies where too few patients have been enrolled.

A handful of researchers noted that they can sometimes find some data for their studies with established clinical databases. Most of the researchers that were interviewed were all familiar with the MIMIC III database and several other clinical databases that are often freely available online. Though MIMIC may be accessible online, other databases are not so readily available. For example, the National EMS Information System (NEMSIS) sends researchers a sample of data via a physical flash drive USB device to prove that the data can be opened on the researcher's computer before sending more. Other researchers default to a quick Google search to determine if the data they need for their study exists. Other databases can only be accessed if a researcher knows someone or is mentored by someone who can sponsor their access.

There can be some "cultural" rules surrounding use of external databases. Data might be copyrighted or protected as property by the stewards. If a researcher is allowed to use a group's data, the group is often included as a collaborator as "payment" for using the data. Additionally, navigating these databases is not a consistent experience. Data often isn't documented formally and the best guides are stumbled upon by finding forums online.

The amount of available data is often considered before starting a study – notably in retrospective analyses.If there isn't enough data, the research question guiding the study is refined in hopes of landing on a question that can be pursued with existing data. In prospective studies, it is typical to under-enroll patients. This leads to a dearth of freshly collected data for a study. Oftentimes a researcher may start with a retrospective study and then write a grant to obtain funding. Finally after securing funding, a prospective study can

be initiated. It isn't uncommon to drop a study entirely after refining the research question several times and still not having enough data to carry-out the study.

Readily available data is not always ideal for a researcher's specific study. Collecting data oneself is the most assured method of getting the exact type of data needed for the research question. This however is not always possible due to various reasons such as lack of funding, nature of the data (i.e. emergency room, ambulatory), existence outside of the researcher's system, etc. If the study is particularly specific, data that would be valuable for the actual study and data that is adjacent to what is needed for the study is sometimes grouped together. For example, in a study for Chronic Obstructive Pulmonary Disorder (COPD) ventilator data where there is little COPD data, Acute Respiratory Distress Syndrome (ARDS) may be used to supplement the COPD dataset. Grouping datasets like this doesn't actually mean the datasets that are being grouped are necessarily similar. Grouping the data like this ultimately loses precision.

For researchers that work with ambulatory and emergency room data, gathering data can be a challenging task. Unlike a clinical study, emergency room patients cannot be "recruited" or volunteer to be subjects. Because of this, the data accumulates slowly over time. Additionally, ambulatory and emergency room data is observational data. Once that information is changed, it is no longer observational and becomes retrospective. Physiological data captured in the emergency room or within an ambulance cannot be retrospectively "recaptured".

If there isn't enough data for a study collected from ambulatory and emergency room patients, researchers must rely on wider confidence intervals. Wider confidence intervals are also used when precise estimation is not possible or when measurements are physically

collected incorrectly often due to human error in intense situations.

Sometimes studies can be discouraged due to lack of data. Only recently have ambulances started collecting the type of data specifically used in certain studies such as in suspected stroke patients. Groups of patients are broad and can now start to be identified within the ambulance thanks to the new technology and standards for paramedics to collect data on the go. However, this technology is still not widely available and sometimes doesn't become available until the patient makes it from the ambulance into the emergency room.

Challenges of Available Data

Even when real data is available, it is not without challenges and issues. Oftentimes, data for small populations (populations with a high sampling rate, but a low number of patients) is too heterogeneous to be useful in any meaningful way to the researchers.

Gaps in data and missing data are not uncommon either. Human error is a monumental issue with clinical data collection. This can be a common case when clinical practitioners are collecting data during an appointment with a patient. Discrete, quantifiable records in a flow sheet are easier to capture accurately than text notes or other categorical data. One researcher estimated that capturing race ethnicity in electronic medical health records is expected to be 30% inaccurate. Race and ethnicity information is important to capture for social equity and inclusion. The malpractice of incorrectly or neglecting to capture this data was observed during the COVID-19 pandemic and obstructed efforts towards improving equity of care.

The format for accessing data across campuses in the University of California system is also varied. There is no uniform format for sharing or accessing medical data. Gaining access

to other campuses' healthcare data is labor intensive and there is no guarantee that another campus will even have enough data for the interested party. For example, one researcher lamented on how out of five UC schools, the initial 60,000 retrieved data points decreased to 600 confirmed case samples. After further filtering, the researcher was left with only 400 valid data samples. This researcher was not able to find data outside of the UC system that fit their needs. To complicate matters further, data wasn't always digital. Without a streamlined method of transferring data, often this researcher had to receive printed-out health records and had to manually enter the data into a digital format.

Validity of Synthetic Data

When asked about what would make synthetic data "good enough" for research, the researchers shared their thoughts. Data that creates new knowledge about the question that is being asked is good enough for one researcher. Sensitivity analysis performed to account for uncertainties in synthetic data would be an agreeable method to determine the quality of synthetic data as input.

One researcher suggested that to compensate for synthetic data not being a true clinical entity, cohort separate consolidation and cohort comparison to true patient data should be put to the test with a high bar to meet. To see how well one predicts the other could potentially give a sense of how valid the synthetic data is.

However, if synthetic data is generated using a base dataset, then the validity of the synthetic dataset might match that of the original data. In order to know when a true dataset has been properly collected without bias and is being used in a publication, the reader must assume that the researcher had no malicious intention. Sometimes with animal studies, the

data might seem "too good". In human studies, good faith is assumed if the research has been properly reviewed and published. If a single study supports a novel conclusion, fellow researchers do want to see these types of studies replicated nonetheless.

One researcher felt strongly that synthetic data can be validated at different levels. For example, a synthetic dataset that is going to be used in published research should be validated at a higher level than a synthetic dataset used to prototype a model in development. This would allow for speed of research and proper validation by the scientific community.

Places for Synthetic Data Use

The researchers were asked about where they could see synthetic data being helpful to generate or augment data in the clinical realm. They also had input into the specifics of where they could see or would want to see synthetic data fitting in.

Data from multiple sources – whether it is synthetic or real – could increase reproducibility of studies. One researcher recalled the Framingham study out of Massachusetts and how a demographic from one geographical region might produce different results when applied to a demographic hailing from a different region.

Bias might also be partially alleviated by using synthetic data to supplement true data. Randomization doesn't get rid of biases but instead equally distributes biases across the groups being compared to each other. If synthetic data could account for the biases that are not yet known, this would be helpful. The challenge would be to anticipate what is unknown.

Synthetic data may be good for hypothesis generation for future studies. It could also give an idea of the variance or power within a large trial. Using synthetic data for proof-of-

concept work, developing methodological approaches to analysis, and training and student learning were popular scenarios that the researchers could see synthetic data being used in.

Acquiring a larger quantity of data or acquiring more unique data is more involved when asking for permissions. One researcher lamented that there is almost a discouragement from speed in acquiring needed data. The "hoops" that a researcher needs to jump through to acquire more data is almost as detrimental as not having enough data to begin with. Sometimes it takes one and a half years to convince stakeholders and committees that more or diverse data is needed. Sometimes this is just enough time for a research partner to lose vested interest in the project. Occasionally there is no alternative if the resource that you have permission to use isn't sufficient in data quantity and quality.

Most of the researchers mentioned that synthetic data would have been of great use during the COVID-19 pandemic. Besides its uses for rapid prototyping, real data collection was often put on pause during the surges of the pandemic. Synthetic data could be a solution to fill in the blanks left behind when the care and treatment became focused on treating COVID-19.

One researcher described a need for data that you can "play with" that isn't the production data set. This researcher notes the importance of being able to embed and tes technologies in siloed environments. Another researcher found that using synthetic data was especially helpful in stress-testing systems and models. Because copious amounts of data can be generated in the blink of an eye, high throughput can be simulated easily. The patient population that is being tested can also be rapidly changed to represent different demographics or health statuses.

Being able to change patient population data to represent different patient demographics

or types serves a dual purpose since algorithms need to be built to be tuned to specific populations. It is the case that generalized algorithms don't work well for predicting conditions such as clinical deterioration or sepsis. Each population subset contained within the general patient population can be "incredibly unique".

Just a step out of the clinical realm, synthetic data has been helpful for co-validation when medical researchers are partnering with start-up companies. Having synthetic data available helps maintain a controlled environment.

Sample size is important for building models and testing them with enough volume of data. Often researchers will use a sample size calculation tool to determine how many patients may be needed to reach a certain N-value. These calculations are also referred to as "power calculations". In addition to calculating, clinical expertise for estimating proper N-values based on previous studies is used to determine the number of patients needed.

The distribution of unknown data is presumably unknown until it is actually collected and analyzed. When power calculations are performed, the researcher must compute them assuming several characteristics that the potential data will have such as it being normally distributed when this is not always the case. Retrospective studies can help with the assumptions but do not provide a guarantee.

In addition to distribution and characteristics of the uncollected data, the effect that will be measured from the data needs to be assumed in order to perform the power calculations as well. A huge effect size will require a smaller N-value and thus fewer patients. Likewise, a slight effect size will require a larger sample of patients. It seems like it would be difficult to predict what has yet to be gathered.

Using larger synthetic sample sizes wouldn't benefit prospective studies other than use

as a placer until more true data can be collected. However, with synthetic data to augment sample datasets, predictive scoring can be implemented on retrospective studies with large gaps in the data. If the validation cohort is a success in these types of studies, then the use of synthetic data to supplement the true data is a success as well.

Concerns and Questions about Synthetic Data

Clinical researchers wanted to know more about how synthetic data is generated and used in research. Questions they had about synthetic data included the following:

How is synthetic data protected from bias? One doctor wanted to know if the biases within a generated synthetic dataset are able to be known and accounted for. If so, they would be most agreeable to using synthetic data in their research. Synthetic data generated from data-driven methods will retain biases found in the original dataset given. To solve the problem of bias in synthetic data, the problem of bias in the real data would need to be solved first. As they say frequently in computer science and research disciplines in reference to the quality of data given: "garbage in, garbage out."

How similar is synthetic data generation to general data imputation? Another doctor wanted to know how similar synthetic data generation is to general data imputation. Imputation is a form of data-driven, as opposed to process-driven, synthetic data generation. The presence of this question might suggest that researchers are currently using synthetic data generation methods and not knowing it by name or how the underlying algorithms operate.

Much of the researchers' concerns with synthetic data match their concerns with real data. If real data can be biased, then synthetic data could be biased or objectively bad as well. If real data can't be used in evaluating how good a classification algorithm is performing, there is a concern that synthetic data wouldn't be a viable remedy. The outcome of the study becomes the most significant concern.

There were also doubts about whether synthetic data can be used for diagnostic research to see how common a certain disease is. The concern would be that the synthetic data wouldn't reflect the real world and have a different measure of prevalence. Specifically, sensitivity analysis ranges from a 5% to 30% prevalence.

More broadly, researchers are concerned about the quality of the data that is produced. Standards for data assessment would be ideal to have for synthetic data to ensure reporting accuracy. One researcher suggested using a system similar to the Equator Network which provides reporting guidelines for medical studies. The goal for these guidelines is to communicate the biases found in data for greater understanding. A standard like this would have to be globally implemented to be most effective. Additionally, to be published in a high-end medical journal, the data used in a study must meet the Equator Network guidelines.

Some of the researchers were very interested in gaining a deeper understanding of how synthetic data is generated. They believed that learning about how GANs, neural networks, deep learning, and artificial intelligence works would earn more trust in the medical community.

The interviewees gave mixed answers when asked about publishing research that used synthetic data. Some interviewees expressed that research that uses synthetic data should be publishable while others were more hesitant. All interviewees agreed that if synthetic data is

used in a study, the fact should be made known in the subsequent publication and reiterated when discussing interpretations of the data. Additionally, instructions for replication of the study using different synthetic or true data must be included in the publication as well.

With some familiarity of synthetic data generation, one researcher ensuring that the techniques for generating synthetic data can handle and correctly acknowledge the statistical properties of non-normal distributed data or skewed data was a concern for some of the researchers. This is a common pattern found in clinical data for features such as length of stay. These types of features are often hand-manipulated to reflect accuracy. Synthetic data generative tools would have to be aware that not all features can be generated following a normal distribution.

All researchers that were interviewed agreed that there needs to be more collaboration between clinical researchers and data scientists. The concern that people who do not have clinical perspectives or expertise may be performing research on publicly available health data. One interviewee offered ideas for solutions to the lack of collaboration including talking a common language and developing Data Provision Units where subject researchers can learn to ask and narrow down precise questions in order to get the data they need.

Another researcher was concerned that synthetic data may not be "good enough" solely because the base real dataset in which the synthetic data would be generated from may have missing values or issues from human error input.

Suggestions for Synthetic Data Use

Some researchers shared how having a sandbox environment would be helpful to play and test synthetic data.

One doctor from the interview was particularly concerned about why synthetic data isn't a subject that is being regularly covered within the UC Davis's own Master of Clinical Research program.

One of the interviewees expressed how it is "unreasonable" to expect clinicians to understand the statistics behind validation or artificial intelligence behind algorithms. They noted that people who do have that expertise can provide that security.

Final thoughts shared in the interview

The researchers all believed that sharing data is the future for advances in healthcare. One particular researcher shared how they believe the culture at UC Davis might actually be conducive to synthetic data use. They believe that as an academic community, we are "hyper-focused" on innovation. The UC Davis health system is already imbued with the culture of a technology start-up. We as a university have access to the "coolest technologies" and we just have to "connect the dots".

Interview Discussion

I had expected all the clinical researchers that were interviewed to be much more familiar with synthetic data by the name of "synthetic data" rather than "imputed values". I suspect that even though imputation and synthetic data generation are essentially the same concept, there is a larger discrepancy other than naming conventions that set them apart from each other in how clinical researchers understand and use them. This could be because of the associated methods for generating or imputing data. It seems that "imputation" implies

statistical modeling whereas “generating synthetic data” is grouped in with a schema related to modern artificial intelligence practices such as machine learning and deep learning.

Definitions

Institutional Review Board (IRB) An administrative body established to protect the rights and welfare of human research subjects recruited for research.

N-value Generally the usable sample size.

Prospective Study A study type in which data is collected as the study progresses. Data definitions are clearly identified and fields with gaps are filled in for all patients in the study. These studies cost a lot of time and money (Example: clinical trial).

Retrospective Study A study type in which the data has already been collected (Example: Electronic Medical Health Record).

Sandbox environment An isolated, virtual code testing space separate from the production environment. New features in a system can be tested here without worrying about affecting the production environment.

Sensitivity analysis Studies how uncertainty in the output of a model can be determined by the source of the input. Recalculates outcomes using alternative assumptions to test the robustness of a model. Helps to understand the relation between the input and output of a model.

TensorFlow An open source platform for machine learning. Contains tools, libraries, and community resources.

5.3 Discussion of Use Case Results

The answer to whether the data generated in the use cases was good enough for clinical research "depends". Expertise is recommended to check if the data makes sense. With synthetic data tools that are easy for clinical experts to use, the data generated can be reviewed as soon as it is appears on the screen. With the PRF use case, generating data was the easy part. Evaluating the data was much more difficult. For the liver oncology use case, generating the data presented more challenges than evaluating the balance of the mix.

Liver Oncology Use Case Discussion

Going into this thesis research, I thought that balancing data would easy enough. In practice, it is much more difficult than I anticipated. There are many vectors in which the "balance" can be disturbed. Using a synthetic data generation library like SDV creatively offers novel solutions in how we generate synthetic data. Because SDV was so flexible, and generally quick enough to run on my old and dubious laptop, I was able to achieve a "balanced" dataset that both filled the requirements for reducing demographic imbalance and broadening the patient outcome results. When efficacy is less important than composition, SDV's flexibility really shines.

I recommend that a researcher keep the limitations of synthetic data generation in mind when using a tool such as SDV in this way. However, by no means should those limitations discourage creative applications and iterations. With a good plan to judge the efficacy of the resulting data and a clear idea of what is important or what the purpose of the resulting data should serve, using SDV is conclusively a viable and valuable resource for researchers.

PRF Use Case Discussion

As explored in the tangible results from chapter 4, the GaussianCopula model frequently produced the best (most similar) results in comparison to the original dataset. Following close behind was the TVAE model. The two GAN models (CTGAN and CopulaGAN) often performed poorly which was to be expected since GANs are generally better-suited for tasks such as imitating image data. However, when looking at the chi-squared results from the initial four generated datasets (table 4.6), it is remarkable how high the CTGAN model scored. It is almost like the CTGAN model might be better suited to generate synthetic data for datasets that have mostly categorical features. Conversely, the other GAN model, CopulaGAN, performed extremely poorly in comparison.

I encourage further research into the performance of the CTGAN model from the SDV library for generating synthetic datasets with many categorical features. If the CTGAN model is truly better at generating this kind of data, I would then recommend that a researcher tries using the CTGAN model instead of the GaussianCopula model for generating synthetic datasets. Otherwise, if there are more quantitative features in the original dataset, I would recommend that the researcher use the GaussianCopula model to generate their synthetic data.

I am not sure of how SDV differentiates between categorical encoding and label encoding. I assume that is it similar to what Scikit-learn can do with ColumnTransformer /citescikit-learn. Even so, ColumnTransformer in Scikit-learn uses one-hot encoding which is what SDV uses by default. However, the SDV default field encoder is not the categorical encode. It is by default set to `one_hot_encoding` [20].

Despite the lack of clarity in SDV's documentation, specifying the field encodings for the model to use was simple. In addition to field encodings, SDV also allows the user to specify distributions to set for any particular column. I did not experiment successfully with this feature solely due to my lack of subject matter expertise in what different distributions should typically look like in medical records.

Chapter 6

Conclusion

To summarize, I explored the current literature on what exists for synthetic data and patient population generation technology for clinical research. I familiarized myself with the methods and theory behind generating synthetic data as well. Because of the background research that I performed I was able to look at some of the open source projects that anyone could get started with. I also explored a few proprietary solutions for synthetic healthcare data generation. I extensively studied how to evaluate synthetic data.

After laying down the initial groundwork, I interviewed medical researchers at UC Davis to find out the extent of their experiences and opinions about using synthetic data in their research. Drawing from these interviews, I chose a synthetic data generation library to test out with two different use cases. After generating and evaluating several rounds of synthetic data with my use case teams, I have a better understanding on how to use and evaluate the data and the tool used to create the data. Going forward, I am now equipped to write a toolkit for the UC Davis DataLab so that future researchers can use synthetic data in their

project if they would like.

Using the Synthetic Data Vault library (SDV) was very accessible and easy to set-up and use. I definitely conclude that it would be a good tool to write about in a toolkit offered by DataLab. When it is used to generate balanced data, it needs to be carefully curated in such a way that the data created is maintaining all components needed to keep the context of the original data. This can be done by iterating upon the generated data and experimenting with whether the GaussianCopula model or the TVAE model work best for the particular use case. For use cases that involve careful numbers and extremely high scores in efficacy, viable data generated by SDV may be trickier to achieve. It really all comes down to what will be "good enough" for the particular use case.

6.1 Implications

My research aimed to answer the question of how we can serve clinical researchers at UC Davis by enabling them to use synthetic data generation tools. Researchers are already accustomed to using synthetic data, but know it under a different name such as "imputed" or "simulated" data. From my small interview sample, most researchers were interested in using it but unsure how to get started. They generally seemed to agree that synthetic data could be used in publishing to medical journals as long as it is clearly communicated that the data used in a study is synthetic. By commencing with the creation of a synthetic data toolkit for the UC Davis DataLab, researchers can learn how to set-up and produce their own synthetic data and go forth to publish their research.

6.2 Limitations

As an individual researcher, I was limited by the synthetic generation tools that were available to me. It would be interesting and valuable research to see a neutral institution pit some of the proprietary tools against each other in a battle for the best synthetic data results. For example, does Accelario's Synthetic Control Arm generate better synthetic data than MDClone can? Additionally, there is no shortage of commercial solutions for synthetic data generation. I have listed out a small handful of other synthetic data solutions that could be researched further in Appendix D.

6.3 Questions and Suggestions for Further Research

While creating my literature review, I kept accumulating more and more questions for further research. Some of them are closely related to the literature while others are tangential to what has been discussed. I want to know how exactly does synthetic data help fight data bias in medical research? What would the creation of universally accepted standards for synthetic data generation and population representation look like? What are the arguments against using synthetic data in healthcare research? Can artificial "sensitive" data be used as a honeypot in secure systems? How is research funding, if at all, affected by using synthetic data?

As I concluded my interviews, I found myself wondering about further research and questions that I would like to see addressed in the future:

- How does the level of prestige of a scientific or medical journal affect whether research

is published or not?

- Would reviewers for a medical journal be biased about use of synthetic data in studies under review?
- Is information about synthetic data being taught in academic programs such as UC Davis's Master in Clinical Research?
- How common is data imputation? Are researchers aware that data imputation is considered synthetic data?
- What makes some data registries more user-friendly than others?

Bibliography

[1] Ahmed M. Alaa et al. "How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models". en. In: *arXiv:2102.08921 [cs, stat]* (Feb. 2021). arXiv: 2102.08921. URL: http://arxiv.org/abs/2102.08921 (visited on 01/28/2022).

[2] Nick Anderson. Personal communication. May 11, 2022.

[3] Parker Bannister. *Synthetic Data Is Enabling Better Healthcare Tools - Here's How.* Feb. 2022. URL: https://www.particlehealth.com/blog/synthetic-data-healthcare-tools.

[4] Jaume Canet and Lluís Gallart. "Postoperative respiratory failure: pathogenesis, prediction, and prevention". In: *Current Opinion in Critical Care* 20.1 (2014). ISSN: 1070-5295. URL: https://journals.lww.com/co-criticalcare/Fulltext/2014/02000/Postoperative_respiratory_failure__pathogenesis,.10.aspx.

[5] Edward Choi et al. "Generating Multi-label Discrete Electronic Health Records using Generative Adversarial Networks". In: *CoRR* abs/1703.06490 (2017). arXiv: 1703.06490. URL: http://arxiv.org/abs/1703.06490.

[6] UC Davis DataLab. *DataLab at UC Davis*. 2022. URL: https://datalab.ucdavis.edu/.

[7] Committee on Economics. *ASA Physical Status Classification System*. Dec. 2020. URL: https://www.asahq.org/standards-and-guidelines/asa-physical-status-classification-system.

[8] Randi E. Foraker, Douglas L. Mann, and Philip R.O. Payne. "Are Synthetic Data Derivatives the Future of Translational Medicine?" In: *JACC: Basic to Translational Science* 3.5 (2018), pp. 716–718. ISSN: 2452-302X. DOI: https://doi.org/10.1016/j.jacbts.2018.08.007. URL: https://www.sciencedirect.com/science/article/pii/S2452302X18302262.

[9] Randi E. Foraker et al. "Spot the difference: comparing results of analyses from real patient data and synthetic derivatives". In: *JAMIA Open* 3.4 (Dec. 2020), pp. 557–566. ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooaa060. eprint: https://academic.oup.com/jamiaopen/article-pdf/3/4/557/36625809/ooaa060.pdf. URL: https://doi.org/10.1093/jamiaopen/ooaa060.

[10] Andre Goncalves et al. "Generation and evaluation of synthetic patient data". In: *BMC Medical Research Methodology* 20.1 (May 2020), p. 108. ISSN: 1471-2288. DOI: 10.1186/s12874-020-00977-1. URL: https://doi.org/10.1186/s12874-020-00977-1.

[11] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[12] UC Davis Health. *Health Data Resources: Tools*. 2022. URL: https://health.ucdavis.edu/data/tools.html.

[13] Jordan Hoffmann et al. "Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets". In: *Science Advances* 5.4 (2019), eaau6792. DOI: 10.1126/sciadv.aau6792. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aau6792. URL: https://www.science.org/doi/abs/10.1126/sciadv.aau6792.

[14] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[15] Plotly Technologies Inc. *Collaborative data science*. 2015. URL: https://plot.ly.

[16] GO FAIR Initiative. *The Fairest of Them All: Using data principles to evaluate open data repositories*. 2021. URL: https://www.go-fair.org/.

[17] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. "Measuring the quality of Synthetic data for use in competitions". In: *CoRR* abs/1806.11345 (2018). arXiv: 1806.11345. URL: http://arxiv.org/abs/1806.11345.

[18] Uri Kartoun. "Advancing informatics with electronic medical records bots (EMR-Bots)". In: *Software Impacts* 2 (2019), p. 100006. ISSN: 2665-9638. DOI: https://doi.org/10.1016/j.simpa.2019.100006. URL: https://www.sciencedirect.com/science/article/pii/S2665963819300065.

[19] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90.

[20] MIT Data To AI Lab. Mar. 2022. URL: https://sdv.dev/SDV/index.html.

[21] Christopher Lucas. Personal communication. Apr. 21, 2021.

[22] *MCD downloads*. URL: https://www.cms.gov/medicare-coverage-database/downloads/downloads.aspx.

[23] *Medicare Claims Synthetic Public Use Files (SynPUFs)*. URL: https://www.cms.gov/research-statistics-data-and-systems/downloadable-public-use-files/synpufs.

[24] Richard E Murray, Patrick B Ryan, and Stephanie J Reisinger. "Design and validation of a data simulation model for longitudinal healthcare data". In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2011 (2011), pp. 1176–1185. ISSN: 1942-597X. URL: https://europepmc.org/articles/PMC3243118.

[25] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[26] Anat Reiner Benaim et al. "Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies". eng. In: *JMIR medical informatics* 8.2 (Feb. 2020). Publisher: JMIR Publications, e16492–e16492. ISSN: 2291-9694. DOI: 10.2196/16492. URL: https://pubmed.ncbi.nlm.nih.gov/32130148.

[27] Mihaela van der Schaar and Nick Maxfield. *Synthetic data: breaking the data logjam in machine learning for healthcare*. Sept. 2020. URL: https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/.

[28] Rittika Shamsuddin, Barbara Maweu, and Balakrishnan Prabhakaran. "Virtual Patient Model: An Approach for Generating Synthetic Healthcare Time Series Data". In: May 2018. DOI: 10.1109/ICHI.2018.00031.

[29] Bill Siwicki. *Is synthetic data the key to healthcare clinical and business intelligence?* Feb. 2020. URL: https://www.healthcareitnews.com/news/synthetic-data-key-healthcare-clinical-and-business-intelligence.

[30] Samarth Swarup and Madhav V. Marathe. *Generating Synthetic Populations for Social Modeling: Second Tutorial at the Sixteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* Tech. rep. NDSSL-TR-2017-17-073. Biocomplexity Institute of Virginia Tech, May 2017.

[31] Dassault Systemes. *Adapt, Innovate, and Scale for the Future with a Synthetic Control Arm.* URL: https://www.medidata.com/en/synthetic-control-arm-infographic.

[32] Dassault Systemes. *Papers and Publications.* URL: https://www.medidata.com/en/life-science-resources/medidata-institute/publications.

[33] The pandas development team. *pandas-dev/pandas: Pandas.* Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[34] Allan Tucker et al. "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software". en. In: *npj Digital Medicine* 3.1 (Dec. 2020), p. 147. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00353-9. URL: http://www.nature.com/articles/s41746-020-00353-9 (visited on 01/28/2022).

[35] Jason Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: *Jour-*

nal of the American Medical Informatics Association 25.3 (Aug. 2017), pp. 230–238. DOI: 10.1093/jamia/ocx079. URL: https://doi.org/10.1093%2Fjamia%2Focx079.

[36] Alex Watson. *Deep dive on generating synthetic data for Healthcare*. Aug. 2020. URL: https://gretel.ai/blog/deep-dive-on-generating-synthetic-data-for-healthcare.

[37] Lei Xu et al. "Modeling Tabular data using Conditional GAN". In: *CoRR* abs/1907.00503 (2019). arXiv: 1907.00503. URL: http://arxiv.org/abs/1907.00503.

Appendix A

Abbreviations

CTSC Clinical and Translational Science Center (at UC Davis)

EMHR Electronic Medical Health Record

EMR Electronic Medical Record

HIPAA The Health Insurance Portability and Accountability Act of 1996

IPF Iterative Proportional Fitting

IPU Iterative Proportional Updating

OSIM Observational Medical Dataset Simulator

OMOP Observational Medical Outcomes Partnership

SMOTE Synthetic Minority Oversampling Technique

GAN Generative Adversarial (Neural) Network

DNN Deep Neural Network

IM Independent Marginals

CDP Conditional probability tables

PCD Pairwise correlation difference

FHIR Fast Healthcare Interoperability Resources Specification

VAE Variational autoencoder