SYNTHETIC DATA SIMULATION FOR PRIVACY-PRESERVING MEDICAL DATA SHARING

By

Ziqi Zhang

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

June 30, 2022

Nashville, Tennessee

Approved:

Bradley A. Malin, Ph.D.

Ipek Oguz, Ph.D.

Jimeng Sun, Ph.D.

Khaled EI Emam, Ph.D.

Zhijun Yin, Ph.D.

Copyright © 2022 Ziqi Zhang
All Rights Reserved

ACKNOWLEDGMENTS

First and foremost, I would like to express my heartfelt gratitude to Dr. Bradley Malin, my Ph.D. advisor, who has provided me with immeasurable help over the past five years. When I first started my Ph.D. program, I was a complete novice in doing research. Dr. Malin assisted me in building everything I rely upon to finish my degree and become the scholar I am today. He is a constant reminder of the exceptional scientist I aspire to become, and a lighthouse guiding me through academia. Furthermore, Dr. Malin provided me with so much encouragement and support in the difficult moments I encountered over the past years, which was exactly what kept me from giving up and inspired me to keep moving forward.

I would also like to give sincere thanks to my committee members, Dr. Ipek Oguz, Dr. Jimeng Sun, Dr. Khaled EI Emam, and Dr. Zhijun Yin, for their encouragement and constructive advice on this dissertation. Further, I want to thank my colleagues at the Health Information Privacy Laboratory at Vanderbilt University, especially Chao, who has been a constant source of support for me since I first joined HIPLAB.

Next, I want to thank my parents, who helped me make the decision to pursue a Ph.D. degree and supported me throughout the process of earning it. Without a doubt, they are the people who love and care about me the most. I am eternally grateful to my parents for everything they have done for me over the last 27 years.

Further, I want to thank my beloved cat, DuDu, for being the most adorable thing in my life. From Beijing to Nashville, she is always there for me and can always soothe my mind and elicit the best part of me. Without her, I wouldn't be able to get out of rock bottom and bring myself together in those tough days.

In addition, I also want to thank Vanderbilt University and the city of Nashville for giving me such a pleasant environment in which to study, research, and reside. Even though it was not a very long time, I left behind so many happy memories that I believe would bring me strength throughout my whole life.

My thankfulness also extends to my friends, who shared cherished experiences with me. Here I would like to acknowledge several important pieces of help I received. First, I want to thank Maolin, Zichen, Kan, and Yuzhi for taking care of DuDu. Second, I want to thank Yubo, Ziran, Zhongwei, Peter and all my friends who helped me arrange things in Nashville. Thirdly, I want to thank Yuqi and Huiwen for their encouragement when I was considering my future path. At last, I want to especially thank my fellows at SDSZ for being a source of happiness.

This acknowledgment is easier to compose than a scientific paper because the points I wish to make are so straight and clear. However, this brief line cannot adequately express all my gratitude to the adorable people I met in the past five years and those experiences from which I learned so much about gain and loss, give and take, as well as joy and sorrow. I hope I will always have the luxury of gratitude in the coming days.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Motivation

Over the past several decades, we have experienced substantial advances in digital and information technologies, notably in their applications to health and health care. The confluence of modern large-scale databases, statistical learning methods, and computational capabilities, has empowered medical record data collection, storage, and analysis at an unprecedented scale. As such, it is increasingly possible to develop data-driven methods to generate insight into the practice of healthcare. Along with precision medicine initiatives [1], the research and analytics built upon the access to enormous amounts of medical data is expected to have a near-term impact on the development and refinement of healthcare delivery and clinical decision support, as well as lay the groundwork for a long-term transformation of healthcare in terms of efficiency and effectiveness.

This opportunity, however, is frequently put in jeopardy by privacy concerns. In 2021, DeepMind, an Alphabet subsidiary, was accused of violating patient privacy in processing medical data from the UK's National Health Service[1]. The concerns of specific stakeholders over privacy violations in data sharing may counteract their belief in the benefits of sharing medical data to enhance the healthcare outcomes. Currently, a multi-party conflict between the patient, heath care executive, researcher, and other stakeholders widely pervades the data sharing process[2] – for example, patients expect a commitment to the privacy and security of their data through more robust regulatory protection and oversight; healthcare executives hesitate to share data for being held accountability of a potential privacy breach; and data users, who are well positioned to reuse medical data for a wide variety of secondary uses, are limited in their ability to make progress to restrictive policies regarding

[1] https://www.cnbc.com/2021/10/01/google-deepmind-face-lawsuit-over-data-deal-with-britains-nhs.html

[2] https://nam.edu/health-data-sharing-special-publication/

data access.

Concerns about privacy intrusions in the data sharing process are not unfounded. In the United States, the Health Insurance Portability and Accountability Act (HIPAA)[3] mandates de-identification of individual-level medical data intended for public distribution. The de-identification could use either the Expert Determination method or the Safe Harbor method (i.e., the removal of 18 specific types of identifiers). However, stripping away identifiers is not always sufficient to protect an individual's privacy. Back in 2006, researchers found that the supposedly anonymized data, if characterized by high dimensionality and sparsity (which is similar to the case of medical data), can be re-identified at a high probability by leveraging background knowledge from an auxiliary outside source [2]. To prevent re-identification, the research community proposed statistical disclosure limitation models, such as k-anonymity [3], l-diversity [4], t-closeness [5], m-invariance [6]. However, these models are often achieved by making edits to the data directly (e.g., through attribute generalization), which induces an inherent tradeoff between the data utility and privacy. As a result, to avoid a significant loss of data utility, theoretically guaranteed privacy protection against re-identification is frequently impossible to achieve.

Over the past decade, the notion of sharing synthetic data [7] as an alternative to real data has gained traction in the medical informatics community. To provide context, in the setting of creating synthetic data, a real dataset is a collection of observations that represent the real-world data distribution over a predefined feature space, and the synthetic dataset is a sample from the simulation of this distribution. Intuitively, if the simulation is sufficiently accurate, a resulting synthetic dataset can be viewed as a statistical equivalence to the real dataset, with similar utility. And, from a privacy standpoint, given that no synthetic data belong to a real person, there is naturally little risk of directly associating the data with an individual's identity, which indicates that using synthetic data for sharing purposes could mitigate privacy concerns.

[3] https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

Synthetic data is not new and methods for its creation and dissemination have a long history. Back in 1993, Rubin [8] has introduced the concept of releasing synthetic micro-data for public use, as well as a multiple imputation approach to create synthetic datasets by replacing actual values of all of the variables of real individual with predicted values. However, the effort to translate methods into practice not been limited until recent times, particularly due to advances in machine learning theory and computational power. Specifically, researchers lacked effective tools for modeling high-dimensional data with complex cross-feature correlations. Furthermore, previous attempts at data simulation heavily relied on strong assumptions about the data distribution (e.g., knowledge abstracted by domain experts), making it difficult to generalize from specific cases. More recently, the machine learning community has focused on the development of a branch of advanced generative models based on deep neural networks. With deep learning, medical record simulation is formulated as a data modeling problem, which, at its core, is about representation learning – whereby the data distribution is modeled in a latent representation space through a set of parameterized mapping functions. As neural networks grow in their depth and number of parameters, their ability to govern complicated distributions improves with high statistical generalizability, scalability, and little reliance upon knowledge drawn from domain experts. As such, new research avenues and opportunities for synthetic data simulation are opened.

1.2 Problems and Research Goals

This dissertation focuses on creating a synthetic data simulation pipeline to enable privacy-preserving medical data sharing, with an emphasis on electronic medical records (EMRs). This goal is decomposed into three sub-problems: 1) developing a statistical learning framework for capturing the real data distribution and sampling synthetic data from the learned distribution, 2) evaluating how synthetic data adapt to utilities, and 3) analyzing the underlying risk of privacy disclosure associated with sharing synthetic data. A high-level abstraction of the problems and research goals is shown in Figure 1.1. The sub-problems

are detailed in the following subsections.

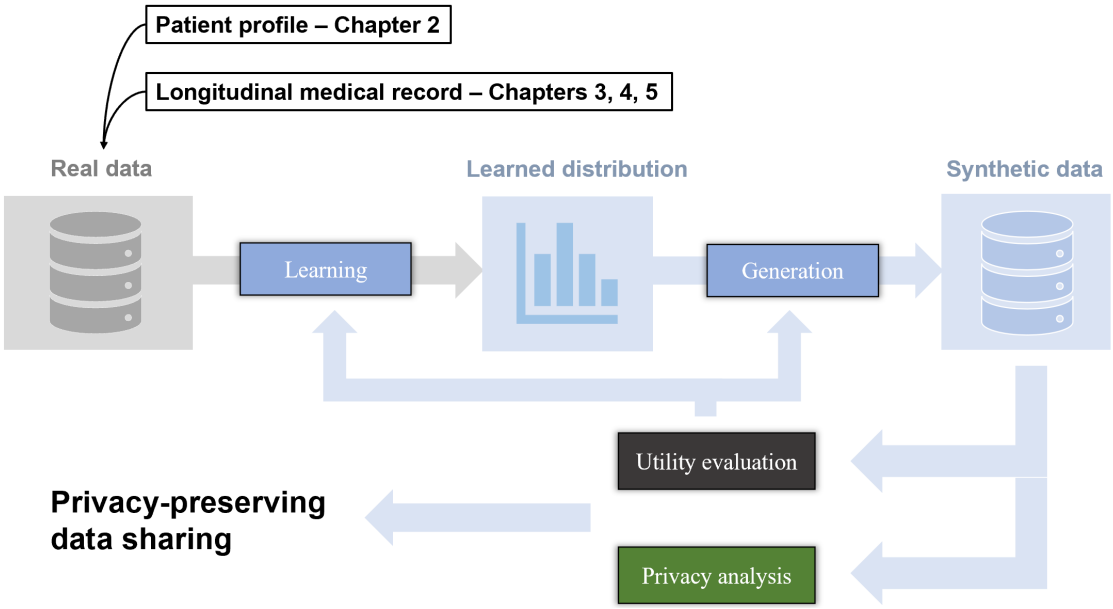


Figure 1.1: An abstraction of the problems and research goals

1.2.1 Electronic medical records Simulation

To begin, I provide a brief overview about how EMRs are represented in this dissertation. The EMR of a patient is represented as a series of episodes (e.g., outpatient visits or hospital stays), each of which is associated with a collection of clinical events (e.g., diagnoses made or procedures performed during the same episode). Clinical events are specified using standardized coding systems that maintain a fix-sized vocabulary (e.g., International Classification of Diseases billing codes). Figure 1.2 provides an architectural depiction of medical data types and structures extracted from the Observational Medical Outcomes Partnership (OMOP) representation that we rely upon to form the patient representation. Doing synthetic data generation at scale for all features simultaneously remains an open problem. This dissertation focuses on a subset of the OMOP representation (e.g., diagnosis, procedure) as a proof of concept, and leaves other types of data (e.g., natural language text) for future work.

The objective of simulation is to create a synthetic version of the real dataset for reuse.

This objective necessitates that the synthetic data have similar utility as their real-world originals. To accomplish this, I consider the utility of synthetic data from two perspectives: completeness and quality. Completeness refers to the amount to which synthetic data represents the attribute space of the real data, which dictates the type of hypothesis or analytics that may be developed on top of the synthetic data. Quality, by contrast, describes how well the synthetic data approximate the real data in a specific attribute space, which determines the fidelity of the hypothesis or analytics derived from the synthetic data. Intuitively, higher data completeness is likely realized at the expense of a lower level of data quality. For example, correlations between attributes in high-dimensional data are difficult to capture completely. In comparison, synthetic records with compressed clinical information are more likely to be constructed with a minor statistical bias toward the real data. As such, to maximize the utility of synthetic data, simulation research regarding various levels of completeness is required.

Building on the aforementioned patient representation, this dissertation explores two distinct types of simulation objectives. The first is referred to as a patient profile, which is a projection of the record into the event coding space that neglects the record's intrinsic temporal structure. The second is referred to as a longitudinal record, which retains the record's original temporal structure. Figure 1.3 provides an abstracted example of the patient profile and longitudinal record.

1.2.2 Utility Evaluation

An answer to the question "Is the synthetic data as reliable as the real data, regardless of any specific usage?" is essential for anyone who intends to share or use the synthetic dataset. Yet, obtaining such an answer is non-trivial. Human experts have intuition and specialized knowledge in determining a synthetic record's authenticity. However, human evaluation is frequently prohibitively expensive, hard to acquire, and subject to within-expert and between-expert variability [9], and thus, requires supplementary evaluation methods for

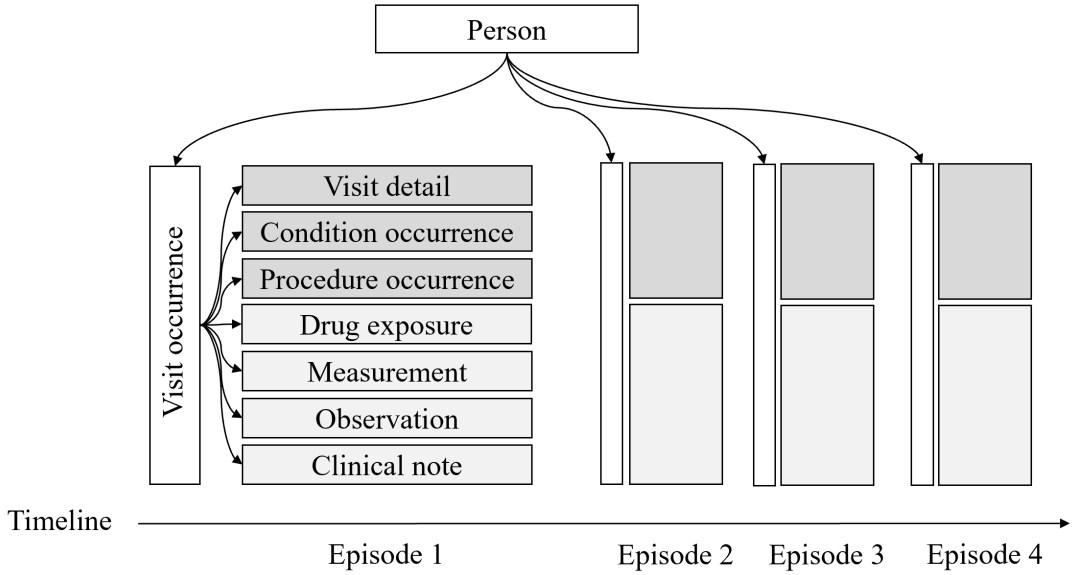


Figure 1.2: An example of the longitudinal medical data structure for a patient with 4 episodes. Each box represents a concept domain for OMOP representation. The darker gray indicates domains that we rely upon in the analysis in this dissertation.

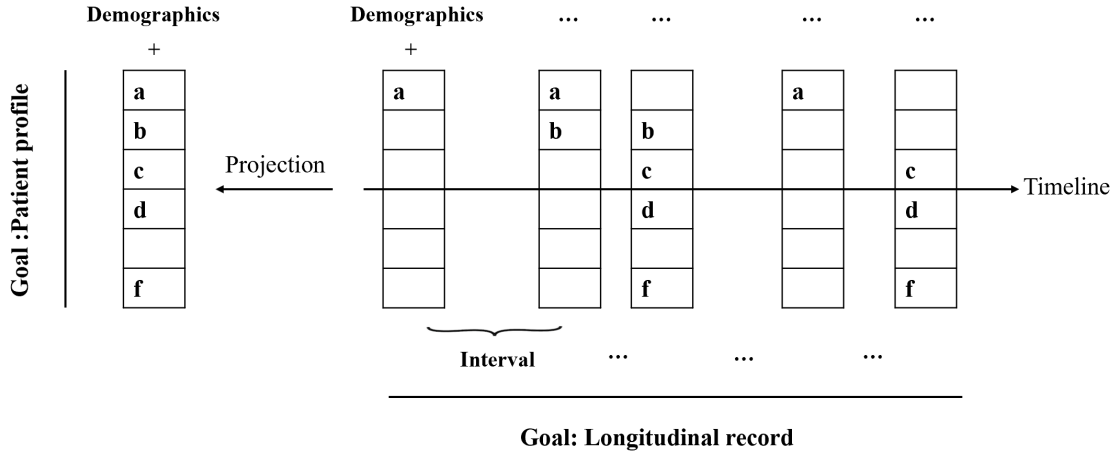


Figure 1.3: An abstracted example of a medical record with 5 episodes. Each letter represents a unique clinical concept.

large-scale studies.

An auxiliary type of method to rely upon for evaluation is machine learning-based analytics. However, this type of method needs careful consideration because it is prone to several shortcomings. First, a measurement may not be comparable across semantically

dissimilar datasets. This is an obstacle to establishing a universal gold standard for the quality of synthetic data. The absence of such a standard may complicate the process of developing policies for synthetic data. Second, it may be challenging to make intuitive judgement about synthetic data's quality in terms of a measurement. Third, a slight divergence of a measurement may not always indicate a high utility for a particular application. As such, a reliable evaluation typically requires the use of a comprehensive and diverse set of methods in conjunction with one another. To achieve this goal, this dissertation investigates application-agnostic evaluation criteria over first-order statistics, structural properties in either original or latent space, and distinguishability of real and synthetic data.

1.2.3 Privacy Risks Analysis

It has been claimed that the use of such approaches poses little risk to the privacy of the individuals whose data are used to compose the models for data synthesis [10, 11]. Such claims are founded on the expectation that synthetic data does not retain an explicit one-to-one match with real individuals (which enables a linkage attack, by which the data are linked to individuals' identity with the assistance of public sources, such as a publicly accessible voter registration database [2, 12]). However, there is evidence that the models involved in data synthesis can leak information associated with the training samples [13, 14, 15, 16, 17], much like how certain types of machine learning models are known to do. For instance, generative models for sequential data can suffer from unintended memorization [18], whereby the synthesized features mimic, or are highly similar to, a specific training sample. As a consequence, an adversary can mount a membership inference attack [19, 20, 21], whereby they infer if targets known to the adversary were used in the synthetic data generation process. Membership inference is a privacy violation in its own right as the targets do not necessarily disclose that they visited a particular healthcare organization or participated in a biomedical research study. Moreover, when membership inference occurs, further compromises, in the form of attribute inference [22, 23], may arise. For instance, the

information associated with these targets reported in the synthetic data, but not known to the adversary a priori, could be revealed as well. The following example further illustrates how it raises privacy risks. Imagine that a malicious attacker Mallory gained access to a patient Bob's medical record history (e.g., via a data broker, self-disclosure by the patient themselves, or a breach of a data warehouse). At some later point in time, Bob received diagnosis **x** (e.g., HIV-positive) and was treated at a healthcare facility, which Bob intends to keep confidential. Then, a researcher at the facility makes public a synthetic cohort of individuals with diagnosis **x** based on its set patient records. Now, imagine that Mallory applies a membership inference strategy to learn that Bob's record was included in the records relied upon to generate the synthetic cohort. At this point, Mallory learns Bob was diagnosed with **x**, which further compromises Bob's privacy.

Although it has been shown that the application of statistical perturbation, such as the mechanisms inherent in differential privacy (DP) [24, 25], may assist in the reduction of such risks; generally, they are not preferred from the perspective of the data user. That is because, for generative models, DP can lead to a significant reduction in the utility of the resulting data [26, 27], rendering the synthetic data relatively inadequate for their intended purposes.

Given such problems, it is in the best interest of a data holder to consider the risk that a privacy attack will be successful. And, based on the analysis, they can then decide if it is appropriate to share the synthesized data or if additional protections (either technical mechanisms, such as DP, or data use agreements) are warranted. This dissertation formulates the membership inference and attribute inference problems from the perspective of the data holder, who aims to perform a disclosure risk assessment prior to sharing any medical data.

1.3 Dissertation Overview

The remainder of this dissertation details our effort in designing synthetic EMR simulation pipelines. My research encompasses all three facets of the learning and generation method,

utility evaluation, and privacy analysis for both patient profiles and longitudinal medical records. The following is the structure of this dissertation.

In Chapter 2, I describe a simulation pipeline for patient profiles. In Chapter 3, I introduce a framework for modeling the temporal characteristics of EMR data and a corresponding simulation pipeline for longitudinal medical records. Both Chapters 2 and 3 cover methods for assessing the utility and privacy of synthetic data. In Chapter 4, I discuss how to improve the pipeline demonstrated in Chapter 3, considering the model training and data generation strategies, as well as the utility evaluation method. In Chapter 5, I introduce a method for longitudinal medical record simulation that is distinct from the one described in Chapters 3 and 4, and I evaluate this method from both the utility and privacy standpoints. Chapter 6 concludes this dissertation with my contributions, the limitations of the current work, and the potential research directions for future work. Parts of Chapters 2, 3, and 5 have been published as peer-reviewed journal articles [28, 29, 30]. Part of Chapter 4 was under review as a peer-reviewed journal article as the time this dissertation was submitted.

CHAPTER 2

Patient Profile Simulation

2.1 Introduction

The research community has attempted to simulate synthetic EMR data through models based on clinical knowledge published in the literature and the data documented in real electronic medical records [31, 32]. Certain contributions can be categorized into knowledge-oriented EMR synthesis, where knowledge is extracted either from real EMR data or external data [33, 34]. For instance, Buczak et al. generate synthetic EMRs of tularemia patients by mining real EMR records to obtain patients' care patterns, frequencies of billing codes and syndromes [35]. These approaches are appropriately designed based on the extracted knowledge and can account for static, as well as temporal aspects of a patient's status and the evolution of disease. However, there are several limitations that are common to these approaches: 1) the knowledge merged into the generation process is often incomplete (or biased), 2) the generation mechanisms are specific to a particular phenotype or process, which lacks generalization ability, and 3) sharing patient-level synthetic data may be vulnerable to another privacy intrusion, such as membership attack. In this situation, the data recipient is able to correctly predict if a real record is part of the training dataset that led to synthetic records. This attack may leak information about features (e.g., diagnoses) of a real patient.

More recently, the machine learning community has focused on another category, which we refer to as data-driven EMR synthesis, through the research into advanced generative models that automatically extract the inherent knowledge within (or between) data in real records. Among various techniques, generative adversarial networks (GANs) [36] have shown a remarkable ability to generate synthetic data with a realistic feel, while simultaneously protecting privacy. This is because the artificial nature of the data has the potential

to mitigate the concerns of re-identification. GANs have potential to be resistant to the attribute inference and membership inference attacks [37]. In general, GANs are notable in that they are designed to address an adversarial environment in which a generator is forced to produce increasingly realistic instances, such that an evolving discriminator, cannot distinguish them from real data. To date, the applications of GANs have been successful in the domains of imaging [38, 39, 40], natural language text [41, 42, 43] and audio generation [44, 45].

Over the past several years, GANs have been customized to generate structured and categorical EMR data (e.g., sets of billing codes) [37, 46]. GANs in this domain adopt the following pipeline. Initially, the system selects a training dataset of EMRs that satisfies the definition of a target population (e.g., type 2 diabetics). Next, the system optimizes for an objective function based on the distance between the distributions of synthetic and real data. Finally, the system evaluates the GAN with respect to data utility and privacy risks. The evaluation of GANs is usually accomplished by characterizing both the distributional similarity of features and predictive similarity on a simple task between the real and synthetic data.

However, this approach to EMR simulation has several drawbacks. First, the current GAN model may not be sufficiently efficient in capturing the data distribution, thus, induce a barrier in the learning task. Second, current measures of data utility fail to characterize if the generated data retain key structural properties of real data in the original and the latent space. Third, relying solely on the EMRs of a population of interest as the training data may cause a loss of certain statistical properties of the real data.

Given the limitations of the current simulation pipeline, we hypothesized that the utility of the data could be enhanced, without scarifying privacy, through a refinement of the learning process and models. Specifically, we aimed to enhance the learning model of GANs through introducing advanced optimization objective, incorporating additional utility measures of key structural properties, and refining the filtering strategy for selecting

training data. This chapter demonstrates the plausibility of this hypothesis by applying the new GAN pipeline with approximately one million real EMRs from Vanderbilt University Medical Center (VUMC).

2.2 Related Work - GANs in the Medical Domain

In the medical domain, several variations of GANs have been developed to generate realistic EMRs of diagnosis and procedure codes. These include medGAN [37], medBGAN and medWGAN [46]. These GANs have several commonalities and their architecture is shown in Figure 1a. First, they are all based on a framework that combines the GAN architecture with an autoencoder (which projects the original data into a low dimension space and then reconstructs them), as shown in Figure 1a. The autoencoder is incorporated to address the limitation that the original GAN cannot generate discrete outputs. This is achieved by concatenating the generator with the pretrained decoder, which is fine-tuned during the training process. Another common characteristic is that they all apply batch normalization [47] and a shortcut technique [48] to the generator to accelerate learning. Yet, these GANs differ in their distance measures between the distributions of real and synthetic data.

medGAN applies Jensen-Shannon Divergence (JS Divergence), which makes it susceptible to mode collapse, where the generator learns to map different inputs to the same output, and mode drop, where the generator only captures certain regions of the underlying distribution of the real data [49]. To stabilize GAN training and solve the mode challenges, medBGAN and medWGAN adopt the distance measures introduced in Boundary-seeking GAN [50] and Wasserstein GAN [51, 52], respectively. The objective function of medBGAN pushes the generator to match the distribution of the real data by continuing to generate samples near the boundary of the discriminator in each optimization iteration. By contrast, medWGAN applies Wasserstein Divergence to formulate the objective function in a manner that the divergence between the distributions can be more accurately measured in a gradient-descent algorithm.

2.3 The EMR-WGAN Framework

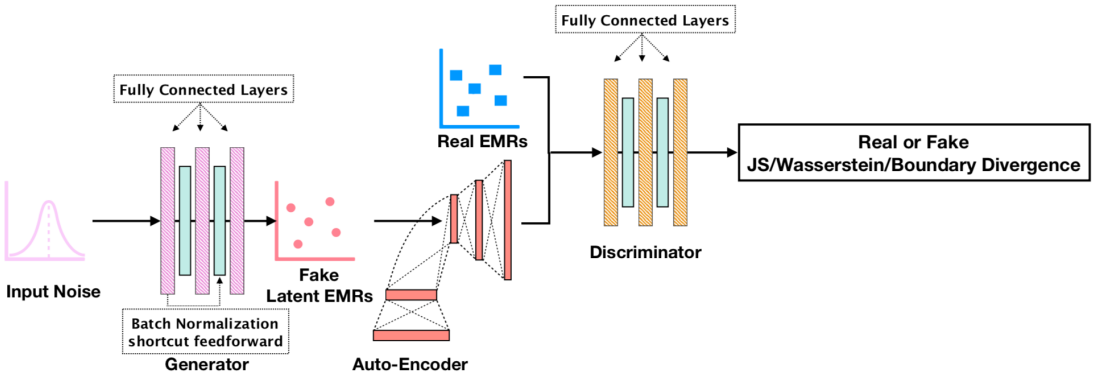
2.3.1 Model Architecture

EMR-WGAN, whose architecture is shown in Figure 2.1b, refines the learning model. It uses the basic structure of the medGAN with several modifications. First, due to the drawbacks of JS Divergence, EMR-WGAN, like medWGAN, adopts the Wasserstein Divergence and employs an corresponding implementation known as WGAN-GP [52]. Second, we remove the autoencoder because pre-training is no longer required to stabilize the training process when Wasserstein distance is used as the optimization objective and the WGAN-GP framework is capable of simulating discrete data. Third, to mitigate the effect of an exploding gradient, a phenomenon in which gradients accumulate large amounts of error (resulting in unstable training), and a vanishing gradient, a phenomenon in which the gradient of the loss function becomes zero (resulting in an inability to appropriately update the network), we additionally apply layer normalization [53] in the discriminator. Batch normalization is not applied because it would change the training objective from penalizing the norm of the discriminator's gradient with respect to each input independently to penalizing the gradient's norm of the entire batch (which is required in WGAN-GP to enforce the Lipschitz constraint to the discriminator). By contrast, layer normalization maintains its computation within each single input, which is suitable for the discriminator to mitigate the training obstacles incurred by unexpected gradient updates.

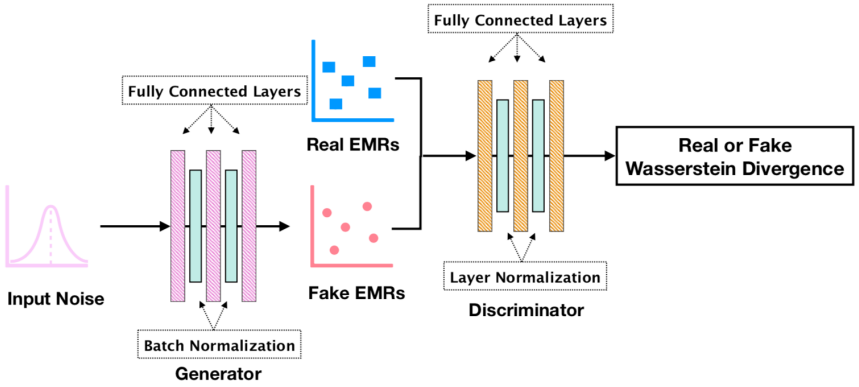
2.3.2 Training Strategy

For the purposes of generating EMR data with a specific concept, it is straightforward to train a GAN model on real records with the same concept. We refer to such a filtering strategy as simple training. However, simple training may cause a loss of certain statistical properties when the size of available real data is small.

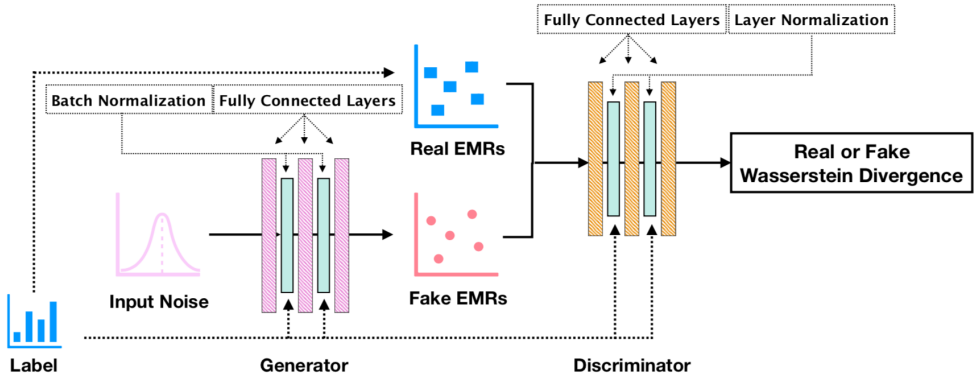
We introduce a conditional training strategy, where we use a conditional version of GANs over the EMR data with various concepts to generate synthetic records of a single



(a) medGAN, medWGAN and medBGAN



(b) EMR-WGAN



(c) EMR-CWGAN

Figure 2.1: Architecture of (a) previous and (b-c) proposed GAN models.

concept. For example, when the target concept is Male, age 18-44, then conditional training will addtionally involve real records of females and other age groups in the training process.

Conditional training needs to explicitly figure out the concept label of each record, thus

we build the conditional version of EMR-WGAN, EMR-CWGAN, whose architecture is shown in Figure 2.1c. We incorporate the concept labels of records as part of the generator and the discriminator. Specifically, the labels are denoted by a set of embeddings, which are integrated into the input of the generator and discriminator. After training EMR-CWGAN with different populations and their labels, one can apply the embedding associated with the label of a desired population to obtain synthetic records.

2.3.3 Evaluation measures

2.3.3.1 Utility Measures

Dimension-wise Prediction (DWP) [37]

This measure evaluates the degree to which a generative model captures the inter-dimensional relationships of real data, following a strategy summarized as "Train on Synthetic, Test on Real (TSTR)" [54]. Specifically, for each binary attribute, two classifiers are trained on real and synthetic data, where the binary status of the attribute serves as the dependent variable and all remaining attributes serve as the independent variables. Specifically, the real dataset is randomly partitioned into two sets: *Training* and *Testing* with ratio 4:1. The generative model is trained on *Training* and, subsequently, generates a *Synthetic* set with the same size. The two logistic regression classifiers for each attribute are trained on *Training* and *Synthetic*, respectively, and then both are evaluated with *Testing*. We then compare the F1 score of the classifiers. For offering a ceiling on the performance of the generative model, we partition *Training* into two equally-sized subsets, train a classifier for each and compare them with *Testing*.

Latent Space Representation (LSR)

This measure evaluates the ability of a generative model to capture the latent factorized representations of real data. It is natural to assume that each real record is generated from a distribution $p(x|w)$, where w represents data generative factors in the latent space $\mathbb{R}^d$ with

m independent and $d-m$ dependent dimensions. We utilize the β-Variational Autoencoder (β-VAE) [55] to discover, among $\mathbb{R}^d$, an efficient representation $w^{'} \in \mathbb{R}^m$ of real data. β-VAE rewrites the objective function of VAE [56] by inserting a weight β to the KL Divergence regularization:

$$\mathscr{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z,x)} \log p_\theta(x|z) - \beta \mathrm{KL}(\mathrm{q}_\phi(\mathrm{z}|\mathrm{x})||\mathrm{p}(\mathrm{z}))$$

where $z \in \mathbb{R}^d$ satisfies the standard Gaussian distribution. A larger β value encourages more dimensions $q_\phi(z|x)$ in to approach their corresponding dimensions in $p(z)$. In other words, the mean of the variance distribution in each of these dimensions is forced to approach 1. At the same time, the remaining dimensions (i.e., $w^{'}$) can be thought of as efficient latent dimensions to characterize, and then reconstruct, the input data. We interpret each of these dimensions as a latent mode. A useful generative model is expected to yield synthetic data with a variance distribution for each latent mode that is similar to real data.

We train a β-VAE model over the real dataset and retain the set of latent modes with a threshold for the mean of the variance distribution less than 0.5. We provide a synthetic dataset of the same size as the real dataset into this β-VAE model. In doing so, we record the variance distributions of the latent modes. We measure the distance between the mean of each mode's variance distribution and the mean of its counterpart in real data. A smaller distance indicates a greater similarity in synthetic and real data.

First-order Proximity (FOP)

This measure investigates whether synthetic data retains the local structure of real data. To do so, we build an undirected attribute graph from a population (with a format of adjacency matrix), where the weight of an edge between a pair of attributes corresponds to their co-occurrence frequency in the population (i.e., the count of patients have positive values for both binary attributes). FOP, denoted by edge weights, is widely adopted to characterize

adjacent structures of networks [57]. We measure the difference in adjacency matrices between the synthetic and real datasets.

Frequent Association Rules (FAR)

This measure investigates the extent to which the patient-level attribute associations in real data are maintained in synthetic data.

The two key criteria in association rule mining are *support* and *confidence*. Support is an indication of how frequently the condition set appears in the dataset, whereas confidence is an indication of how often a condition rule is true. With respect to rule mining in EMR, the support of condition set X (e.g., a set of diseases and procedures) T is defined as the proportion of records in T that contain X. The confidence of a condition rule, $X \Rightarrow Y$, with respect to T, is the proportion of records that contain X that also contain Y.

We first obtain all frequent condition sets, FCS for abbreviation, (forming a set $\mathscr{S}$) with frequency larger than a threshold min_s such that any subset of any FCS is not in $\mathscr{S}$. For each FCS $f \in \mathscr{S}$, we then determine the set of association rules $R : f' \Rightarrow f - f'$, where each rule satisfies that the number of records which have f' also have f is greater than a threshold min_c. By applying such a process to both real and synthetic EMR data, we measure the proportion of the association rules that are from the synthetic data that are in the real data and vice versa, which we refer to as recall and precision, respectively. We use the well-known association rule mining technique *Apriori* [58] to learn FCSs and the association rules from the real and synthetic EMRs. It is notable that FAR can be regarded as an expansion on the structural measure FOP. This is because FAR does not limit the number of features to consider and, thus, consider deeper and broader dependencies between features. By contrast, FOP focuses on the condition sets containing only two features.

2.3.3.2 Privacy Measures

Membership Inference

An attacker committing a membership attack could be motivated in numerous ways. Here, provide several as an illustration of the potential problems. First, the attacker may execute the attack to gain new knowledge about a known person. It is often the case that a training dataset is composed of a cohort based on some rigorously defined criteria (e.g., all patients have HIV or share a certain sexual orientation). In this case, if this knowledge is not known to the attacker a priori, then proving a targeted individual's membership would lead to a clear disclosure about the individual. Even if the attacker had some prior belief about the status of the targeted individual, proving their membership would provide absolute certainty, which would be a boost in their knowledge. Second, the attacker might not be interested in targeting the individuals in the dataset, but rather, discrediting the organization that shared the simulated data. Consider, it is likely that healthcare organizations will claim that such simulated data is de-identified. At the same time, they may promise the individuals to whom the real data corresponds that their inclusion in such a dataset will not be made evident. However, if the attacker can prove the presence of one or more targeted individuals, then they may claim that the organization is failing to adhere to its promises and might be in violation of federal regulation (particularly if they did not obtain consent from the patients prior to creating the synthesizer).

We assume that an attacker is in possession of the complete set of diagnosis of a set of real patients. The attacker will attempt to infer which patients are in the training dataset. We calculate the Hamming distance between each known and synthetic record. Given a distance threshold, the attacker claims that all records less distant to any real one than the threshold are the targeted real patient. We assess the precision and recall of this claim.

Attribute Inference

This attack is accomplished by inferring an unknown attribute value of a set of compromised patients via the generated data. Attribute inference may infringe upon a patient's privacy when an attacker gains knowledge that is only accessible in the training dataset.

We assume the attacker possesses a subset of attributes of some real records and attempts to infer the value of the missing attribute. This is accomplished by applying a k-nearest neighbors algorithm, where for each real record, the k nearest neighbors in synthetic data help decide the feature value of interest. We measure the F1 score of attribute inference as a function of k.

Reproduction Rate

The portion of reproduced records among synthetic records helps evaluate the risk of identity disclosure. Specifically, if the re-identification risk of the real data is high, a high reproduction rate would indicate the use of synthetic data cannot prevent identity disclosure. However, it should also be noted that a high reproduction rate does not necessarily equal a high privacy risk in case of the real data themselves are not subject to identity disclosure. In addition, reproduction rate also evaluates the ability of a generative model to create new instances rather than memorizing the training data.

2.4 Materials

The data in this study is derived from the VUMC Synthetic Derivative (SD), a de-identified warehouse of over 2.2 million EMRs. We extracted all ICD-9 diagnosis codes for each patient, which were rolled up to their subcategories by removing the portion of the codes to the right of the "." and retained the distinct set codes. This process led to 944 codes. We refer to this dataset as SD.

Table 2.1: Summary statistics of the EMR datasets.

Dataset	Patients	Gender	Age Distribution 0-17, 18-44, 45-64, >64
SD	2,246,444	M:47% F:53%	21%, 32%, 24%, 23%
CSD	1,045,634	M:47%: F:53%	17%, 29%, 26%, 28%
Dataset	**ICD-9 Codes**	**Codes Per Patient**	**Patients Per Codes**
SD	944	8.113	19,298
CSD	854	14.76	18,080

Summary statistics for this dataset, including age and gender, are provided in Table 2.1. Note that we discretized age into four groups 0-17, 18-44, 45-64 and >64 based on U.S. Census 2010 criteria [1] for presentation purposes (more fine-grained age groups could be applied). In doing so we treat the dataset categorically, facilitating the evaluation of training strategies.

It was observed that a portion of the records, as well as a subset of the billing codes, were not suitable for EMR synthesis. For example, EMRs with too few codes may not be informative during learning and, instead, may lead to biased (or even incorrect) models. The same is true for ICD-9 codes with very low prevalence. As such, we refined the data (details in Appendix A) to compose a cleaner dataset, which we refer to as CSD.

This dataset is composed of 854 billing codes. It has approximately half the patients in the SD dataset, but maintains roughly the same distribution of age and gender. The number of distinct codes per patient and the number of patients per code is approximately 15 and 18,000, respectively, compared with 8 and 19,300 for SD.

Each patient's record is represented as a binary vector over the codes, where a cell value is one if the corresponding code is in an EMR and zero otherwise.

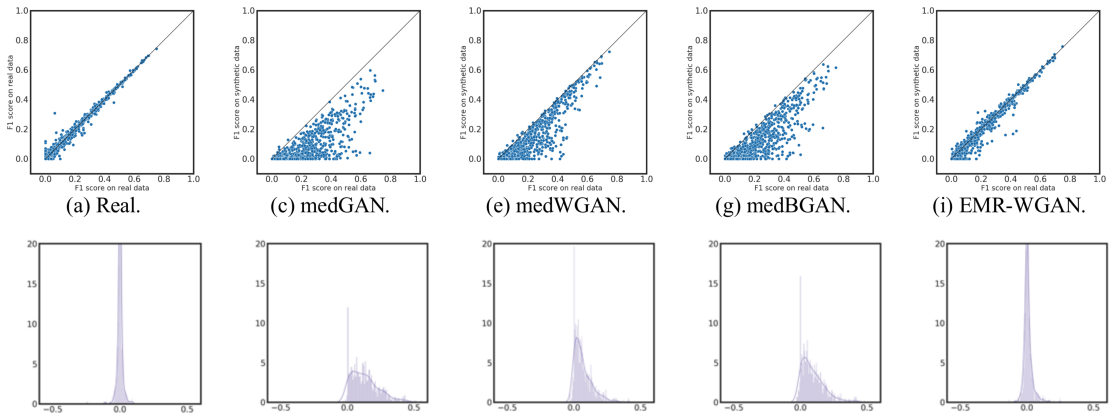
[1] U.S. Department of Commerce. *Age and Sex Composition: 2010.*

2.5 Experimental Results

2.5.1 Evaluating EMR-WGAN

Dimension-wise Prediction

The results for DWP are shown in Figure 2.2. There are several findings worth highlighting. In Figures 2.2a and 2.2b, it can be seen that the F1 scores for billing codes in the real *vs.* real setting are close to the diagonal without obvious bias. Additionally, the distribution of dot-to-diagonal distances is roughly symmetric, which indicates the stability of the inter-dimensional relationship in the original system. As depicted in Figures 2.2c and 2.2d, the distribution of dot-to-diagonal distances are heavily biased towards real data, which suggests that medGAN fails to capture the inter-dimensional relationship of real data. Third, Figures 2.2f and 2.2h show that medWGAN and medBGAN achieve similar performance, but are still biased in a manner similar to medGAN. Fourth, EMR-WGAN outperforms all alternatives demonstrate similar patterns as the real *vs.* real setting, as presented in Figures 2.2i and 2.2j. As such, it appears that EMR-WGAN is more apt at simulating the inter-dimensional relationships in real data.



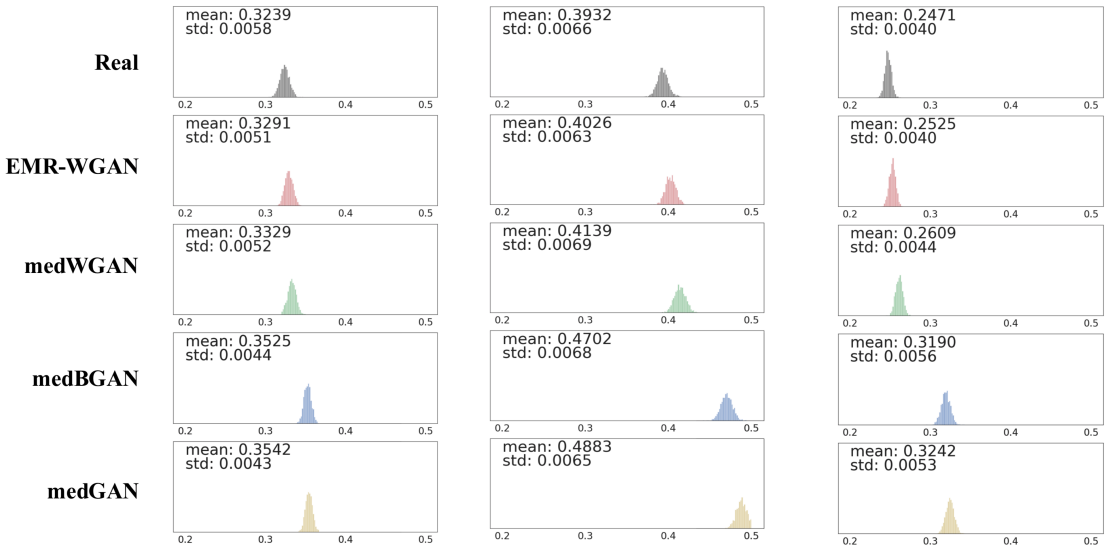
(a) Real. (c) medGAN. (e) medWGAN. (g) medBGAN. (i) EMR-WGAN.

(b) Distribution of dot-to-diagonal distance of (a). (d) Distribution of dot-to-diagonal distance of (c). (f) Distribution of dot-to-diagonal distance of (e). (h) Distribution of dot-to-diagonal distance of (g). (j) Distribution of dot-to-diagonal distance of (i).

Figure 2.2: Dimension-wise prediction. Subfigure a presents the F1 scores of logistic regression classifiers in real vs real setting. Subfigures c, e, g and i show the results of real vs synthetic setting of four GANs. Subfigures b, d, f, h and j demonstrate the distributions of perpendicular distances from dots to the diagonal line for a, c, e, g and i, respectively.

Latent Space Representation

Figure 2.3 shows the LSR results in all three latent modes. The generative models are sorted according to the mean of the variance distributions. EMR-WGAN achieves the smallest distance to real data. By contrast, there are relatively large gaps between the medBGAN and medGAN distributions and real data. To assess the reproducibility of this finding, we generated data 10 times for each generative model and confirmed that EMR-WGAN had a smaller mean than each of the alternative methods at a 0.01 significance level (via t-test). This result suggests that EMR-WGAN can better capture the latent structural properties of the data.



(a) Latent mode 1. (c) Latent mode 2. (c) Latent mode 3.

Figure 2.3: Latent space representation. Each subfigure illustrates the distribution of the variances in one latent dimension (with mean less than 0.5). The first row corresponds to real data. Each subsequent row corresponds to synthetic data generated by a particular type of GAN

First-order Proximity

Figure 2.4 shows the FOP graph distances between synthetic and real data. EMR-WGAN clearly achieves the smallest distance and outperforms all other approaches. medWGAN and medBGAN are less likely to capture the patterns of local structures in real data.

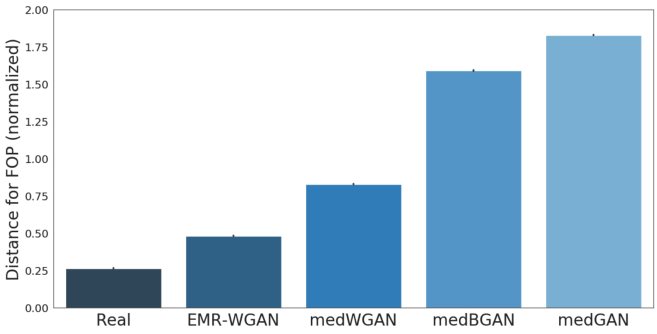


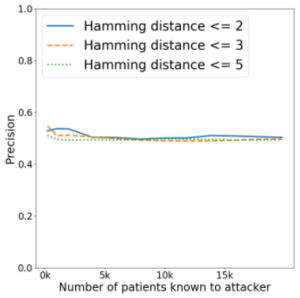
Figure 2.4: First-order proximity. The normalized graph distances between the billing code networks learned from real and synthetic data with respect to FOP. We compute the graph distances in four settings: real *vs.* real, medGAN *vs.* real, medWGAN *vs.* real, medBGAN *vs.* real and EMR-WGAN *vs.* real. We sort the generative models according to the normalized distance values

Membership Inference

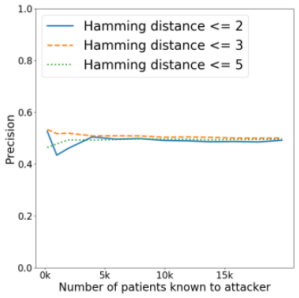
For each subpopulation, We randomly sample w records from *Training* and another w records from *Testing*. We view the mixture of the $2w$ records as the fully compromised dataset by an attacker. By varying the number records known to the attacker, we present the precision and recall of membership inference of medBGAN, medWGAN and EMR-WGAN in Figure 2.5 and Figure 2.6, respectively. We use three different thresholds of distance in the form of a Hamming distance of 2, 3 and 5. There are several observations. First, for each subpopulation, the precision for all GANs are similar (around 0.5). This implies that an inference is no better than random. Second, when the number of known real records is small, the precision is unstable for all GANs. This is particularly the case when the Hamming distance less than 2. Third, we cannot observe increase on the recall for membership attacks against EMR-WGAN in any distance threshold in comparison to medBGAN and medWGAN. And, fourth, when the Hamming distance threshold is 2, only about 10% of the known records to an attacker can be found in the training set. However, there is no obvious indication of which records would be found, which suggests that the rate of success of such an attack would be on the order of 0.1, which is in line with privacy

risk thresholds that have been put into practice. Based on this evidence, we believe that EMR-WGAN induces no greater privacy risks with respect to membership inference than the state-of-the-art approaches. However, we would like to clarify that the result given by the state-of-the-art inference methods might not necessarily indicate the upper bound of the privacy risk. As such, we do not assert that EMR-WGAN, as well as other GAN models are not subject to any membership inference risk.

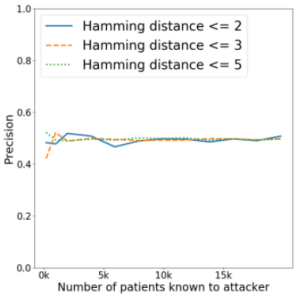
0-17, M



(a)

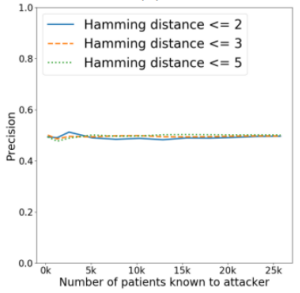


(b)

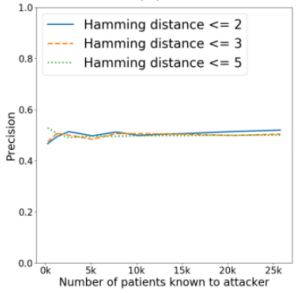


(c)

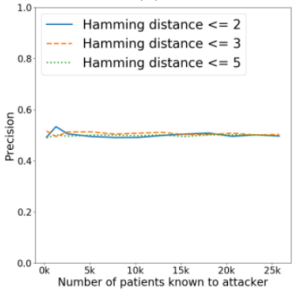
18-44, M



(d)

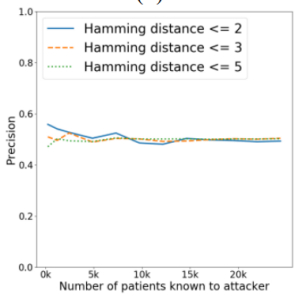


(e)

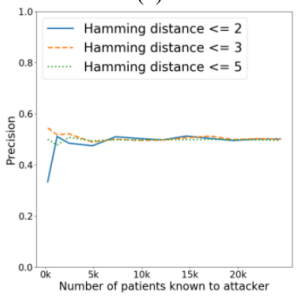


(f)

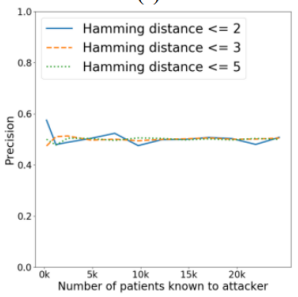
45-64, M



(g)

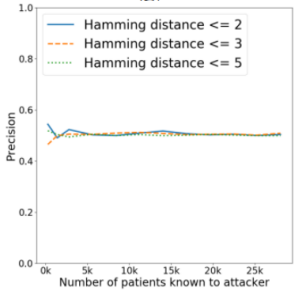


(h)

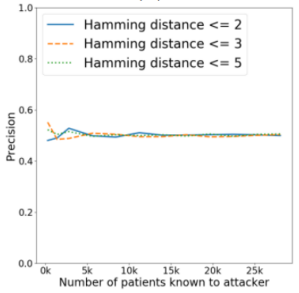


(i)

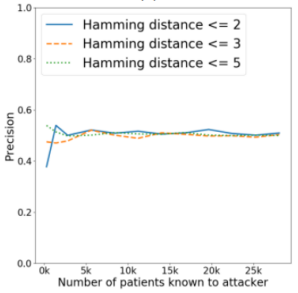
>64, M



(j)



(k)



(l)

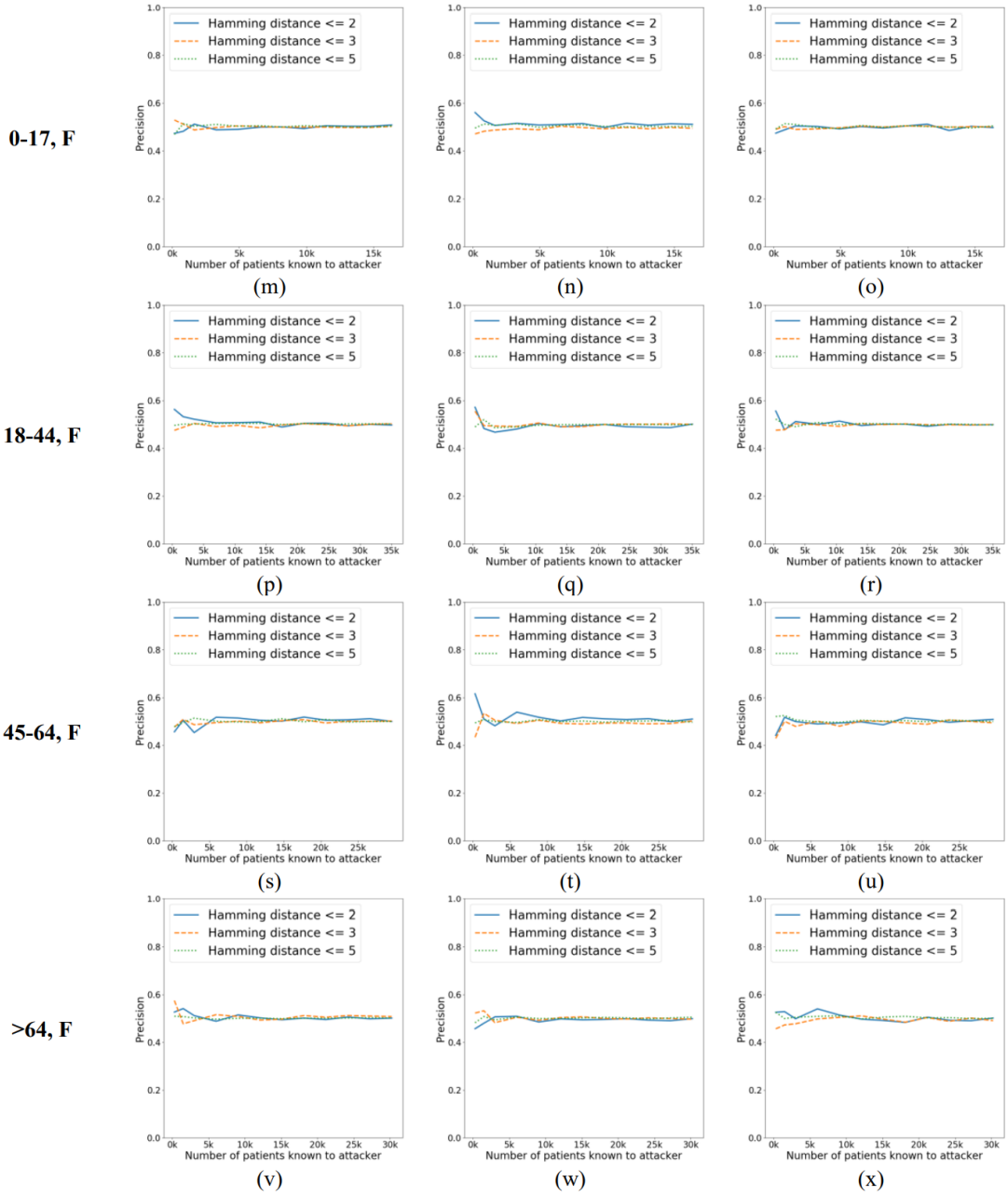
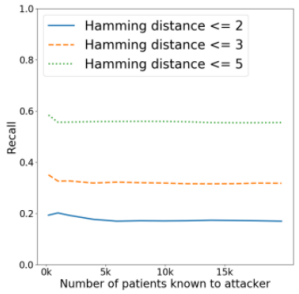
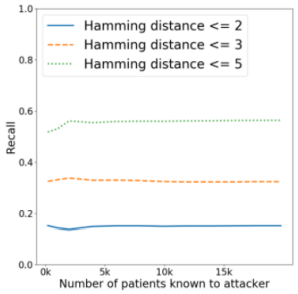


Figure 2.5: Precision of membership inference in subpopulations as a function of the number of patients' records known to an attacker. The first, second, and third column of subfigures correspond to medBGAN, medWGAN, and EMR-WGAN, respectively.

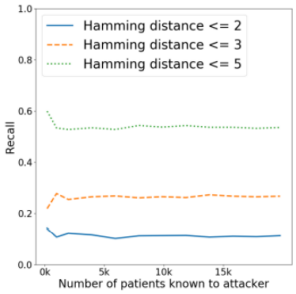
0-17, M



(a)

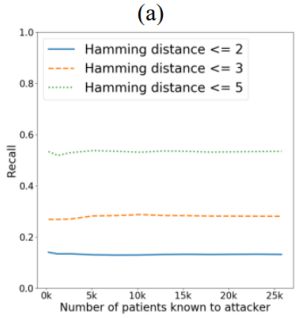


(b)

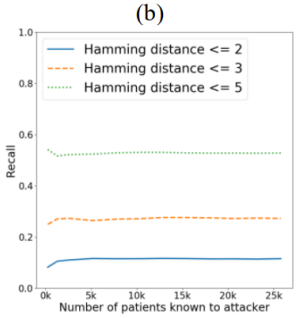


(c)

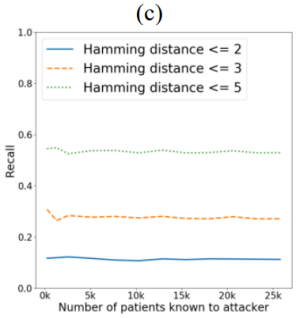
18-44, M



(d)

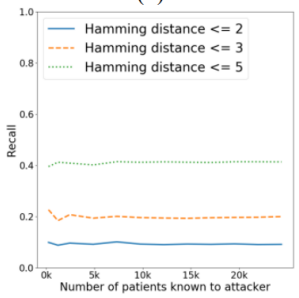


(e)

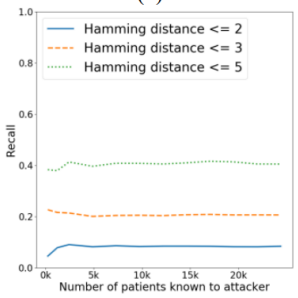


(f)

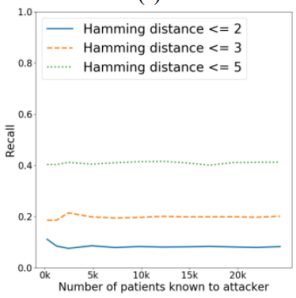
45-64, M



(g)

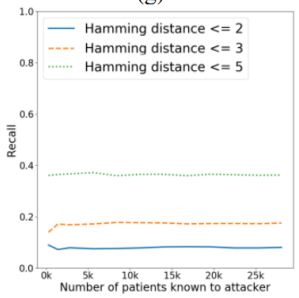


(h)

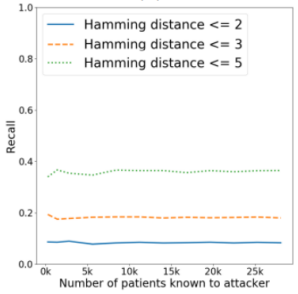


(i)

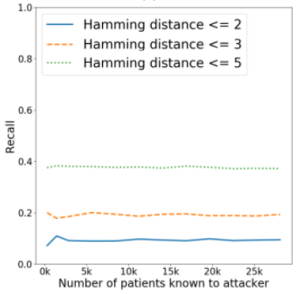
>64, M



(j)



(k)



(l)

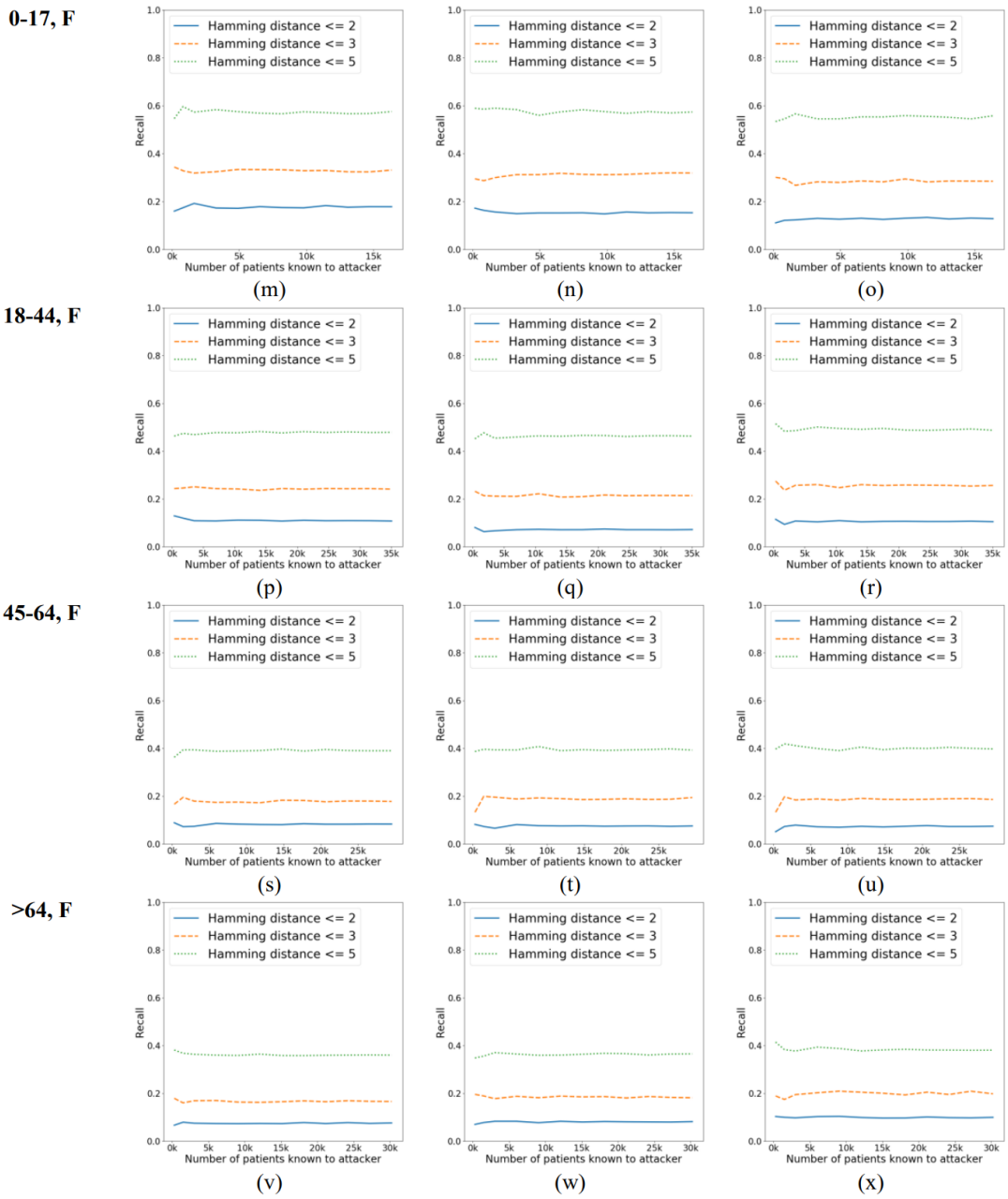


Figure 2.6: Recall of membership inference in subpopulations as a function of the number of patients' records known to an attacker. The first, second, and third column of subfigures correspond to medBGAN, medWGAN, and EMR-WGAN, respectively.

Attribute Inference

We randomly sample 10% records from the training dataset as the partially compromised records. Note that for mimicking real attacks, we constrain the number of ICD-9 codes in any of the sampled records to be greater than a percentage threshold (4%) with respect to

the number of known binary status of codes to an attacker.

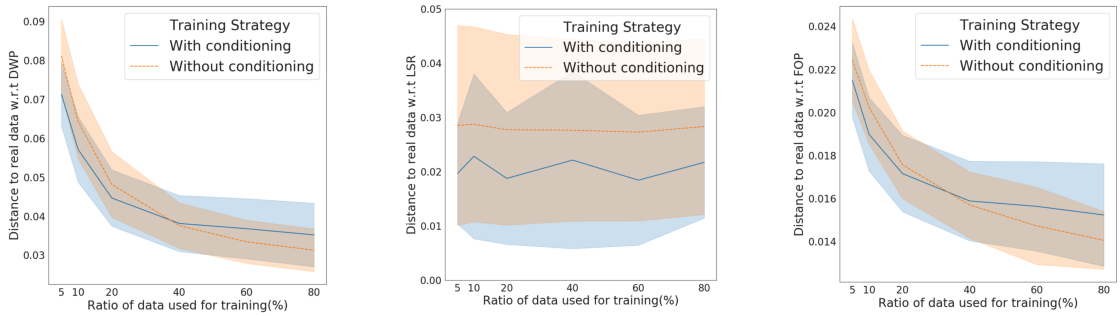
Table 2.2 presents the F1 scores of attribute inference of three different GANs in 8 populations, where we vary the number of known features to an attacker and the number of nearest neighbors. There are several observations to note. First, for all GANs, the risk of privacy disclosure is larger when k is small. This is because more noise will be incorporated by considering a larger number of neighbors in inferring features. Second, as expected, for all GANs, in F1 scores grow with the number of features known to the attacker. In other words, the inference will be more precise when an attacker knows more information about patients. Third, on average (across all populations), the risk of privacy disclosure by EMR-WGAN is less than the risk induced by medWGAN, whereas there is no obvious dominance between medBGAN and EMR-WGAN. Thus, it can be concluded EMR-WGAN achieves a similar privacy risk level as other GAN models with respect to attribute inference. However, we acknowledge that other GAN models are not guaranteed to pose a low privacy risk. As such, this result does not necessarily imply that EMR-WGAN is immune to attribute inference.

Reproduction Rate

Table 2.3 shows the reproduction rate (as well as standard deviation) of medBGAN, medWGAN and EMR-WGAN for the eight subpopulations. For each GAN and each population, we generate synthetic data 10 times and report the average. It can be seen that the reproduction rate for EMR-WGAN is less than 1%. This indicates that our model is capable of producing novel medical records rather than simply remembering the training data. As a consequence, this implies that the risk of EMR-WGAN causing identity disclosure through data reproduction is small. It should also be recognized that medBGAN has higher reproduction rates than EMR-WGAN for all populations, whereas medWGAN shows lower reproduction rates.

2.5.2 Evaluating the Training Strategy

We compare the simple training (based on EMR-WGAN) and conditional training strategy (based on EMR-CWGAN) by assessing utility (including DWP, LSR and FOP). We varied the training set to determine how it influenced the utility. The results are in Figure 2.7, where we present the average and standard deviation of each utility measure across each subpopulation.



(a) DWP (b) LSR (c) FOP

Figure 2.7: A comparison of three utility measures on two training strategies.

In Figure 2.7a, we report the difference in the mean of the dot-to-diagonal distribution between the simulated and real data in DWP. For LSR, as shown in Figure 2.7b, we report the difference in the mean of variance distribution between the simulated and real data. In Figure 2.7c, we show the FOP distance between simulated and real data.

As can be seen, when the training dataset is small (towards the left of the figures), conditional training outperforms traditional training. Specifically, when the size of available data for training is less than 35% of the original, conditional training can achieve higher data utility with respect to DWP and FOP. With respect to LSR, conditional training leads to a better utility than simple training.

We believe this is because the complex associations between diseases may cross the boundary of populations with different concept labels. In other words, when the available training dataset size is small, real EMR records with their concept labels different from the simulation task can help strengthen the signals characterizing the statistical properties

between code features.

2.6 Discussions

This study has several notable implications for the design and deployment of GAN pipelines. First, applying Wasserstein Divergence as well as the affiliated optimization techniques provides a GAN model with the ability to solve the problem of generating categorical data. Second, this work illustrates the importance of a comprehensive set of utility measures. Utility measures that only characterize basic statistics can lead to biased or incorrect conclusions. Third, conditional training is more useful in generating EMR data than simple training strategy when real data for training is small. Such finding makes a clear suggestion for the EMR generation tasks in the application domain, where the volume of real data is often a bottleneck for learning.

There are several limitations that should be acknowledged. First, we focused on only binary features (e.g., positive assertion or lack of a diagnosis). Further investigation is needed into EMR simulation when continuous features are taken into account. Second, we modeled the EMR in a static manner, yet the trajectory of a disease evolves, often punctuated by various interventions. For example, in the event lab test results should be generated, then time should be considered and modeled accordingly. Third, we focused on an application-agnostic evaluation pipeline of synthetic data and demonstrated the synthetic data quality from the perspective of statistical properties. Yet our result does not sufficiently imply the performance of synthetic data in any specific real-world applications. Fourth, we conducted experiments on a curated version of the EMR dataset. It is unknown if the findings in this chapter from either the utility or privacy perspective would hold when the simulation is performed on the original uncurated feature space. We expect further experiments into the scalability of the proposed methods. In addition, the generalizability of our findings should also be investigated with data from multiple EMR sources. Finally, we note that this analysis considered only the statistical validity of the synthetic records. It is

possible that the synthetic records conflict with known phenomena that a clinical specialist might recognize in the data. It is critical for the data from such synthesis methods to be adjudicated by clinically knowledgeable individuals to aid in their uptake in practice, though it should be recognized that EMR data is inherently noisy, such that generating records that are not in alignment with clinical expectations would not necessarily imply that the proposed methodology has failed to accomplish its goal of creating EMR data.

Table 2.2: F1 scores of attribute inference of GAN models in the subpopulations.

Know Features	Neighbors	GANs	Subpopulation								
			0-17,F	**0-17,M**	**18-44,F**	**18-44,M**	**45-64,F**	**45-64,M**	**>64,F**	**>64,M**	**Average**
128	**1**	**medBGAN**	0.20	0.20	0.16	0.16	0.20	0.19	0.21	0.22	0.19
		medWGAN	0.20	0.20	0.16	0.16	0.18	0.19	0.20	0.21	0.19
		EMR-WGAN	0.20	0.18	0.16	0.18	0.18	0.19	0.20	0.21	0.19
	10	**medBGAN**	0.14	0.14	0.09	0.09	0.10	0.13	0.15	0.16	0.13
		medWGAN	0.16	0.15	0.10	0.11	0.10	0.11	0.15	0.17	0.14
		EMR-WGAN	0.14	0.13	0.09	0.12	0.11	0.13	0.14	0.15	0.12
256	**1**	**medBGAN**	0.23	0.23	0.19	0.18	0.18	0.21	0.24	0.25	0.21
		medWGAN	0.26	0.26	0.20	0.21	0.21	0.23	0.24	0.25	0.23
		EMR-WGAN	0.24	0.23	0.20	0.22	0.21	0.23	0.24	0.23	0.23
	10	**medBGAN**	0.14	0.14	0.08	0.08	0.09	0.11	0.14	0.16	0.12
		medWGAN	0.19	0.19	0.11	0.12	0.11	0.14	0.15	0.18	0.15
		EMR-WGAN	0.16	0.15	0.09	0.13	0.11	0.14	0.15	0.16	0.14

Table 2.3: Reproduction rate in the subpopulations.

GANs	**Subpopulation**				
	Gender	**0-17 (%)**	**18-44 (%)**	**45-64 (%)**	**>64 (%)**
medBGAN	**M**	1.85	0.43	0.66	1.20
	F	2.39	0.72	0.38	0.84
medWGAN	**M**	0.21	0.14	0.10	0.24
	F	0.25	0.03	0.06	0.17
EMR-WGAN	**M**	0.32	0.24	0.21	0.68
	F	0.37	0.09	0.15	0.33

CHAPTER 3

Longitudinal medical record Simulation

3.1 Introduction

The GAN-based simulation techniques for coded event data (e.g., insurance billing codes) [28, 37, 46, 59] introduced in Chapter 2 are limited in that they only generate patient profiles in a static manner, ignoring their temporal characteristics. This is problematic for several reasons. First, current techniques do not accurately reflect how EMRs are recorded, organized and utilized in practice. If synthetic coded data included timestamps for clinical events (e.g., dates or duration from a reference point), they would be better oriented for modeling more complex phenotypes and supporting predictions about outcomes that are time-aware. Second, current techniques lack the capacity to model temporal features. Though the machine learning community explored this problem [60], the resulting approaches focus on partially revising the original records (via GANs) for the purposes of refining the prediction tasks, instead of generating entirely new records.

There are several factors need to be considered to ensure a meaningful simulation of a sequence of episodes. First, healthcare Organization episodes contain a variable number of clinical events (e.g., the number of billed diagnoses changes from episode to episode). As such, it is necessary to design a compact representation for each episode that compresses the space and preserves information in a computable form. Second, we need to learn the temporal correlations between episodes in EMRs. Researchers have successfully leveraged various recurrent neural networks (RNNs) to model patient trajectories and, as an artifact, make predictions about patient outcomes [61, 62, 63, 64, 65]; however, they cannot be directly applied to simulate sequences of episodes. This is because there is often more than one event per episode (e.g., a episode will likely be associated with multiple diagnosis codes). In this setting, the recurrent unit needs to output the joint distribution

of the feature space, instead of the marginal distribution that is utilized by existing models. Third, generative models (e.g., GANs) are often used in this setting to approximate multivariate distributions, but they suffer from the problems of mode collapse (i.e., the generator maps different inputs to the same output) and mode drop (i.e., the generator only captures certain regions of the underlying distribution of the real data) [66, 67]. These problems are magnified when the distributions are characterized by non-convexity (i.e., the probability density function has multiple local maxima). In EMRs, this could happen, for instance, when diagnosis codes exhibit a nonmonotonic probability density over a sequence of episode. To address such challenges, several simulation techniques incorporate advanced divergence measures between the real and synthetic distributions [51, 68, 69] and reorient the optimization strategy [52, 66, 70]. However, there is little evidence that these techniques sufficiently address the mode collapse and mode drop problems in distributions with non-convex densities.

To address these issues, we developed a simulation framework, called Synthetic Temporal EMR Generator (SynTEG), to generate timestamped diagnostic events. This chapter introduces the SynTEG architecture and illustrates its performance by experiment with data from over 500,000 patient records at Vanderbilt University Medical Center (VUMC). The experiment result shows that the system maintains temporal relationships between diagnoses while thwarting membership inference and attribute inference on patient privacy.

3.2 Method

3.2.1 Intuitions into Sequence Modeling

We represent each longitudinal record as $e_1, e_2, ..., e_N$, where e_i denotes the ith episode and N is the total number of episodes. Note that N may vary across records. We can achieve a sequential simulation of the episodes by estimating the probability of e_i given the set of

prior episodes $e_1, e_2, ..., e_{i-1}$:

$$P(e_1, e_2, ..., e_N) = P(e_1) \prod_{i=2}^{N} P(e_i | e_1, e_2, ..., e_{i-1}) \tag{3.1}$$

where $P()$ denotes a probability density function. Equation 3.1 can be decomposed to enable generative modeling as follows:

$$s_i = f(s_{i-1}; e_{i-1}) \tag{3.2}$$

$$e_i \sim P(e_i | s_i) \tag{3.3}$$

where s_i can be thought of as an implicit representation of a patient's health status. Equation 3.2 represents the status transition that transpires when a new episode occurs. Formula 3.3 represents the creation of a single episode based on s_i, which is usually realized by either of an explicit-density method or an implicit-density method. To illustrate both methods, we define $C_i = \{c_{i,1}, \ldots, c_{i,M_i}\}$ as the set of clinical event concepts associated with e_i, and $\overline{C_i}$ as its complementary set (i.e., the set of concepts that do not exist in e_i).

The explicit-density method in the case of episode simulation can be implemented through a sequencing model [71]:

$$P(e_i | s_i) = P(c_{i,1} | s_i) \prod_{j=2}^{M} P(c_{i,j} | c_{i,1}, c_{i,2}, ..., c_{i,j-1}; s_i) \tag{3.4}$$

In modeling this equation, Vinyals et al. [72] demonstrate that if the training dataset is sufficiently large, the order to apply the chain rule can be arbitrary. However, there are two major obstacles to applying this method. First, modeling equations 3.2 and 3.4 collectively necessitates using a nested autoregressive training paradigm, which may introduce significant exposure bias. Second, sampling from the learned distribution is a non-trivial problem that the natural language generation community has extensively investigated but has yet to find a satisfactory solution [73]. Our experiments indicate that the explicit-density method

does not achieve a satisfactory level of performance.

The implicit-density method performs reparameterization to transform samples drawn from an ordinary latent distribution $P(z)$ (e.g., Gaussian) into the target distribution through a deterministic function:

$$e_i = g(s_i; z); z \sim P(z) \tag{3.5}$$

Reparameterization is typically realized through GANs. However, the GANs require significantly more effort to implement in longitudinal simulation than patient profile simulation introduced in Chapter 2, because the model has to account for both equations 3.2 and 3.5. In theory, the adversarial training process for GANs is guaranteed to converge to the global optimal under two conditions: a) it is modeled as a convex-concave game; and b) there is arbitrarily strong representation power to ensure that the discriminator to be optimized at each update [36]. However, in practice, neither of these assumptions holds true. As indicated by Nagarajan et al. [74], the training objective of GANs is not necessarily convex-concave, even when using a one-layer generator and discriminator. As a result, the training of GANs is frequently not even locally convergent [75], rendering training instabilities or even failures. Gradient-based regularization [52, 76] on the discriminator is often adopted to mitigate this issue, but such regularizations limit the expressivity of the model required to support the state transition modeling of 3.2. To address this concern, we propose a framework using a locally supervised training paradigm, in which the training process is split into two gradient-isolated stages that separately model $f()$ and $g()$.

3.2.2 The SynTEG Framework

SynTEG, the structure of which is shown in Figure 3.1, uses two learning stages to model the transition function $f()$ and the episode generator $g()$. We refer to these as the *dependency learning* stage and the *conditional simulation* stage.

Stage 1: Dependency learning. We parameterize $f()$ using an autoregressive model,

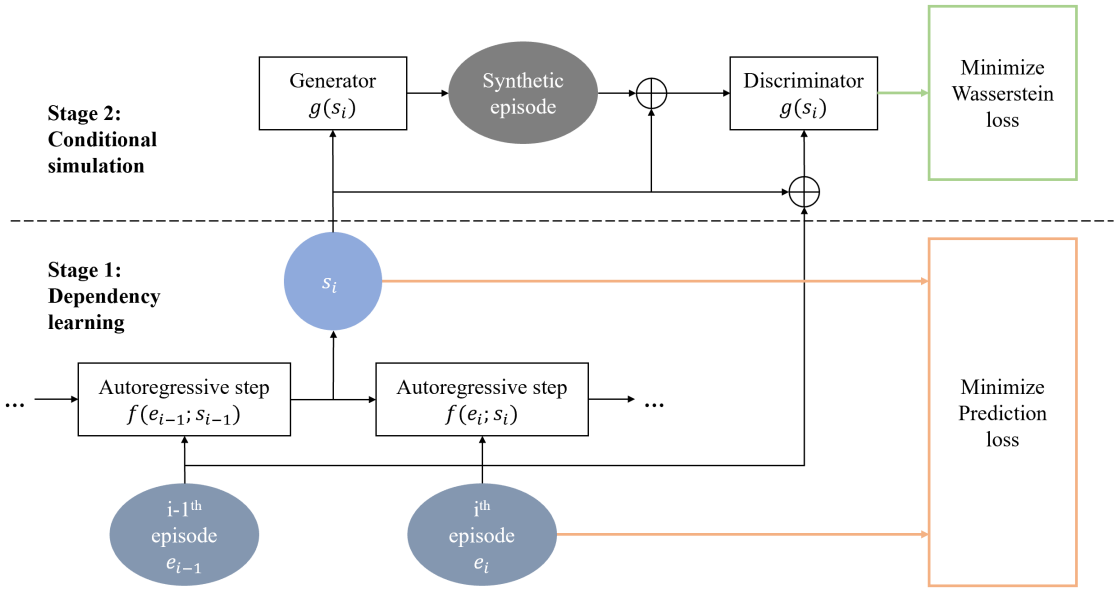


Figure 3.1: A high-level overview of the SynTEG architecture. Each uncolored square box represents a function and each colored oval represents a variable. The parameters of the autoregressive model is optimized to minimize prediction loss, for Dependency Learning (Stage 1). Next, the hidden state of the autoregressive model is extracted as the conditional input of the GANs in the Conditional Simulation (Stage 2). Here the objective is to minimize the Wasserstein divergence between the real and synthetic episodes.

which is trained through a self-supervised learning task with the objective to predict which set of diagnoses will appear in a patient's next episode. Specifically, the training objective can be formalised as

$$\text{Loss}(e_i) = \log\left(1 + \sum_{c \in C_i} \exp\left(-\tilde{c}(s_i)\right)\right) + \log\left(1 + \sum_{c \in \overline{C_i}} \exp\left(\tilde{c}(s_i)\right)\right)$$

where $\tilde{c}(s_i)$ is the logit corresponding to the concept c, which is derived from s_i (e.g. through a Multi-layer perceptron). In doing so, the model is forced to learn a compact representation of status s_i given input e_{i-1} and its previous state s_{i-1}. In addition, we derive a sample dataset that represents the marginal distribution of s as an approximation of the $P(s)$ distribution:

$$P^*(s) = \{f(s_{i-1}, e_{i-1}) | i \in \{2,3,...,N\}\}$$

Stage 2: Conditional simulation. The goal of the second stage is to simulate a multivariate conditional distribution $p(e|s)$ given the condition $P^*(s)$ derived in the first stage. We accomplish this by applying conditional GANs, which includeds the generator $G(e|s)$ and discriminator $D(e,s)$. Specifically, we use the condition version of the EMR-WGAN [28]. The optimization objective with respect to the Wasserstein divergence between $P(e|s)$ and $G(e|s)$ is formalized as

$$\max_{|D|_2<=1} \mathbb{E}_{P^*(s)} \mathbb{E}_{e\sim P(e|s)} D(e,s) - \mathbb{E}_{e\sim G(e|s)} D(e,s)$$

where $|D|_2$ corresponds to the Lipschitz constant of D.

3.2.3 Modeling Time Interval between Episodes

Simulating the timestamp of each episode is a critical component of creating realistic synthetic medical records. Within the SynTEG framework, we model the timestamp associated with each episode by learning the distribution of the interval between adjacent episodes

$$P(t_1, t_2, ..., t_N) = P(t_1) \prod_{i=2}^{N} P(t_i - t_{i-1} | s_{i-1}, e_i). \tag{3.6}$$

The distribution $P(t_i - t_{i-1}|s_{i-1}, e_i)$ is realized through a density function $h(\Delta t_i|s_{i-1}, e_i)$, which we denote as $h(t)$ to reduce clutter. We further define a corresponding intensity function

$$\lambda(t) = \frac{h(t)}{1 - H(t)} \tag{3.7}$$

where $H(t)$ is the cumulative probability that a new episode will occur before duration t. It is worth noting that $\lambda(t)dt$ is equivalent to the expected number of episodes occurring in

an infinitesimal interval dt. Based on $\lambda(t)$, $h(t)$ can be reversely derived as

$$h(t) = \lambda(t)\exp\left(-\int_0^{\mathrm{t}} \lambda(\mathrm{u})\mathrm{du}\right) \tag{3.8}$$

However, this function is intractable without specific assumptions on $\lambda(t)$ that result in a resolvable integral. As such, we adopt the approach proposed by Omi et al. [77], using a neural network to model the cumulative intensity function $\Lambda(t) = \int_0^t \lambda(u)du$ instead of $\lambda(t)$. In doing so, the optimization objective becomes

$$\log h(t) = \log \frac{\partial \Lambda(t)}{\partial t} - \Lambda(t),$$

and we can directly model $\Lambda(t)$ with a neural network. Notably, considering $\Lambda(t)$ is monotonically increasing and ranges from zero to one, the neural network is constrained to use positive kernels associated with t, and to use the softplus function as the activation of the final layer. After obtaining $h(t)$, we can calculate $H(t)$ with Equations 3.7 and 3.8.

3.2.4 Utility Evaluation

To the best of our knowledge, there are no standard utility functions for synthetic temporal EMRs. Thus, we introduce several measures, which characterize the extent to which the simulated data retains 1) correlations between temporal features and 2) a general representation capacity with respect to forecast future diagnosis. We further measure the extent to which the trajectory of well-known chronic diseases are represented in the synthetic data.

Bernoulli Success Probability (BSP)

This measure investigates the degree to which the distribution of each clinical event concept among the generated records is similar to real data. We compare the Bernoulli Success Probability of each concept in an episode, and each concept in a episode conditioning on the presence of any other concept, in real and synthetic data.

First-order Temporal Statistics (FTS)

This measure evaluates the extent to which the synthetic data retains the time-related characteristics of diagnosis features of the real data. Specifically, for each unique diagnosis code, we calculate the mean and standard deviation of 1) occurrence age (i.e., the age associated with an episode containing the diagnosis code) and 2) the time between the episode containing this diagnosis code and the following episode, which we refer to as inter-episode interval. We refer to these as the occurrence and recurrence statistics, respectively. The larger the difference in the statistics learned from the real and synthetic data, the more biased the model is in the time-related characteristics of the diagnoses.

Diagnosis Forecast Analysis (DFA)

This measure evaluates the extent to which the synthetic data remains useful for the secondary uses (e.g., predictions about what will happen to a patient in the future). To do so, we follow the train on synthetic and test on real (TSTR) [54] strategy. Specifically, we train two models - one on real and one on synthetic data to predict which diagnoses will be realized at a patient's next episode, given the history of previous episodes. When the two models achieve sufficiently similar prediction performance with respect to AUROC when tested on another part of real data, we claim the synthetic data has the same level of representation capacity as the real EMR data.

Latent Temporal Statistics (LTS)

This measure evaluates how well the trajectory of chronic disease is modeled in the synthetic data. Specifically, this is done by comparing the distribution of real and synthetic data over latent variables that were not explicitly modeled.

To perform this analysis, we select four common chronic diseases: 1) Type-2 diabetes (T2D), 2) heart failure, 3) hypertension, and 4) chronic obstructive pulmonary disease

(COPD), which exhibit more prolonged patterns over time than acute diseases. For each disease subpopulation, we draw uniformly at random without replacement two equal-sized matrices M_r and M_s, where each row represents a record and each group of columns represents the diagnoses over a time window, from the real and synthetic data, respectively. The definitions for each subpopulation, as well as the details for the construction of the feature matrices, are provided in Supplemental Appendix B. To compare the temporal patterns for a disease of interest, we decompose M_r into latent factors and assess how well the distributions over those factors are retained in the synthetic data.

To do so, we apply singular value decomposition (SVD) on M_r to obtain its right singular vectors and the corresponding singular values. We then project M_r and M_s to the new (low-dimensional) space whose bases correspond to the selected singular vectors to generate a set of latent features. Finally, we compute the Kolmogorov-Smirnov statistic (which is the maximum vertical distance between the empirical cumulative distribution functions) between the two projections as a measure of the distance between the real and synthetic distributions along each vector. We compute the weighted average of the statistic across all latent features, weighted by their corresponding singular values. We refer to this value as the *weighted latent difference*. The closer the weighted latent difference is to zero, the closer the distributions of the two datasets are. We compare the weighted latent difference for real vs. real subsets against real vs. synthetic subsets to understand how well the temporal patterns are preserved. To investigate the stability of this weighted latent difference, we repeat this 100 times by randomly sampling both the real and synthetic data.

3.2.5 Privacy Evaluation

Privacy risk measures have been defined for structured billing codes simulated in a static setting, as introduced in Chapter 2, but not a temporal setting. As such, we adapted privacy risk measures for two known adversarial scenarios: membership inference and attribute disclosure attacks.

Membership Inference

Though designed to generate synthetic clinical event data, a generative model may reveal membership information for real records. More specifically, an attacker who has information about a set of real patient records may leverage the synthetic records to infer whether the corresponding records were in the training dataset of the generative model. Once a patient's membership is known, additional information associated with the dataset (which may be sensitive) would be revealed. Thus, we investigate the extent to which an attacker can leverage a synthetic dataset to distinguish between records used in the training set and those not in the set. This attack is evaluated in the following manner. First, we define split CSD into a training dataset D_1 and a holdout dataset D_2. we use SynTEG to generate a synthetic dataset with size equal to that of D_1. Second, we train an autoregressive model (which is the same model used in the dependency learning stage) over S. This yields a probabilistic model $p_S(e_i|s_{i-1}, e_{i-1})$. Third, we compute the perplexity of records in D_1 and D_2 according to this model, which is defined as:

$$perplexity_S(r) = -\frac{1}{N}\log\prod_{i=1}^{N} p_S(e_i|s_{i-1}, e_{i-1})$$

In this sense, perplexity serves as a proxy of the log likelihood of a record. Finally, we compare the perplexity distributions between D_1 and D_2 by assessing the R^2 of the quantile-quantile regression and estimated KL-divergence between the distributions.

Attribute Disclosure

It is possible that a generative model, when poorly designed or trained, can leak information about the patients' records in the training data. In this scenario, it is assumed that an attacker is aware of the identities of certain real records, referred to as partially compromised records. The attacker then attempts to learn about attributes that they were not

aware of (e.g., a particular diagnosis). We investigate the risk that an attacker can infer the unknown attributes by leveraging the synthetic dataset.

Previous approaches to attribute disclosure make inferences through a majority vote of the synthetic records that have shortest distance to the partially compromised record [28, 37, 46]. However, this strategy is likely to underestimate the risk because it does not consider the prior knowledge an adversary may have with respect to the attribute. Thus, in this chapter, we assume the worst-case scenario, whereby the attacker has prior knowledge about each of the diagnosis codes in this study (i.e., the dependencies between diagnosis codes derived from statistical inference on the real dataset).

To measure the attribute disclosure risk induced by a temporal clinical event data simulation model, we assume that the attacker determines an attribute is realized for a patient if the predicted likelihood leveraging synthetic data is a threshold greater than the value given by prior knowledge. Since the prior knowledge derived from the real data has some natural level of variance due to sampling, this could lead to a biased risk estimation (for both the true positive rate and false positive rate of an attacker's inference). To address this issue, we add a *Control* group, which simulates risk estimation in the situation where no information is leaked. The evaluation process is illustrated as follows.

We first define the projection operation of "$\circ$" and the mask operation "[]" on the record level: $r \circ \text{Attr}$ represent the binary status of an attribute Attr(i.e. a phecode) in record r; $r[\text{Attr}]$ is the partial representation of r such that the presence status of the phecode Attr is masked to be negative (i.e., 0). We train a probabilistic model P_r to maximize

$$\mathbb{E}_{r \in D_1'} P(r \circ \text{Attr} | r[\text{Attr}])$$

We use a early stop strategy such that we end the training process when

$$\mathbb{E}_{r \in D_1'} P(r \circ \text{Attr} | r[\text{Attr}]) = \mathbb{E}_{r \in D_1''} P(r \circ \text{Attr} | r[\text{Attr}]),$$

where $D_1^{'}$ and $D_1^{''}$ are subsets of D_1. We also train P_s through a similar process, but without early stop, on the synthetic dataset as the attack model. Then we evaluate the attribute inference attack as follows. Given Attr and r, if

$$P_s(r \circ \text{Attr} = 1 | r[\text{Attr}]) - P_r(r \circ \text{Attr} = 1 | r[\text{Attr}]) > t,$$

where t is a threshold, the attack model predicts Attr as positive for r. To set up the *Control* group, we train another model with the same structure on D_2. Considering this model has no access to D_1, it cannot leak private information in D_1 beyond prior knowledge, and thus can be used as a calibration to evaluate the risk of attribute inference attack.

3.3 Materials

The clinical event data for this study was collected from the Synthetic Derivative (SD) at Vanderbilt University Medical Center, which contains over 2.1 million de-identified EMRs.

We extracted all diagnosis codes (initially encoded as International Classification of Diseases (ICD) billing codes), their timestamps, and the demographics of the corresponding patients from 2,187,629 records. The ICD codes were mapped to Phenome-wide Association Studies (PheWAS) codes, or phecodes, which aggregate billing codes into clinically meaningful phenotypes [78, 79]. The phecodes for each record were then grouped into episodes according to the corresponding timestamp at billing (i.e., each group contains all phecodes billed on the same calendar day). In doing so, each record was represented as a sequence of episodes, each of which was represented by 1) a binary vector over the attribute space, indicating the presence/absence of diagnoses, and 2) the corresponding timestamp. We refer to this as the SD dataset.

To mitigate noise in the data, we refined SD in several ways to obtain a subset that we refer to as the clean SD (or CSD). First, we removed records with fewer than 10 episodes. This was done because, upon review, it was evident that the majority of such records lacked sufficient syntactic structure between episodes (i.e., ordinal pattern of phecodes) and, thus,

would not assist in temporal modeling. Moreover, such records might correspond to individuals who received a non-trivial amount of their healthcare out of VUMC system. Second, we removed low prevalence phecodes (less than 2000 occurrences) because they contributed little to the learning process while inducing high sparsity and bias. Third, we removed records with more than 200 episodes or 35 different phecodes in a single episode for computational efficiency. This resulted in a dataset of 580,054 records (covering 1,276 unique phecodes).

The summary statistics for the SD and CSD datasets are shown in Table 3.1.

Table 3.1: Summary statistics for the clinical event datasets used in this study.

Dataset	Patient Records	Gender	Age Distribution 0-17, 18-44, 45-64, >64		
SD	2,187,629	M:47% F:53%	31%, 30%, 23%, 16%		
CSD	580,054	M:46%: F:54%	21%, 27%, 28%, 24%		
Dataset	**Phecodes Codes**	**Codes Per Patient**	**Patients Per Codes**	**Episode Per Record**	**Codes Per Episode**
SD	1797	9.79	12,031	12.12	2.27
CSD	1276	23.17	16,575	32.56	2.26

3.4 Experimental Design and Result

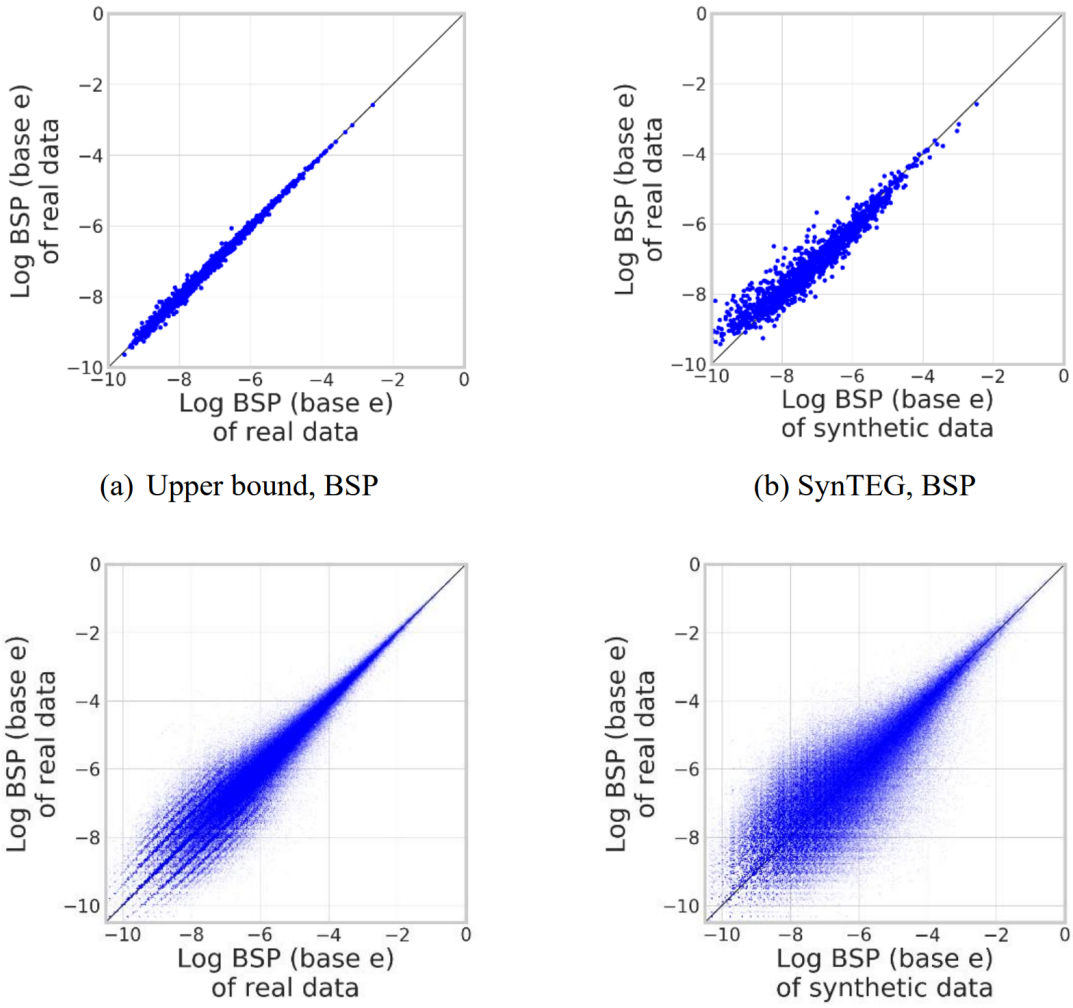
We randomly split CSD into 85% for training and 15% for testing (i.e., holdout) sets, referred to as D_1 and D_2, respectively. We applied the former to train the SynTEG model, which generates a synthetic dataset. We assess the utility and privacy using the three sets, including the testing set, and random samplings of training and synthetic set with the same number of records as the testing set. We use the similarity between the synthetic and testing set as indication of synthetic data's quality, and the similarity between the training and testing set as the upper bound of our measurements.

3.4.1 Utility Analysis

Bernoulli Success Probability

The BSP results are shown in Figure 3.2. As can be seen from Figure 3.2a, all dots distribute closely along the 45-degree diagonal line, which suggests that the BSP of phecodes in the real data is highly stable. As shown in Figure 3.2b, the synthetic data achieves a similar pattern; however to a slightly lesser degree than the real *vs.* real setting (the mean relative differences are 0.9% and 3.9% for Figures 3.2a and 3.2b, respectively). It is also notable that there is no obvious bias in the real *vs.* synthetic setting with respect to the BSP. Similar observations can be made with respect to the conditional BSP, which is illustrated

in Figures 3.2c and 3.2d, except that the phecodes with lower presence frequency show much less stability in both the real *vs.* real and real *vs.* synthetic settings than on the BSP (the mean relative differences weighted by log frequency of phecodes are 5.5% and 9.6% for Figures 3.2c and 3.2d, respectively).



(a) Upper bound, BSP

(b) SynTEG, BSP

(c) Upper bound, conditional BSP

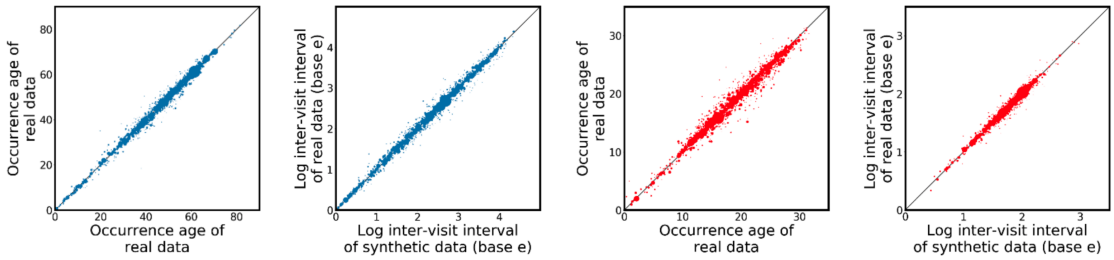
(d) SynTEG, conditional BSP

Figure 3.2: Bernoulli Success Probability (BSP) in the a, c) real *vs.* real setting and b, d) real *vs.* synthetic setting.

First-order Temporal Statistics

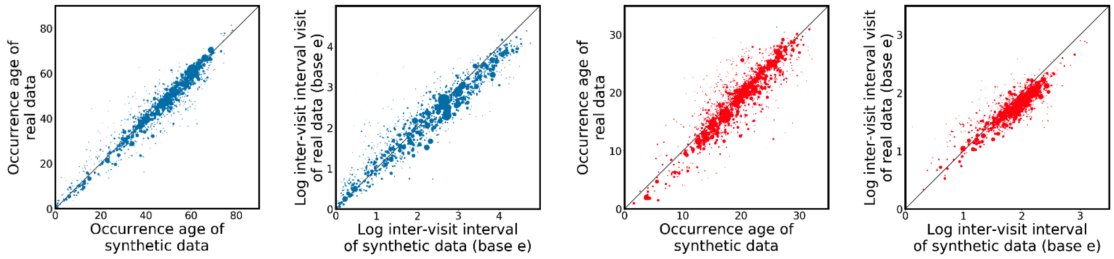
The FTS results are shown in Figure 3.3, where each point corresponds to a phecode. As

can be seen in the four subfigures in the top row, both the occurrence age and the time until the next episode are stable (with respect to the mean and standard deviation), indicating a lack of bias in the real *vs.* real setting. By comparing Figures 3.3e-h with 3a-d, it can be seen that the real *vs.* synthetic setting exhibits a similar pattern, though with a slightly higher variance (the mean absolute relative difference weighted by the log of number of cases in Figures 3.3a-d and 3.3e-h are 3.7% *vs.* 4.9%, 11.9% *vs.* 14.2%, 2.5% *vs.* 4.2%, 10.3% *vs.* 15.2%, respectively). This suggests that SynTEG can capture the distribution of occurrence age and inter-episode interval of each phecode with little bias. It further suggests that the temporal characteristics of the synthetic data are highly similar to the real data.



(a) Occurrence mean (b) Recurrence mean (c) Occurrence standard deviation (d) Recurrence standard deviation

Upper bound (real EHR data only)



(e) Occurrence mean (f) Recurrence mean (g) Occurrence standard deviation (h) Recurrence standard deviation

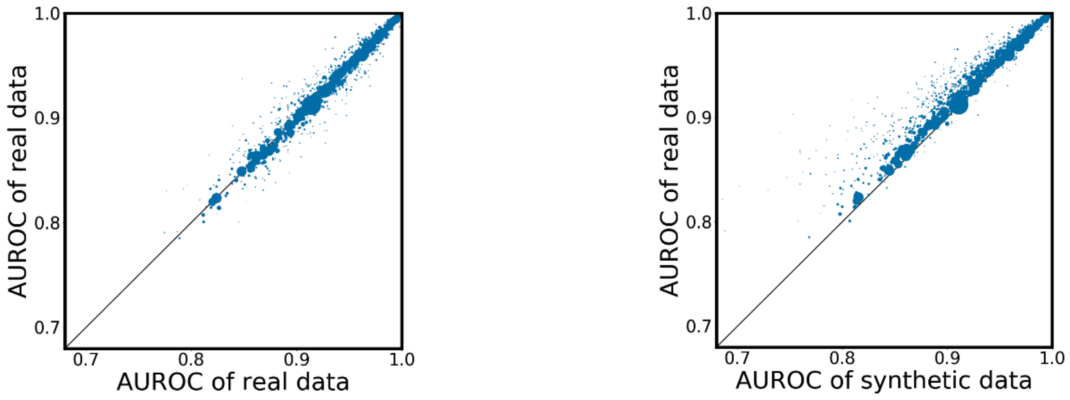
SynTEG

Figure 3.3: First-order temporal statistics for 1,276 phecodes in the real *vs.* real setting (a – d) and the real *vs.* synthetic setting (e – h). The size of each dot represents the number of records with the corresponding code.

Diagnosis Forecast Analysis

The DFA results are shown in Figure 3.4, where each point corresponds to a phecode. It can be seen that most points are close to the 45-degree diagonal line (which is where a perfect statistical replication would present). As can be seen by the size of the dots, the phecodes that diverge from this line correspond to those lacking a sufficient number of training instances.

The mean and standard deviation of absolute relative difference (weighted by the log of the number of patient records affiliated with a phecode) for the real *vs.* synthetic setting are 1.6% and 3.8%, compared to 0.7% and 0.9% for the real *vs.* real setting, which indicates the model trained on synthetic data achieves a similar prediction performance on most of phecodes as the model trained on real data. This result suggests that the synthetic data generated by SynTEG has close capability to real data on predicting future diagnosis.



(a) Upper bound (real EHR data only) (b) SynTEG

Figure 3.4: Disease forecast results in the a) real *vs.* real setting and b) real *vs.* synthetic setting. The size of each dot represents the number of records containing the corresponding code.

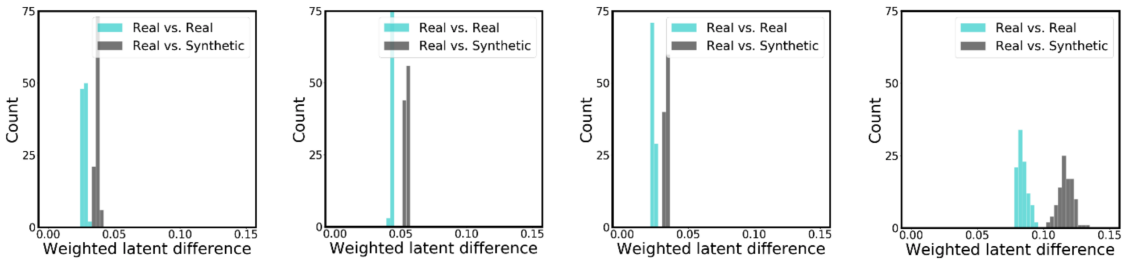
Latent Temporal Statistics

The LTS results are shown in Figure 3.5, where the histograms represent the results of 100 independent samplings. For the real *vs.* real setting, M_r is drawn from D_2, while M_s is

drawn from D_1. For the real vs. synthetic setting, M_r is drawn from synthetic data, while M_s is drawn from D_1. There are several notable findings to highlight. First, the real vs. real histograms show that the weighted latent differences fall in a narrow distribution centered below 0.1 but above zero (medians of 0.027, 0.024, 0.043, 0.084 for each subpopulation, respectively). This indicates that the latent features discovered in each of the real data samples are relatively stable, and gives an idea of how much of a difference we should expect due to sampling variation alone.

Second, we observed that there is more variation in the COPD subpopulation (Figure 3.5d) than the in other disease subpopulations. One possible reason is that there are not a sufficient number of records for the COPD subpopulation to sufficiently represent the latent space (there are only 611 records in the selected subpopulation of COPD, while T2D, heart failure, and hypertension were affiliated with 4969, 4161, and 8836 records respectively).

Third, as can be observed from the real *vs.* synthetic histograms in all subfigures, the distribution of latent features in synthetic data has a modest difference from the real data (the medians of the weighted latent difference are 0.037, 0.033, 0.054, 0.117 for each subpopulation, respectively), usually less than twice the difference expected from random sampling alone. These results suggest that our model can capture reasonably well the long-term dependencies in clinical event data and simulate temporal patterns of diseases.



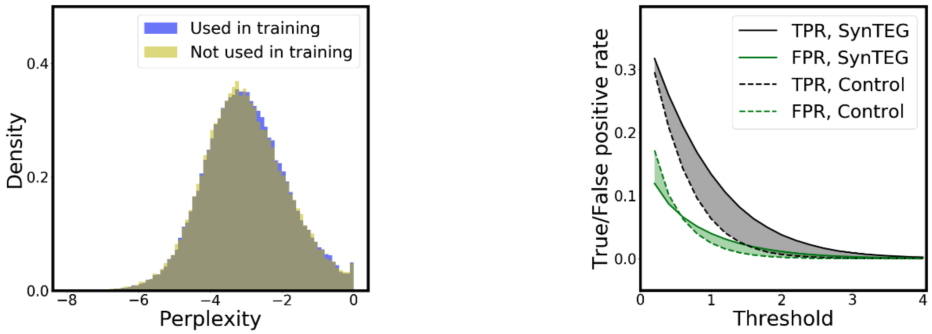
(a) Type-2 diabetes (b) Heart failure (c) Hypertension (d) COPD

Figure 3.5: Histograms for the latent temporal statistics from the experimental results of 100 independent samplings.

3.4.2 Privacy Analysis

The membership inference results are shown in Figure 3.6a. It can be seen that the perplexity distributions for datasets D_1(used for training SynTEG) and D_2 (not used for training) are almost the same. The R^2 of the quantile-quantile regression is 0.9997, while the estimated KL-divergence, based on 1,000 samples, is 0.0093. This indicates that the model learned from the synthetic dataset provides similar likelihoods for the real data used in training the generative model and the real data held out of training. As a consequence, it is highly unlikely that an attacker could determine if a certain real record was in the SynTEG training cohort.

Figure 3.6b illustrates the results of the attribute inference attack. It can be seen that when the threshold is small, less than 0.6, SynTEG has a higher true positive rate (TPR) and lower false positive rate (FPR) than Control. However, the differences are both less than 0.05. With a threshold larger than 0.6, SynTEG still exhibits a higher TPR but the differences are never greater than 0.07, while its FPR is also higher. The difference between SynTEG and Control in FPR and TPR are both not statistically significant, which suggests the potential risk of attribute inference leveraging synthetic data generated by SynTEG is at a low level.



(a) The perplexity distributions for real EHR data used in the training and data not used in training.

(b) True and false positive rates (TPR and FPR) of the attribute inference attack.

Figure 3.6: The privacy risk results for the (a) membership inference attack and (b) attribute inference attack.

3.5 Discussion

This chapter has several notable implications with respect to the simulation of temporal coded medical data. First, the experimental findings suggest that a two-stage learning process, based on deep learning, and GANs in particular, can support the generation of realistic diagnosis trajectories with temporal dependencies. Specifically, the patient status representation from stage-1 is informative, such that it can serve as the condition of the temporal generation process. The utility analysis demonstrates that synthetic data enables the prediction of future diagnosis in a highly similar manner to the real data. Moreover, a generative model trained using the entire population can retain temporal relationships for specific subpopulations of patients with chronic diseases. This suggests that the synthetic data may be useful for various applications, such as future disease forecast and clinical phenotyping.

Second, this study indicates that, though real temporal EMRs have more complicated structures and individual-specific features, the proposed generative model, when applied to simulate synthetic sequences of coded diagnoses, leads to negligible privacy risks with respect to membership and attribute inference attacks. Even though we assumed a worst-case scenario for an attribute inference attack (that is, when the attacker has prior statistical knowledge about all diagnosis codes), the privacy risk remains at a very low level. Still, it should be recognized that these results are specific unto SynTEG and it should not be assumed that all generative models will be devoid of privacy risks.

Given these findings, we believe there are several opportunities to resolve the current limitations of this research. First, we focused on the simulation of diagnosis codes events only. However, there is a need to simulate EMRs with various types of medical data, including the combination of discrete (e.g., procedure and diagnosis codes) and continuous features (e.g., laboratory test results, vital signs, and medication dosages). Further investigation will be required to capture the inherent dependency between feature types. Second, the dataset used for experiments in this chapter was curated. The scalability of the proposed generative model to a larger uncurated feature space necessitates further investigation in

terms of utility and privacy. On one hand, simulating phecodes, though beneficial to phenotype related tasks, may overgeneralize certain disease groups (e.g., infectious diseases), leading to reduced utility in the synthetic data. One the other hand, when representing diagnoses using a larger feature space, such as ICD-9 or ICD-10 (which are approximately 7 and 37 times larger than the phecode space, respectively), the data becomes quite sparse, such that the patterns within, as well as between, features could be washed out. We believe that an appropriate granularity of the diagnosis feature space is important for both data utility and learning effect, but is outside the scope of this specific investigation. Third, we performed experiments on data from only one EMR resource. It remains to investigate whether our findings are generalizable to datasets from other sites, which may have a distinct patient distribution. Fourth, as noted earlier, the primary goal of this study is to develop and evaluate a simulation framework for sequences of diagnosis codes, as opposed to the actual health status of patients. To achieve the latter, we suspect that the framework will need to be augmented to account for uncertainty in a patient's condition. For instance, such a representation should, at the very least, should allow for attribution in the form of 1) definitely (not) have, and 2) might (not) have a certain diagnosis. We suspect that this can be accomplished by expanding the feature space, such that each diagnosis is represented as multiple variables (e.g., one variable definite presence of a diagnosis and another variable to represent the potential presence of the diagnosis). Given that these variables would be mutually exclusive, the framework would need to incorporate constraint-based training [46] to ensure that conflicting representations are not simulated. Finally, in measuring the utility of synthetic records, we only investigated their statistical validity in comparison with real data, rather than the clinical reasonableness. It is possible that in a synthetic record the order of two events may conflict with medical knowledge. It is important for the synthetic data to be evaluated by clinical specialists for the purpose of discovering the wrongly generated combination of features. In addition, we also acknowledge that the statistical validity is not sufficient to predict the performance of synthetic data in specific

real world applications. Therefore, further research is required to study the relationship between application-agnostic utility measurements and the actual utility of synthetic data.

3.6 Conclusion

This chapter introduced a generative framework for simulating temporal clinical event data. The framework consists of two primary components: dependency extraction and conditional generation. We designed utility measures focused on temporal statistics and diagnosis forecasting capacity, as well as privacy risk measures for membership and attribute inference in the temporal setting. We illustrated this framework retains data utility while mitigating known privacy threats by training models using approximately half a million patient records. We believe this investigation sets the stage for further investigation with clinical event simulation, with near term opportunities to extend this model to account for multiple types of clinical events (e.g., diagnoses and procedures) with in a scalable fashion (e.g., thousands of variables).

CHAPTER 4

An Enhanced Longitudinal Simulation Framework

4.1 Introduction

The longitudinal simulation strategy introduced in the previous chapter is capable of creating synthetic datasets that preserve the first-order statistics of EMR data (e.g., the frequencies of diagnosis codes and the time interval between adjacent events). However, as we show in this chapter, there are two fundamental problems with these methods. First, they generate synthetic data that becomes less realistic over time. This is because the synthetic sequences gradually drift away from a realistic representation. Second, the current quality evaluation approaches fail to notice the drift, tending to suggest the synthetic and real records are more similar than they actually are.

In this chapter, we address these issues through a new longitudinal EMR simulation framework that incorporates several amendments to the current longitudinal simulation pipeline, as illustrated at the lower area of 4.1. First, to uncover the drift problem and enable a robust synthetic data quality assessment, we refine the evaluation process, which uses a critic to tell the difference between real and synthetic data. Specifically, we provide the critic with knowledge about the generative process to ensure the critic is not easily fooled, which we show is a dilemma in the current evaluation process. Second, to dampen the compounding of errors over time when synthesizing long sequences of clinical events, a key contributor to the drift problem, we amend the generative modeling framework upon which the learning process relies. This is accomplished by leveraging recent findings from the machine learning community, which indicate that an appropriately designed training strategy (e.g., [80]) can mitigate insufficient modeling that leads to poor generalizability. We further introduce a feedback mechanism into the data generation process, in the form of a rejection sampling strategy, to improve the quality of the resulting data.

To illustrate the benefits of these enhancements, we perform a systematic series of experiments with two distinct datasets. The first, which enables reproducibility of our findings, is a publicly accessible data source, the Registered Tier of the NIH-sponsored *All of Us* Research Program [81]. The second is a dataset derived from the EMRs of Vanderbilt University Medical Center patients [82]. We empirically demonstrate that 1) the new generative modeling and the rejection sampling strategy collectively yields synthetic datasets with less drift and better quality and 2) the evaluation process reduces overestimation on the quality of synthetic data.

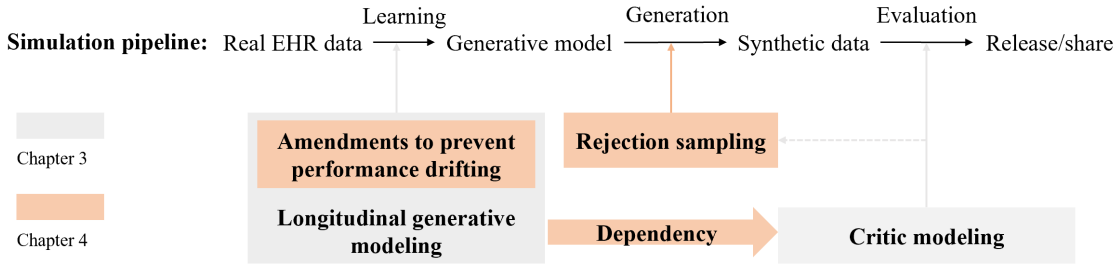


Figure 4.1: A summary of (top) the current longitudinal medical data simulation pipeline and (bottom) the refinements described in this chapter.

4.2 Mitigating Drift with Condition Fuzzing and Regularization

The SynTEG framework introduced in Chapter 3 uses a model training paradigm that includes two gradient-isolated stages for longitudinal record simulation. Specifically, in the *dependency learning* stage, an autoregressive model, denoted by $f()$, is trained to represent the status transition that transpires when a new episode occurs; while in the *conditional simulation* stage, a conditional generative model, denoted as $g()$, is trained to create single episode at each autoregressive step based on the output of $f()$. It should be noted that, in the model training phase, the input for $f()$ is a sequence of real episodes, whereas in the generation phase, the input for $f()$ corresponds to previously generated episodes. Figure 4.2 provides an illustration of this training paradigm.

This paradigm enables the synthesis of medical records with an arbitrary number of episodes. However, as our experimental results below illustrate, under the current imple-

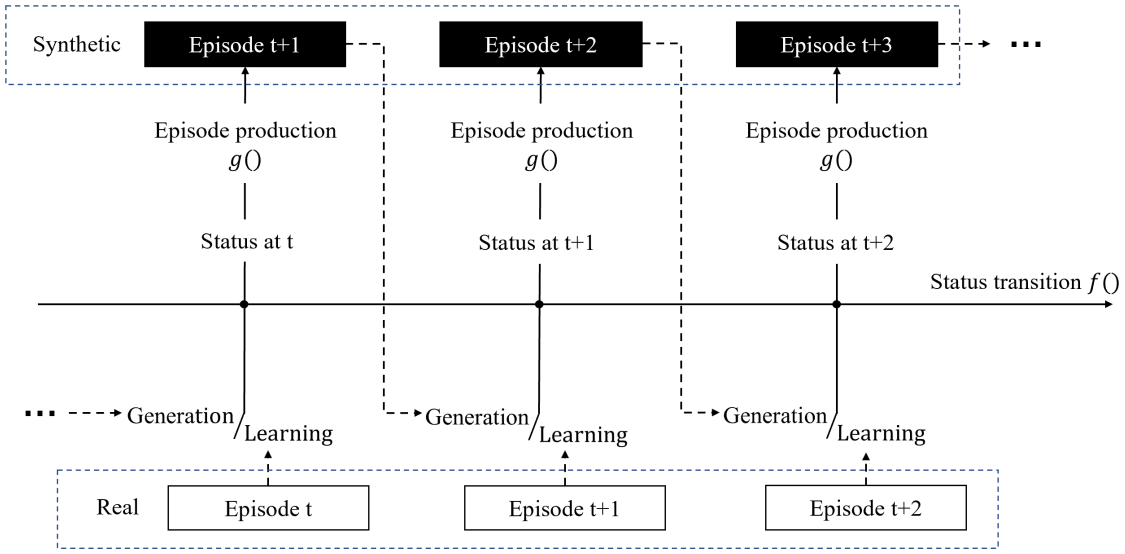


Figure 4.2: The longitudinal medical data synthesis process. The white and black boxes represent real and synthetic episodes, respectively. The Generation/Learning switch indicates that the status representation of the model is updated by 1) previously generated synthetic episodes in the generation phase and 2) ground truth real episodes in the training phase.

mentation of this paradigm, the quality of synthetic records rapidly worsens as the number of episodes grows. This problem is the manifestation of a phenomenon we refer to as drifting – $g()$ makes errors (i.e., generating unrealistic episodes) with a non-negligible and continually increasing rate at each step of the synthesis process.

We mitigate drift by enabling the synthesis process to enhance its resilience against the self-reinforcement of error that occurs from exposure bias [83, 84]. This occurs when a model is exposed to real medical data but not episodes generated by itself in the learning phase. Specifically, this is achieved through two amendments to the learning process.

First, we orient s_t to preserve the mutual information between the sequence of episodes $e_1, \ldots, e_t$ and the subsequent episode e_{t+1}. In doing so, s_t can provide more assistance to $g()$ for capturing the episode distribution. This decreases the overall chance that $g()$ makes an error at an arbitrary step. This is achieved by minimizing the contrastive loss [85] in training $f()$:

$$L = -\mathbb{E}_t \log \frac{\exp(\text{sim}(h(e_{t+1}), s_t))}{\sum_{e' \in V, e' \neq e_{t+1}} \exp(\text{sim}(h(e'), s_t))},$$

where V is the set of episodes from all patients, $h()$ is an auxiliary encoder, $\text{sim}(u,v) = \frac{1}{\varepsilon} \frac{uv^T}{\|u\|\|v\|}$ is the scaled dot product between the L2 normalized u and v, and $\varepsilon > 0$ is an adjustable hyperparameter. According to Oord and colleagues [85], minimizing L is equivalent to maximizing the upper bound of mutual information. Given the computational challenges associated with considering all of the episodes of all patients, we randomly sample a subset of patients and select episodes associated with these patients as an approximation of V in each training step.

Second, we add controlled noise to s_t before it is fed to $g()$ in the training phase. This makes the support of the s_t distribution (i.e., the set of possible values of s_t) larger in the learning phase, which reduces the difference between the support of the s_t distributions in the learning and generation phases. The noise addition procedure is formulated as:

$$\alpha \sim \text{Uniform}(0, \alpha_0);$$

$$s_t \leftarrow s_t + N(0, \alpha I s_t).$$

where α_0 is a predefined threshold, I is the identity matrix, and $N(0, \alpha I s_t)$ denotes a multivariate normal distribution with mean 0 and diagonal covariance $\alpha I s_t$. In doing so, the cosine distance between the fuzzed and original s_t is in a uniform distribution determined by α_0. In addition, we further normalize s_t in the Euclidian space as this practice improves the performance of downstream tasks [86, 87, 88].

4.3 Auditing the Generation Process

In practice, it is difficult to train generative models that precisely represent the target data distribution, particularly when the data is of high dimensionality. By contrast, measuring the distributional divergence between the real and synthetic data is a much easier task. This

allows us to enhance the quality of synthetic datasets even if the learning process cannot be further improved. We accomplish this by auditing the generation process through a feedback mechanism based on rejection sampling [89]. Formally, if real and synthetic data distributions P and Q exhibit the same support, then a sample of P can be obtained by repeatedly accepting an instance r from Q with a probability proportional to $P(r)/Q(r)$.

In practice, $P(r)/Q(r)$ can be approximated by the marginal distribution over a limited set of features. For example, we only consider a binary variable v, such that:

$$P(v=1) \sim \text{Bernoulli}(a),\ Q(v=1) \sim \text{ Bernoulli}(b).$$

This means that we can use $P(v=1)/Q(v=1) = a/b$ to perform rejection sampling. However, we can also perform a more powerful rejection sampling (regarding the quality of the resulting datasets) by including more features. We provide a more refined implementation to estimate $P(r)/Q(r)$ with critic modeling in the Critic implementation subsection of this chapter (See Equation 4.2).

Following the implementation proposed by Azadi and colleagues [90], r is accepted into the final synthetic dataset if, and only if:

$$\phi(r) < \left(1 - \exp(\gamma) + \exp(-\frac{P(r)}{Q(r)} + M + \gamma)\right)^{-1} \tag{4.1}$$

where $\phi(r) \sim \text{Uniform}(0,1)$, $M = max_r P(r)/Q(r)$, and γ is a constant used to normalize the computation and to adjust the overall acceptance rate.

However, due to drift, the sampling process is more likely to reject records containing more episodes. As such, using this implementation may lead to synthetic datasets composed only of a small number of episodes. To mitigate this problem, we refine the sampling process by calculating the overall acceptance rate of the records with the same number of episodes using criteria 4.1, denoted as $\alpha(n)$ for any number of episodes n. We then revise

$\phi(r)$ to be:

$$\phi(r) \sim \text{Uniform}\left(0, \frac{\alpha(N(r))}{\max_r \alpha(N(r))}\right)$$

where $N(r)$ represents the number of episodes for record r. The detailed implementation of rejection sampling for medical record generation is provided in Algorithm 4.1.

Algorithm 4.1: Rejection Sampling for Longitudinal medical record Generation

Input : Trained generative model G;
Real dataset R;
Model $c(r)$ as an estimation of $P(r)/Q(r)$;
pre-defined constant γ

Output: Synthetic dataset S

Create S with G so that $len(S) = len(R)$;
Calculate $M = \max_{r \in S} c(r)$;
Reset S to an empty set;
Create empty dictionaries *acceptedCount* and *totalCount*;
while $len(S) < len(R)$ **do**
  Synthesize a synthetic record r with G;
  $totalCount[N(r)] += 1$;
  if $\phi(r) \sim \text{Uniform}(0,1) < 1/(1-\exp(\gamma)+\exp(-c(r)+M+\gamma))$ **then**
    $S.add(r)$;
    $acceptedCount[N(r)] += 1$;
  end
end
Calculate $\alpha(n) = \frac{acceptedCount(n)}{totalCount(n)}$ for $n \in 1, 2, \ldots, \max_{r \in R} N(r)$;
Reset S to an empty set;
while $len(S) < len(R)$ **do**
  Synthesize a synthetic record r with G;
  $totalCount[N(r)] += 1$;
  if $\phi(r) \sim \text{Uniform}(0, \alpha(N(r))/\max_{r\alpha(N(r))}) <$
  $1/(1-\exp(\gamma)+\exp(-c(r)+M+\gamma))$ **then**
    $S.add(r)$;
    $acceptedCount[N(r)] += 1$;
  end
end

4.4 Synthetic medical data Quality Evaluation

A reliable tool for quality evaluation is as important to a medical data simulation framework as a capable generative approach. We specifically focus on an application-agnostic evaluation because the downstream application of synthetic medical data might not be known at the time of simulation.

This section starts with a brief introduction to related evaluation methods. We next introduce the notion of *critic modeling*, which determines the distinguishability between real and synthetic health data. This notion which has been widely used for application-agnostic evaluation [91, 92, 93]. At last, a new implementation strategy of critic modeling is presented, which leads to a more reliable evaluation for medical data as our experimental results illustrate.

4.4.1 Related Evaluation Methods

This subsection summarizes the quality evaluation methods for the synthetic medical records that are commonly used in prior investigations. We also discuss the limitations of these methods, which the critic method is designed to address.

Generally speaking, there are three types of methods. The first type measures whether synthetic data can support analytics over $p(y|x)$ as real data, where x and y are different sets of attributes of data. This is realized by comparing $\mathbb{E}_{x',y'} \log P(y|x,z=1)$ and $\mathbb{E}_{x',y'} \log P(y|x,z=0)$, given $P(x',y') = P(x,y)$, where the binary label z indicates if a record r is real or synthetic (e.g., $z=1$ when r is real). It is notable that the comparison is similar to the evaluation of the Kullback-Leibler (KL) divergence between $P(y|x,z=1)$ and $p(y|x,z=0)$. The DWP [37] measure introduced in Chapter 2, the DFA measure introduced in Chapter 3, and the TSTR method introduced by Esteban et al. [54] all can be regarded as measures of this type. The replication analysis on the bivariate multivariate feature correlations introduced by Azizi et al. [94] also falls into this category.

The second type compares real and synthetic data in the latent space. The method of this type is based on an assumption of an underlying generative process to compose the record distribution:

$$P(r) = \sum_{w} P(w)P(r|w),$$

where w is a set of latent factors that sufficiently characterize a record. Under this assumption, it is meaningful to assess whether $P(w)$ is the same in the real and synthetic data. The LSR and LTS measures introduced in Chapters 2 and 3, respectively, are in this category.

The clustering analysis method [92, 95] shares the same intuition in that each cluster center can be regarded as an unique latent factor.

The third type calculates the statistical distance between real and synthetic data, such as the maximum mean discrepancy (MMD) [96] and the Wasserstein distance [51]. This type of method is frequently used to evaluate the synthetic data generated by deep generative models (e.g., GANs) [97]. The univariate distribution of each categorical variable in the data (e.g., Bernoulli Success Probability) is also an example of this type.

The first two types of method partially sketch the quality of the synthetic data, but fail to provide a comprehensive assessment regarding the distributional discrepancy between real and synthetic data. Both $P(y|x,z=1) = P(y|x,z=0)$ and $P(w|z=1) = P(w|z=0)$ are necessary, but not sufficient conditions, of $P(r|z=1) = P(r|z=0)$. Thus, even perfect outcome (e.g., the first two equivalence above) ascertained by these methods cannot ensure that the synthetic data will exhibit high quality. The third type of method usually cannot provide an intuition into the data utility straightforwardly, and cannot be directly compared between semantically different datasets, thus, lack a certain level of interpretability and universality.

4.4.2 Evaluation Based on Discrimination

An application-agnostic evaluation of synthetic medical data should consider both the Type 1 and Type 2 errors. In this setting, the Type 1 error is a fidelity measure in that it characterizes the extent to which a model generates data (at a high probability) that are unlikely to be produced in the real world. By contrast, the Type 2 error is a diversity measure in that it characterizes the extent to which a model fails to generate data that are frequently observed in the real world. Both types of error are covered by the Jensen-Shannon (JS) divergence between real and synthetic data,

$$\mathrm{JS}(P||Q) = 1/2\mathrm{KL}(P||M) + 1/2\mathrm{KL}(Q||M)$$

where P and Q indicate the distributions of real and synthetic data, respectively, and $M = (P+Q)/2$. Specifically, the first component of the righthand side corresponds to the Type 1 error and the second component corresponds to the Type 2 error.

Meanwhile, an evaluation based on critic modeling can be interpreted as measuring the JS divergence between the real and synthetic data distributions. Consider a dataset D with

the same number of real and synthetic records, we can derive:

$$\mathrm{JS}(P||Q) \propto -\mathbb{E}_{r|z=1} \log P(z=1|r) - \mathbb{E}_{r|z=0} \log P(z=0|r) \propto -\mathbb{E}_{(z,r)\in D} \log P(z|r).$$

This means that if we have a quantitative critic $c(r)$ that represents the posterior $P(z=1|r)$, then the JS divergence is approximately equivalent to the error of using $c(r)$ to classify $r \in D$ as real or synthetic. As such, an evaluation based on critic modeling shares the benefits of measuring JS divergence, which covers both data fidelity and diversity. In practice, the critic is typically obtained by training a machine learning model, whose empirical error on a testing set is relied upon for evaluation.

It should be noted that the notion of a critic is also used in training a GAN. However, the critic model for a GAN (i.e., the discriminator) cannot be directly reused for record-level evaluation because of the fact that a GAN works only on episode production (due to the discrete nature of sequence synthesis). In other words, the critic model, which we elaborate upon in the following sections of this chapter, does not represent a component of a GAN.

4.4.3 Critic Implementation

Various prior investigations [91, 92, 93] train machine learning models through naive supervised learning for the critic. However, this practice often leads to an approximation of $P(z=1|r)$ with a very loose estimation of the lower bound of $\mathrm{JS}(P||Q)$. As a result, the quality of synthetic data is likely to be greatly overestimated.

Zellers and colleagues [98] provided a strategy for distinguishing human-written from machine-generated texts. Specifically, well-performed critic requires the same inductive bias (i.e., a set of explicit or implicit assumptions implied by the learning algorithm) as the model to create synthetic data. In doing so, the critic is less likely to settle on a local optima in its training process. In practice, this strategy can be realized through transfer learning whereby the critic model is built on top of the feedback of the trained generative model on the data to discriminate (i.e., the features in both real and synthetic instances extracted from the generative model). Inspired by this strategy, we further propose a finetuning strategy wherein the critic is built by refining the trained generative model (i.e., $f()$) to be aware of the difference in the real and synthetic data distributions through a training process to resolve $P(z|r)$. We anticipate that the finetuning strategy will lead to a tighter lower bound of $\mathrm{JS}(P||Q)$, which would be more reliable for evaluation purposes. Figure 4.3 illustrates the training process of the critic model.

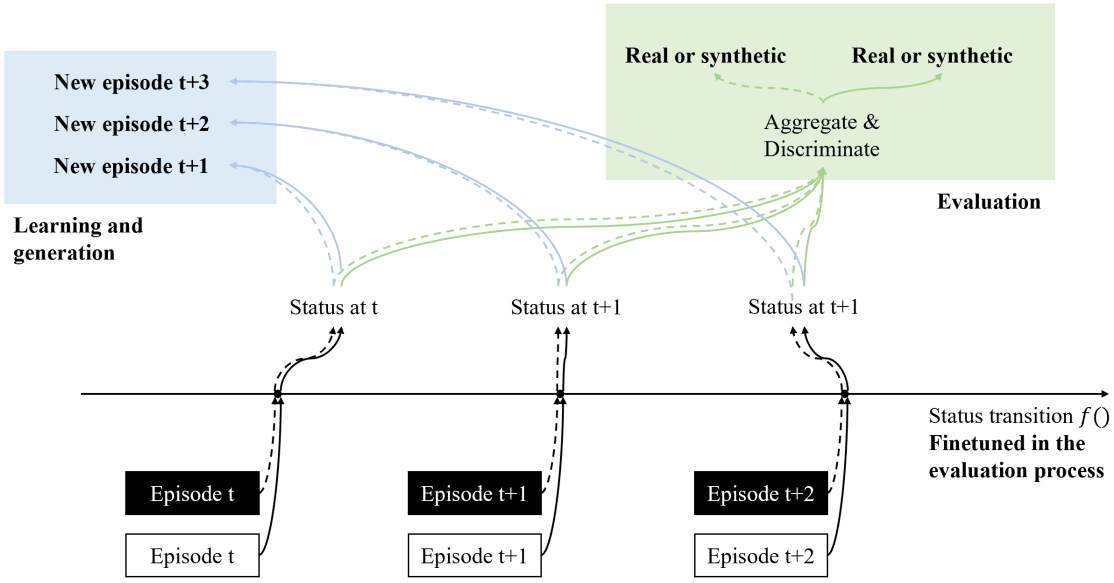


Figure 4.3: The process of developing a critic model development process. The white and black boxes represent real and synthetic episodes, respectively. The vertical dashed and solid arrows (except for the one corresponding to $f()$) represent calculation over real and synthetic data, respectively. The blue and green colors represent calculations performed in the learning and generation process and evaluation process, respectively; while the black color represents calculations performed in both processes.

It is further worth mentioning that the $c(r)$ obtained through this approach can support a fair estimation of $P(r)/Q(r)$ for rejection sampling as

$$\frac{P(r)}{Q(r)} \approx \frac{c(r)}{1-c(r)} \tag{4.2}$$

4.5 Experiments

4.5.1 Materials

In this chapter, we aim to derive findings that are generalizable in different settings of real world scenarios where the target data are represented differently. We conducted an empirical analysis with EMR data from two distinct resources. The first dataset is derived from de-identified EMRs from Vanderbilt University Medical Center (VUMC) [82]. For this dataset, we follow the approach outlined in Chapter 3, such that we focus on diagnoses that have been mapped into Phenome-wide Association Studies (PheWAS) codes [79, 99] and remove the codes with low frequency (i.e., smaller than 0.1%) in the set of records to ensure a feature space that has sufficient evidence to support the simulation objective. The

second dataset is derived from the publicly available Registered Tier of the NIH-sponsored *All of Us* Research Program. In this dataset, we focus on both diagnosis and procedure billing codes mapped into their Clinical Classifications Software (CCS) form[1]. It is worth mentioning that the datasets are not constrained to the same coding system.

We further refine the datasets in several ways. First, we define a clinical episode as a group of diagnoses and procedures that were documented on the same day. For each patient, we partition the sequence of episodes if there is a more than two years interval between consecutive episodes and retain the final partition only so that each patient record contributes only one sequence of events that are likely to be related to one another. Second, this chapter focuses on the simulation of records that preserve a meaningful longitudinal structure. As such, we retain only the patients who exhibit a relatively large number of episodes. Specifically, we chose a threshold of 25 for the *All of Us* data, and 10 for the VUMC data, considering that the VUMC data exhibits a substantially lower number of episodes per patient. Additionally, for computational efficiency, we limit the total number of episodes per patient to their most recent 200, and randomly downsample the VUMC dataset to the same number of records as the *All of Us* dataset. Table 4.1 provides summary statistics for the resulting datasets. Each dataset contains medical records from 59,617 patients. The VUMC dataset covers 1276 PheWAS codes and the *All of Us* dataset covers 526 CCS codes.

Table 4.1: A summary of the datasets used in this Chapter.

Dataset	**Patients**	**Episodes**	**Episodes Per Patient (Mean; 25th, 50th, 75th quartile)**	**Event Concepts**
VUMC	59,617	2,116,628	35.5; 14, 22, 43	PheWAS: 1276
All of Us	59,617	4,731,317	79.4; 38, 59, 105	CCS diagnosis: 282 CCS procedure: 244

4.5.2 Experimental Design

We perform longitudinal simulation with two implementations that utilize a combination of a GAN with a recurrent neural network. The first, corresponding to the SynTEG framework

[1] https://www.hcup-us.ahrq.gov/tools$_software.jsp$

introduced in Chapter 3, is used as a baseline for comparison. The second, referred to as the CFR implementation, is built upon the baseline and additionally incorporates condition fuzzing and regularization. Next, for each trained model, we generate a synthetic dataset with the same size as the real dataset. We also construct a third synthetic dataset using the CFR implementation with rejection sampling to study the influence of generation strategy on the quality of resulting datasets. Specifically, the rejection sampling is achieved through critical modeling mentioned in the previous section.

We evaluate the quality of the resulting synthetic datasets as follows. First, we randomly split each real and synthetic dataset into a training set and a testing set according to a 4:1 ratio. Second, we merge the two training sets to train three critics to distinguish between the real and synthetic data. Each of the discriminators corresponds to one of the following strategies:

- **Naive:** Train the critic using randomly initialized parameters;
- **Transfer learning:** Train the critic based on representations of real and synthetic records derived from the generative model;
- **Finetuning:** Use the trained generative model (i.e., $f()$) to initialize the critic's parameters.

It should be noted that for the critic model, we use the same architecture of $f()$ as the generative model, followed by an additional classifier. Specifically, the classifier is composed of an attention layer and multiple fully connected layers with layer normalization and residual connection between each layer. The status maintained by $f()$ at all timesteps are collectively used as the input to the additional classifier. Under the transfer learning and finetune strategies for critic modeling, the parameters of $f()$ are shared between generative and critic models at the beginning of the training.

Third, we report the performance of the critics on the merged testing sets in terms of discrimination accuracy and area under the receiver operating characteristic curve (AUROC). Since the critics based on transfer learning and finetuning can utilize either of the trained generative models (i.e., baseline or CFR), we report the best discrimination performance to conduct a fair comparison between strategies.

In addition, we also performed an experiment to investigate the drift problem and demonstrate how the new techniques mitigate drift. In this experiment, we extract episodes

from the records of each synthetic dataset and train critics to distinguish between real and synthetic episodes instead of complete records. We then order the extracted episodes by their relative positions in complete records. The ordered episode set is split into 20 equal-sized consecutive partitions. We report on the averaged discrimination performance per partition.

4.5.3 Results

Here, we present the discrimination performance between the real and synthetic data. To orient the reader, it should be recognized that a lower discrimination performance indicates better synthetic data quality. We observed that our results demonstrate the same patterns for both AUROC and accuracy on both datasets, such that we refrain from specifying the dataset and performance measure in the following presentation.

Critic Modeling

We report the performance of the various critic modeling strategies in Tables 4.2 and 4.3, where each row corresponds to a distinct strategy. It can be seen that the critics that leverage generative models perform substantially better than the critics trained with the naive strategy. Moreover, the critics trained with the finetuning strategy perform better than the transfer learning strategy. Based on these findings, it appears that finetuning achieves a relatively accurate estimation of the difference between real and synthetic data and, thus, we selected it for further evaluation purposes.

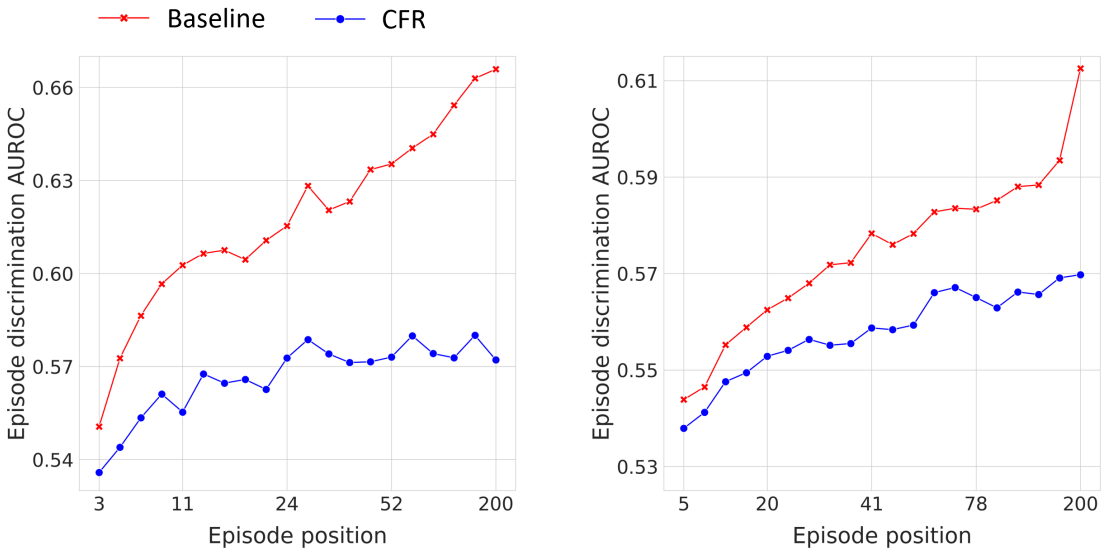
Learning and Generation

We assess the performance of model training and data generation methods by comparing the columns in Tables 4.2 and 4.3, respectively. Specifically, we focus on the rows corresponding to the finetuning strategy (shown in bold font). It can be seen that the CFR implementation produces synthetic data that are more difficult to distinguish from real data than the baseline. This suggests that the refined learning process induces synthetic records with higher statistical similarity to real records. Notably, rejection sampling further improves the quality of the resulting synthetic datasets without changing the generative model. For the VUMC data, it can be seen that incorporating both approaches into the synthesis pipeline induces a 17.0% and 16.6% quality improvement in terms of discrimination accuracy and AUROC, respectively. Similar performance is achieved for the *All of Us* data with improve-

ments of 14.1% and 12.2% in discrimination accuracy and AUROC, respectively.

Longitudinal Drift

Next, we illustrate the drift problem in longitudinal simulation. Figure 4.4 shows the episode discrimination AUROC as a function of the position of an episode in a generated record. There are two notable findings. First, the discrimination AUROC for both the baseline and the CFR implementation are substantially larger than 0.5. Also, there is a positive correlation between the discrimination AUROC and the episode position. This observation is strong evidence of the drift problem – the later an episode is produced in a sequence, the more likely the real and synthetic can be distinguished from one another. Second, this positive correlation is significantly weaker for the CFR implementation than the baseline. Specifically, the slope of a linear regression between the x-axis and y-axis values in Figure 4.4 is 0.0048 (Baseline) vs. 0.0016 (CFR) for the VUMC data and 0.0027 (Baseline) vs. 0.0014 (CFR) for the *All of Us* data, which illustrates that the proposed condition fuzzing and regularization technique is effective at mitigating error accumulation in the stepwise generation process.



(a) VUMC

(b) *All of Us*

Figure 4.4: Discrimination AUROC as a function of episode position.

4.6 Discussion

This chapter shows how to improve the quality of synthetic longitudinal health data, but there are several open issues that remain.

First, we wish to point out that this chapter focuses on building an automated pipeline for synthetic health data generation and evaluation. As such, we demonstrated an evaluation process from the perspective of computational indistinguishability. Yet we did not involve clinically knowledgeable experts in the assessment of the new learning method. While we have applied objective criteria to assess a synthetic record's authenticity, it is possible that human experts may leverage their intuition and specialized knowledge or intuition to appraise the extent to which the synthetic data is representative of real patient trajectories. Still, our evaluation is not subject to within-expert and between-expert variabilities [9], and thus, benefits large-scale studies. Nevertheless, more research is needed to investigate the consistency between machine and human evaluation and the integration of clinical experts' domain knowledge into the machine evaluation.

Second, there is an opportunity to further improve the simulation pipeline with the feedback mechanism. At the very beginning of a simulation process, it should be decided what attributes of real data should be preserved (i.e., the completeness of the resulting synthetic data with real data as the reference). This decision is essential in that data completeness is likely realized at the cost of a certain level of data quality. For example, correlations in the attributes of high-dimensional data are difficult to be completely captured. By contrast, synthetic records with fewer episodes or less detailed clinical information are more likely to be generated with a small statistical bias to the real data. As such, to achieve optimal utility of synthetic data, analysis is needed to determine how best to balance the completeness and quality in determining the objective of simulation. We believe that the feedback mechanism can be extended to assist in the decision process.

Lastly, we acknowledge that there are several limitations of this work that should be considered in future research. First, although we orient the generative modeling framework to mitigate error compounding in the generation process, the drifting problem is not completely resolved. Synthetic records with a large number of episodes are not guaranteed to have the exactly same distribution as the real records. Second, the scalability of our findings needs further evaluation. We conduct experiments with data projected into the curated PheWAS or CCS coding space. However, it is unknown how the proposed methods would perform given data with a larger or uncurated feature space and, thus, higher sparsity (e.g.,

medical records encoded with the ICD-10 system, which is composed of almost 68,000 codes, or medical records with the PheWAS codes with extremely prevalence in the population preserved), or data of various types, including the combination of categorical (e.g., procedure, diagnosis) and continuous features (e.g., laboratory test results, vital signs, and medication dosage). Further, it remains to investigate whether the findings in this chapter could generalize to a broader scope of EMR datasets.

4.7 Conclusion

This chapter shows that the longitudinal medical record simulation paradigm introduced in Chapter 3 leads to synthetic data that drifts from the distribution of real data over time. Our findings show that this occurs because of self-reinforced errors in episode generation. We further show that the problem of drift can, to large extent, be mitigated by incorporating the conditional fuzzing and regularization methods into model training process and a feedback mechanism into the generation process. In addition, we introduce a strategy for critic modeling that leads to more reliable assessment of the quality of synthetic data. The experiments conducted on EMRs from two real clinical data resources demonstrate the effectiveness of our approaches, but we acknowledge that our assessment relies solely on quantitative assessments and neglects feedback from clinically knowledgeable experts.

Table 4.2: Discrimination performance for the VUMC dataset (n = 59,617). CFR = condition fuzzing and regularization; RS = rejection sampling.

Discrimination Strategy	**Simulation and Generation Strategy**					
	Baseline		**Baseline + CFR**		**Baseline + CFR + RS**	
	Accuracy	**AUROC**	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**
Naive	0.605	0.655	0.566	0.591	0.538	0.556
Transfer learning	0.699	0.767	0.698	0.772	0.650	0.708
Finetune	**0.830**	**0.909**	**0.752**	**0.838**	**0.689**	**0.758**

Table 4.3: Discrimination performance for the *All of Us* dataset (n = 59,617). CFR = condition fuzzing and regularization; RS = rejection sampling.

Discrimination Strategy	**Simulation and Generation Strategy**					
	Baseline		**Baseline + CFR**		**Baseline + CFR + RS**	
	Accuracy	**AUROC**	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**
Naive	0.637	0.692	0.600	0.643	0.539	0.559
Transfer learning	0.788	0.874	0.724	0.802	0.674	0.741
Finetune	**0.844**	**0.918**	**0.777**	**0.858**	**0.725**	**0.806**

CHAPTER 5

Partially Synthetic Medical Data Simulation

5.1 Introduction

Chapters 3 and 4 introduced synthetic data generated in a full synthesis manner. In this chapter, we consider another type of synthetic data, namely partially synthetic data [100, 101]. In the full synthesis setting, a generative model is learned to simulate the real data distribution, and synthetic data are then sampled from this distribution. By contrast, in the partial synthesis setting, a transformation function is learned to map each real record into a synthetic record through feature perturbation. The resulting synthetic data can be considered as sampled from the vicinal distribution of the real dataset (i.e., the distribution of the vicinities of the instances in the semantic space that defines them). It is worth mentioning that the idea of partially synthetic data is conceptually related to data augmentation. Data augmentation is commonly deemed as a regularization technique for improving the generalizability of machine learning models. It implies that each instance in the dataset can be approximated as an interpolation of its neighbors in the semantic space, which collectively constructs the instance's vicinal distribution. Machine learning models trained on an augmented dataset frequently perform as well or even better than models trained exclusively on the original dataset [102]. As such, it is expected that the synthetic data created through a data augmentation process will maintain a similar statistical property to the real data, thus retaining a high degree of utility.

However, such a superiority of utility could be attained at the expense of maintaining a one-to-one relationship between real individuals and synthetic records. Intuitively, sharing partially synthetic data should not raise privacy disclosure risks because the linkage retained between the synthetic records and the real data upon which it is based is implicit (as a result of feature perturbation). However, such risks might be exposed through state-of-the-art machine learning frameworks. Given this issue, it is in the best interest of a data holder to perform a privacy risk assessment prior to sharing any partially synthetic medical data. And, based on the analysis, they can then decide if it is appropriate to share the synthesized data.

Among a variety of privacy attacks, membership inference attacks have received a significant amount of attention over the past several years [21, 103, 104, 105, 106]. However,

such an attack in the context of sharing synthetic data is quite different from the traditional scenario of targeting machine learning models. Notably, potential adversaries can only gain access to a synthetic dataset of a certain number of records, as opposed to the trained model that generates synthetic data (A summarized illustration of the comparison is provided in Figure 5.1). Therefore, most research findings regarding membership inference in the traditional scenario cannot be applied to the synthetic data scenario. Several approaches have been developed to simulate the membership inference attack against synthetic data [13, 14, 15, 16, 17]; however, they are limited in several ways when being used for risk evaluation. First, many of these methods rely on assumptions about specific deep learning frameworks, such as generative adversarial networks (GANs [36]) and variational autoencoders (VAEs [56]). Second, these methods tend to assume that the synthetic data has a well-defined structure, such as those encountered in visually inspectable images. Yet data about one's health are often longitudinal, which are not structured in a perfectly aligned manner. Specifically, each patient's record includes multiple episodes of care, which are irregularly distributed across a timeline. Additionally, the number of such episodes and events can vary across patient records. As a result, methodology for medical data synthesis is increasingly realized in a manner that episodes in the same record are sequentially generated (in a pre-defined order) based on their antecedents rather than being generated altogether at one time by a GAN or VAE [29, 93]. As a consequence, the assumptions inherent in current approaches are not always valid, rendering them less useful.

In this chapter, we first introduce a method for partially synthetic medical record generation. Then we introduce a framework for effective membership inference against partially synthetic data to support privacy risk assessment, leveraging the principles of contrastive representation learning. With this framework, we aim to determine the upper bound of risk brought by an adversary who invokes an optimal strategy. We assess the effectiveness of our method through systematic experiments with longitudinally structured diagnosis and procedure code data derived from two large clinical datasets: one from Vanderbilt University Medical Center (VUMC), and the other from the NIH-sponsored *All of Us* Research Program [81]. We empirically demonstrate that partially synthetic data has potential to achieve higher quality than fully synthetic data, but is vulnerable to applications of the proposed membership inference framework.

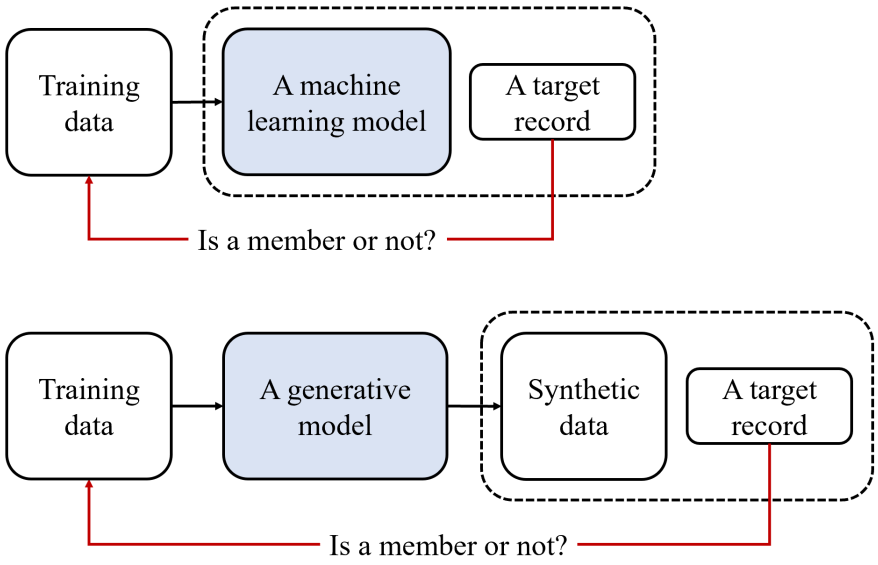


Figure 5.1: An illustration of membership inference against a machine learning model (upper), and against synthetic data (lower). The dashed box indicates the resource that can be used for inference. The shaded box represents the machine learning models.

5.2 Partially Synthetic Data Generation

This section introduces a method for partially synthetic medical record generation using a variation of the multiple imputation strategy [8, 107], which is implemented by iteratively replacing episodes of a real individual with simulated synthetic episodes. This method is also inspired from the work of Kobayasi and Lewis and colleagues [108, 109] on data augmentation in the natural language processing domain.

We first organize each record r as a sequence of consecutive episodes (e.g., outpatient visit or inpatient hospital stay), denoted as $e_1, e_2, ..., e_N$, where e_t corresponds to the t^{th} episode and N is the total number of episodes, which can vary across records.

We represent r from each e_t's view as (e_t^-, e_t, e_t^+), where e_t^- and e_t^+ represent the previous and following episodes of e_t in the sequence, respectively. We then learn a transformation to map each (e_t^-, e_t^+) to a fixed-length vector representation h_t through a pre-training task. Next, we train a conditional generative model to simulate each proxy episode $\widetilde{e_t}$ of e_t given h_t. For brevity, we represent all of the steps of the process as $\widetilde{e_t} \sim G(e_t^-, e_t^+)$.

After training, we obtain a proxy $\widetilde{r}$ for each r through the process described by Algorithm 5.1. Briefly, this is accomplished by iteratively replacing e_t with $\widetilde{e_t}$ for each t in a random shuffling of $(1, 2, ..., N)$, where the number of iterations n is determined through empirical calibration.

Algorithm 5.1: Partially Synthetic medical record Generation

Input : Trained model G;
Record $(e_1, e_2, \ldots, e_N)$;
Number of iterations n

Output: Proxy $\widetilde{r}$

$\widetilde{e_t^0} \leftarrow v_t$;

for $k \leftarrow 0$ **to** n **do**

  Randomly shuffle the sequence $(1, 2, \ldots, N)$;

  Let $o(t)$ represent the order of t in the shuffled sequence;

  for $t \leftarrow 1$ **to** l **do**

    $\widetilde{e_{o(t)}^{k+1}} = G(\widetilde{e_1^{a_1}}, \ldots \widetilde{e_{o(t)-1}^{a_{o(t)-1}}}, \widetilde{e_{o(t)+1}^{a_{o(t)+1}}}, \widetilde{e_N^{a_N}})$,

    where

$$a(i) = \begin{cases} k & \text{if } o(i) \leq o(t) \\ k+1 & o.w. \end{cases}$$

  end

end

$\widetilde{r} = (\widetilde{e_1^n}, \widetilde{e_2^n}, \ldots, \widetilde{e_N^n})$

5.3 Membership Inference Against Partially Synthetic Data

5.3.1 Preliminaries

In this section, we describe the data holder's perspective regarding how an adversary conducts membership inference against synthetic data.

We begin by providing context for the adversarial setting. The synthetic medical data generation process aims to produce data that serves as a substitute for real patient data. Since the model involved in the synthesis process, referred to as the target model, does not need to be shared, in this dissertation, we assume that membership inference functions in a black-box setting. Thus, the adversary is provided access to the synthetic dataset only and not the target model. Given this setting, we define membership inference against synthetic data as follows.

We assume the adversary possesses full knowledge for a collection of records $X = \{r_1, r_2, \ldots, r_n\}$, referred to as the known target dataset. X is partitioned into two mutually exclusive datasets. The first dataset, X_{source}, is involved in the synthesis process, while the second dataset, $X_{holdout}$ is not. The membership status of each record is maintained in a set of Boolean values $\{m_1, m_2, \ldots, m_n\}$, such that $m_i = 1$ if $r_i \in X_{source}$ and $m_i = 0$ if

$r_i \in X_{holdout}$. The adversary's model is thus defined as:

$$\mathscr{M} : (r_i, X_{syn}) \rightarrow \{0, 1\},$$

where X_{syn} corresponds to the synthetic dataset.

The adversary's goal is to resolve a maximal subset of X, $X^{'}$, such that

$$\frac{1}{|X^{'}|} \sum_{r_i \in X^{'}} \mathscr{M}(r_i, X_{syn}) \cdot m_i > p,$$

where p is a pre-defined threshold, representing the precision (which equals to $1 - false\ positive\ rate$) of the adversary's inference committed against X.

5.3.2 Related research

To date, there have been several investigations into the feasibility of a generic approach to membership inference through models trained in an unsupervised manner - particularly for generative models. The typical approach creates local copies of the generative model, G, with parameter $\theta(X_{syn})$ using the synthetic data. This model is then applied to assign each known record with a likelihood that is either generated or accepted by the local copies [13, 14]. $\mathscr{M}$ is typically formulated as

$$\text{sign}[P(r_i | G_{\theta(X_{syn})}) > t],$$

where sign$[\cdot]$ is a signum function that returns either 0 or 1 and t is a pre-defined threshold.

In the attack formulated by Chen and colleagues [15], it is assumed that, if the synthetic data pose a membership inference risk for a known record, then it must be possible to observe that the generative model overfits the record (i.e., the model assigns a higher likelihood to records that are in the training set than those that are not):

$$P(m_i = 1 | r_i, X_{syn}) \propto P(r_i | G_{\theta(X_{syn})}).$$

However, this formulation requires an explicit density function from the generative model, which is not always available. Chen et al. [15], as well as Bilprecht et al. [16], thus propose a more generic membership inference framework. They specifically utilize the property that a membership inference risk can be observed when the synthetic data demonstrate a certain

level of similarity to a target record:

$$P(m_i = 1|r_i, X_{syn}) \propto L(r_i, X_{syn}),$$

where $L(\cdot,\cdot)$ denotes a general notion of the similarity between r_i and X_{syn}. Yet, this approach is hindered in practice because it either 1) relies only upon a simple non-parameterized metric for $L(\cdot,\cdot)$ [16], rendering the approach insufficient for data with complex structure and high dimensionality, or 2) relies on specific assumptions about the target generative model [15].

Recent advancements in representation learning, however, provide an opportunity to alleviate both problems by defining the distance between the latent representations of the records. Typically, the approaches designed to support this endeavor fall into one of two groups: generative or contrastive. The former simulates the data distribution and then derives a latent form that represents the semantic features as decodable factors [110, 111]. Training generative models, which requires simulation of the data in a lossless manner, is often computationally intensive and requires excessively large quantities of data, particularly when in a sequential form (e.g., generative models for natural language: T5 [112] required 34 billion tokens to train 11 billion parameters, while GPT-3 [113] required 300 billion tokens to train 175 billion parameters). Yet, from the perspective of membership inference, acquiring a lossless representation might be unnecessary. A representation composed of a limited number of features may be sufficient to recognize a unique record instance. As such, contrastive learning [114, 115], is well-aligned with the objective of membership inference. Still, one of the challenges in applying a contrastive learning approach is how to design the augmentations (i.e., slightly modified copies of already existing records) needed for training that will maximally promote membership inference as a downstream task of representation learning.

5.3.3 Membership Inference Algorithm

In this section, we introduce a two-step process to perform membership inference. In the first step, we learn the representations of records. In the second step, we apply a measure to calculate the distance between the representation of a known record and the representation of the synthetic records, which can be effectively exploited for membership inference. These two steps collectively represent an implementation of function $L()$ described in section 5.3.2. Additionally, $L()$ followed by a heuristic algorithm to perform inference based

on $L()$ can be regarded as the adversarial model $\mathscr{M}()$. A flow chart of the membership inference process is shown in Figure 5.2. We refer to this method as *CRL-proxy*.

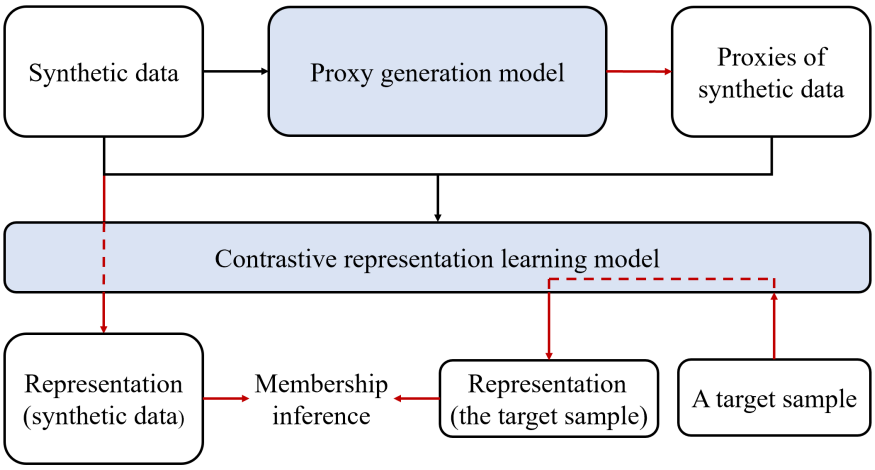


Figure 5.2: A procedural depiction of the membership inference framework. The black arrows indicate the training process while the red arrows indicate inference using the trained models.

5.3.3.1 Training

We leverage contrastive learning to obtain record representations that can be used in downstream membership inference. First, given the synthetic dataset X_{syn}, we reuse the method for partially synthetic data generation to create a proxy dataset X_{proxy} of X_{syn}. Specifically, for each $r \in X_{syn}$, we apply Algorithm 5.1 to generate $\widetilde{r}$ as the proxy for r. Then, we use an encoder to extract each synthetic record's, as well as each proxy's, fixed-length vector representation v_r and $v_{\widetilde{r}}$. The contrastive training's objective is to minimize the following function:

$$-\mathbb{E}_r \log \frac{\exp(\text{sim}(v_{\widetilde{r}}, v_r))}{\sum_{r_c \in X_{proxy}} \exp(\text{sim}(v_{r_c}, v_r))},$$

where $\text{sim}(u,v) = \frac{1}{\varepsilon} \frac{uv^T}{\|u\|\|v\|}$ is the scaled dot product between the L2 normalized u and v, and $\varepsilon > 0$ is an adjustable hyper-parameter. This objective is precisely the NT-Xent loss as proposed by Chen and colleagues [88] and is equivalent to the infoNCE loss [85].

5.3.3.2 Inference

In this subsection, we introduce an algorithm to infer the membership status of a known record. For each $r_i \in X$, if $L(r_i, X_{syn})$ is greater than a certain threshold τ, we assert X_{syn} retains information of r_i and further claim r_i is in the source set to generate synthetic data. Given v_{r_i} and $\{v_r | r \in X_{syn}\}$, we consider the following heuristics to calculate $L(r_i, X_{syn})$

$$L(r_i, X_{syn}) = \max_{r \in X_{syn}} \text{sim}(v_{r_i}, v_r).$$

This function corresponds to the largest similarity between r_i and all records in X_{syn}.

5.4 Experimental Design and Result

To investigate the performance of the partially synthetic data generation and membership inference methodology, we performed an empirical analysis with the same data introduced in Chapter 4, which are derived from two distinct electronic medical record (EMR) resources. The first dataset corresponds to de-identified data from VUMC. The second dataset corresponds to the publicly available Registered Tier data from the NIH-sponsored *All of Us* Research Program. We refer the readership to Chapter 4 for specifications of the datasets.

We perform longitudinal simulation with two implementations. The first, corresponding to the full synthesis method introduced in Chapter 4, is used as a baseline for comparison. The second is built upon the partial synthesis method introduced in Section 5.2. Next, for each trained model, we generate a synthetic dataset with the same size as the real dataset.

We evaluate the quality of the resulting synthetic datasets with the discrimination method introduced in Chapter 4, which is to train critic models to distinguish between the real and synthetic data. Specifically, we use the finetune strategy to train critics as it achieves a relatively accurate estimation of the difference between real and synthetic data.

5.4.1 Risk Assessment

Instead of articulating the risk of membership inference in terms of adverse consequences, we frame it with quantitative values by which the data holder could make decision of whether or not to share the data. This is achieved by providing a topology between the compromised proportion of the population and the attack's precision.

We randomly split each of *All of Us* and VUMC datasets into a source set X_{source}, holdout set $X_{holdout}$. The source set is applied to train the generative model, from which a synthetic set X_{syn} of the same size is generated. We set the size of X_{syn} to be the same as X_{source} given that the synthetic data is meant to be a substitution of the real data. We define $X_{source} \cup X_{holdout}$ as the known target set.

Next, we perform proxy generation and contrastive training on the synthetic set. We split the known target set into 10 partitions of the same size, according to the number of episodes associated with each record. It should be noted that the partitioning is only performed across records but not within a record. The intuition behind this step is to investigate how the number of episodes in a record (i.e., the amount of information provided by the record) influences the precision of membership inference. We apply the inference algorithm (in section 5.3.3.2) to each partition.

We assume the adversary performs an attack on each partition separately for an optimal attack performance. To illustrate the risk, we select the top 20%, 40%, 60%, 80%, and 100% of the records in the known target set with the highest risk score $L()$ (see section 5.3.3.2 for the definition) and calculate the proportion of the selected records that are in the source set. In doing so, we obtain a topological depiction of the relationship between the percentage of individuals targeted and the attack's precision. For instance, a precision of 1 indicates that all of the selected records are correctly inferred as members of the source data, while a precision that is no greater than 0.5 indicates that the adversary is no more successful than a random guess (due to the fact that 50% of the members of the known target set are in the source set).

5.4.2 Membership Inference with Incomplete Knowledge

In the most simple attack scenario for an adversary, the complete medical record of a target individual is available. However, this is an extreme scenario that is unlikely to happen in the real world. It is possible that the adversary only has access to incomplete knowledge of the target individual. To provide a comprehensive analysis of the risk of membership attack against partially synthetic medical data, we also consider the scenario where the adversary possesses incomplete medical records of target individuals. Specifically, we perform experiments in three different settings:

Binary profile: the adversary has access to the profile of the target individual, in which the data only indicate which medical events occurred to the target individual, but not the

temporal trajectory of the events.

Count profile: the adversary has access to the number of times that each diagnosis or procedure was made over the target individual's medical history.

Longitudinal record snippet: the adversary has access to a snippet of the longitudinal record, in which the data cover 10 consecutive episodes of each target individual.

5.4.3 Baseline Method for Membership Inference

To assess how well the proposed *CRL-proxy* performs, we compare with an alternative method for membership inference. The following provides a summary of the baseline model.

The baseline is based on a pretext task performed on the synthetic data to detect the target model's overfitting to the source data. We use the synthetic data to perform masked modeling (which is based on masked language modeling [116]), with which we calculate the approximated likelihood of each record in the known target dataset, conditioned on the synthetic data:

$$\mathbb{E}_t \log P(e_t | e_t^-, e_t^+; X_{syn}).$$

We claim that a record is in the source data when its likelihood is greater than a predefined threshold. This baseline share the same intuition with the membership inference method against fully synthetic medical record introduced in Chapter 3.

5.4.4 Results

5.4.4.1 Quality Evaluation

The experimental results with respect to discrimination performance between real and synthetic data is shown in Table 5.1. It should be noted that a lower discrimination performance indicates better synthetic data quality. We observe that the partial synthesis method produces synthetic data that are more difficult to distinguish from real data than the full synthesis method. For the VUMC data, partially synthetic data demonstrate a 4.5% and 4.3% higher quality than fully synthetic data, in terms of discrimination accuracy and AUROC, respectively. A similar result is observed for the *All of Us* data with a performance disparity of 7.3% and 6.1% in discrimination accuracy and AUROC, respectively. This finding suggests that the partial synthesis method enables synthetic records with higher statistical similarity to real records.

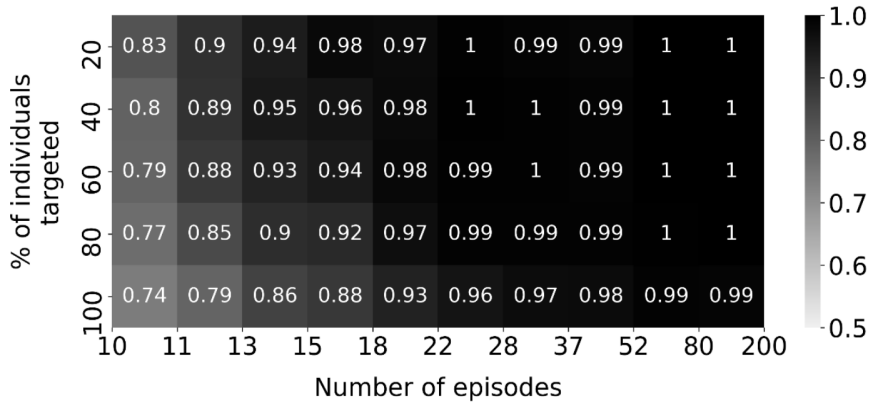
Table 5.1: Discrimination performance.

Dataset	Simulation method			
	Fully synthetic data		Partially synthetic data	
	Accuracy	AUROC	Accuracy	AUROC
VUMC	0.752	0.838	0.718	0.802
All of Us	0.777	0.858	0.720	0.806

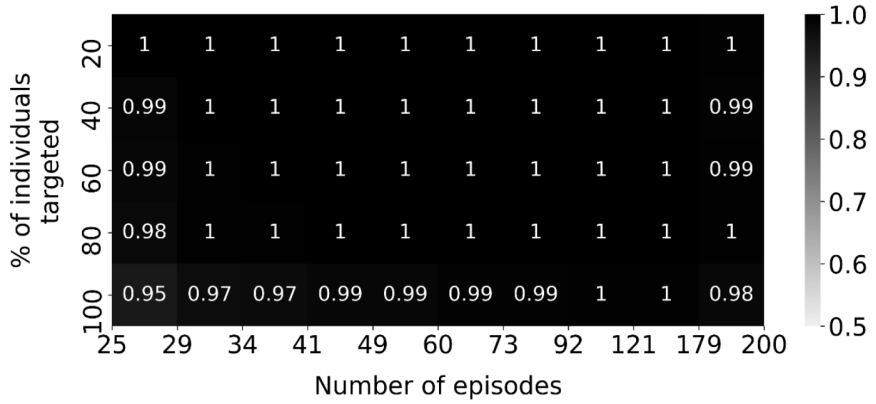
5.4.4.2 Membership Inference Against Partially Synthetic medical data

Figure 5.3 illustrates the membership inference risk based on the experiments with *CRL-proxy*. In this figure, the x-axis represents the number of episodes exhibited by a patient, while the y-axis represents a cumulative percentage of the patients the adversary conducts a membership inference attack upon. The color in the heatmap corresponds to the inference precision of the targeted subset. For example, in Figure 5.3a, the cell on the upper left corner represents patients with 10 to 11 episodes. When the top 20% of the targeted patients, ranked according to $L()$, are inferred as in the source set, the inference precision is 0.83.

There are several findings on partially synthetic data worth highlighting. First, it can be seen that the risk of membership inference is non-trivial for both VUMC and *All of Us* data. As shown in subfigures 5.3a, and 5.3b, multiple cells achieve a precision that is close to 1. The size of the subpopulation vulnerable to membership inference (presented as the percentage of the entire population under consideration) with precision beyond 0.9 is 100% for the *All of Us* data; and 78% for the VUMC data. Now, if the adversary reduces the precision threshold to 0.8, the size of the subpopulation is 100% for the *All of Us* data; and 92% for the VUMC data. Second, the precision for the subpopulation with a greater number of episodes is higher. This is expected and indicates that the records that are more informative are more vulnerable to membership inference. Third, there is a trade-off between the inference precision and the size of the compromised population. The adversary could obtain results with higher confidence by conducting inference on a subpopulation of smaller size.



(a) VUMC



(b) *All of Us*

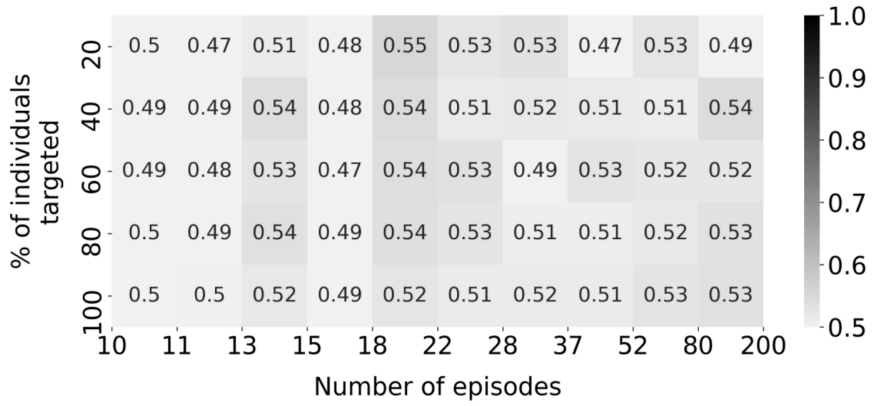
Figure 5.3: A summary of the membership inference risk against partially synthetic medical record. Each cell corresponds to a subset of all individuals who could be targeted by the adversary.

5.4.5 Comparison with the Baseline

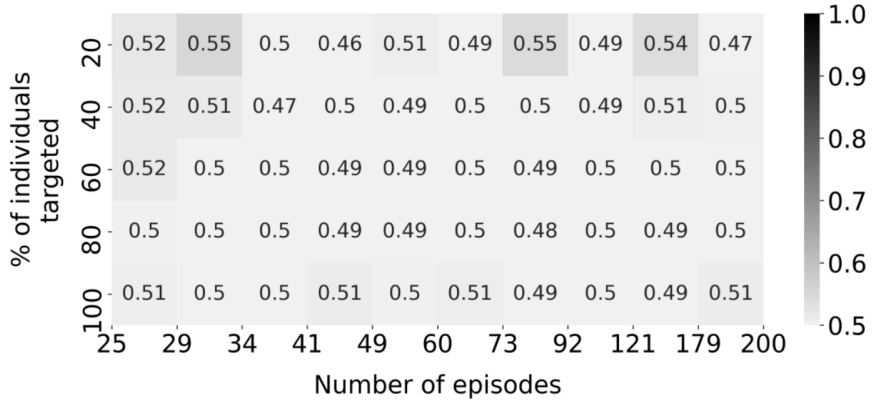
Figure 5.4 illustrates the result achieved by the baseline method. For both datasets, no subpopulation is vulnerable to membership inference risk with precision greater than 0.7. As such, *CRL-proxy* has the significantly better performance. This finding indicates that the contrastive learning method is more reliable for assessing the risk of membership inference.

5.4.6 Membership Inference with Incomplete Knowledge

This subsection illustrates the result when the adversary has incomplete knowledge of target individuals for membership inference attack. Figure 5.5 shows the result of the scenario



(c) VUMC



(d) *All of Us*

Figure 5.4: Membership inference results for the baseline

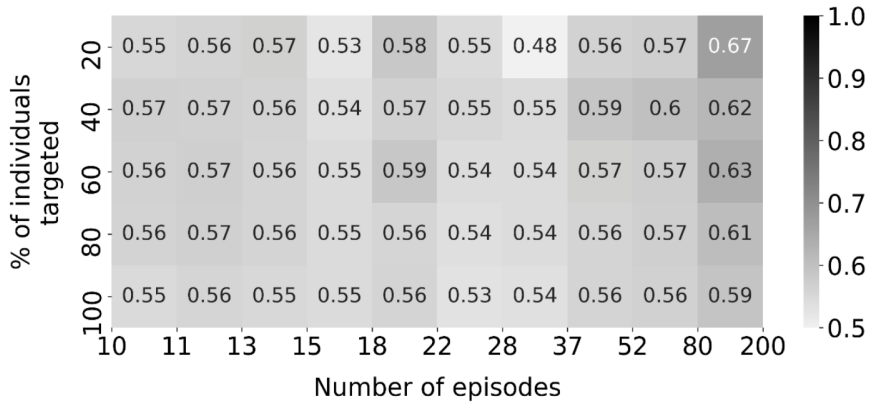
where the adversary has access to the binary profile of the target patients. For both partially synthetic *All of Us* and VUMC data, no subpopulation is vulnerable to membership inference with precision greater than 0.7.

Figure 5.6 shows the result of the scenario where the adversary has access to the count profile of the target patients. When the precision threshold was 0.8, *CRL-proxy* achieved inferential success on 16% of the population with partially synthetic *All of Us* data; while for VUMC data, no subpopulation is subject to membership inference. When the precision threshold was lowered to 0.7, *CRL-proxy* achieved inferential success on 40% and 24% of the population with partially synthetic *All of Us* and VUMC data, respectively.

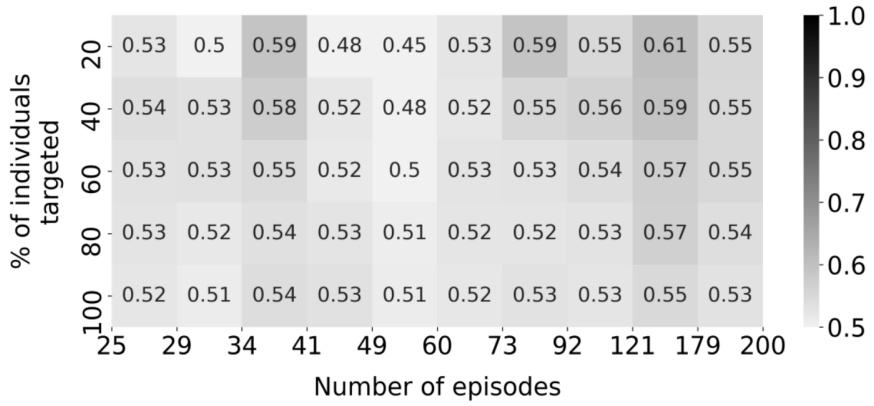
Figure 5.7 shows the result of the scenario where the adversary has access to the snippet of the longitudinal record of the target individuals. For both partially synthetic *All of Us*

and VUMC data, no subpopulation is vulnerable to membership inference with precision greater than 0.7.

It can be seen that the inference with incomplete knowledge substantially drops for both the VUMC and *All of Us* data. Particularly, in the scenarios if binary profile and longitudinal record snippet, partially synthetic data are only marginally susceptible and, in most cases, could be deemed sufficiently protected from membership inference.



(a) VUMC

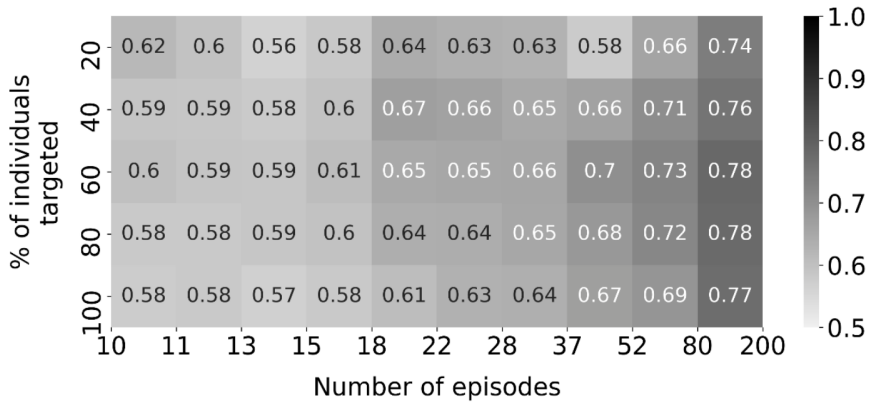


(b) *All of Us*

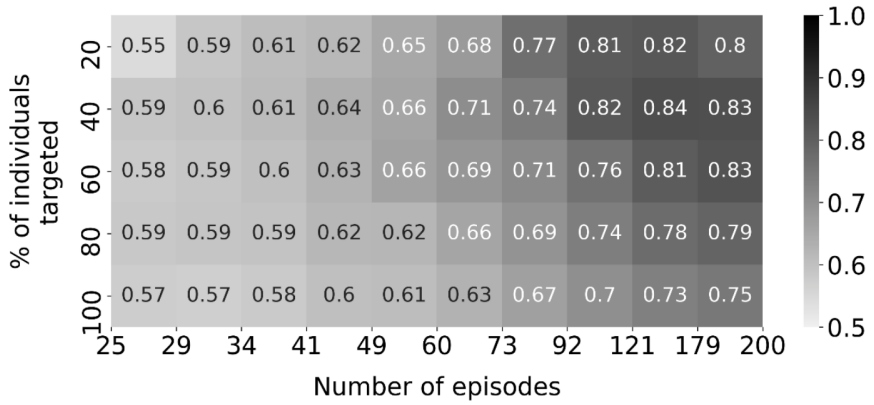
Figure 5.5: An illustration of membership inference risk against partially synthetic medical record, when the adversary has incomplete knowledge of target individuals (Binary profile).

5.5 Discussion and Conclusion

This chapter introduces a novel approach for partially synthetic medical record generation, which enables higher quality of resulting synthetic data than the full synthesis method. This



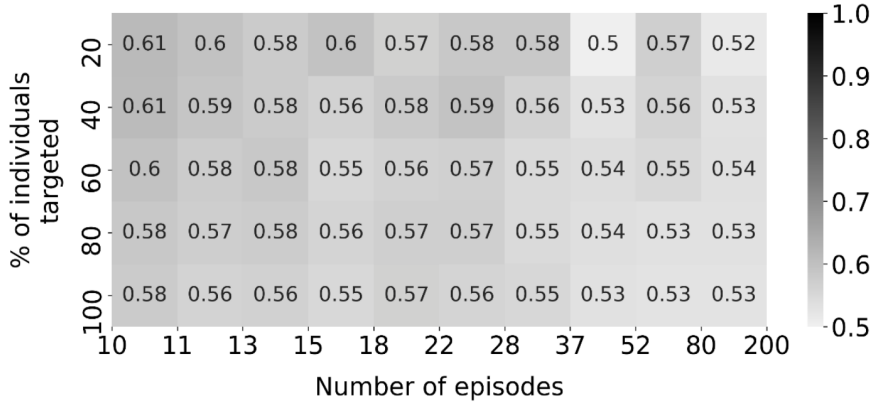
(a) VUMC



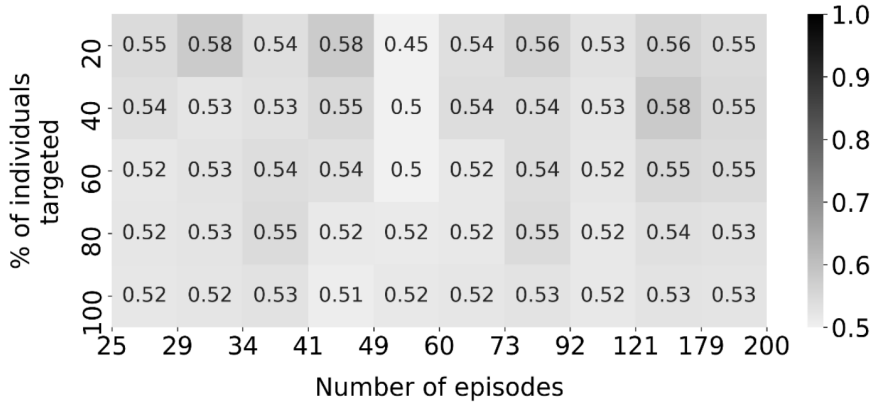
(b) *All of Us*

Figure 5.6: An illustration of membership inference risk against partially synthetic medical record, when the adversary has incomplete knowledge of target individuals (Count profile).

chapter also introduces a membership inference framework based on contrastive representation learning for privacy risk assessment on partially synthetic data. To the best of our knowledge, the proposed membership inference framework is the first approach that is not reliant on assumptions about the model involved in the synthetic data's generation process. The results of our experiments (with two distinct collections of real world medical data) show that partially synthetic data has the potential to retain a higher level of utility than fully synthetic data, but is susceptible to membership inference, especially when the adversary has complete knowledge of target individuals' medical record. Further, the method presented in this chapter could be applied as a preliminary privacy risk evaluation if any partially synthetic dataset is considered for release.



(a) VUMC



(b) *All of Us*

Figure 5.7: An illustration of membership inference risk against partially synthetic medical record, when the adversary has incomplete knowledge of target individuals (Longitudinal record snippet).

It should be noted that the membership inference model works from the data holder's perspective. When evaluating the risk reported in the results section, we use the knowledge that an adversary may not possess: 1) the membership distribution in the known target set (e.g., half of the records are in the source set), and 2) prior knowledge about the topological patterns of relationship between the percentage of individuals targeted and inference precision (e.g., the precision of *CRL-proxy* is higher for records with a larger number of episodes). Therefore, the reported results could lead data holders to slightly overestimate the level of specificity in their synthetic data. A tighter approximation of the risk can be obtained when a better understanding of what knowledge adversaries have access to and

what the behavioral limitations of such adversaries are is available.

We also wish to indicate that there are several opportunities for future refinements of this work. First, we relied upon the CCS and curated PheWAS coding system for EMR data, which has a more coarse feature space in comparison to other systems, e.g., the International Classification of Diseases (Tenth Revision). It is unknown how the experimental results observed in this chapter will hold in other coding systems or the original uncurated coding system. It also remains to be seen how the size and granularity of the feature space influence the quality of partially synthetic data and their corresponding risk of membership inference. Second, our investigation considered only a subset of the available types of medical data that are of interest for synthesis purposes. Specifically, we only considered well-structured diagnosis and procedure codes. However, it is important to investigate how our risk estimation methodology fares in the face of other types of medical concepts (e.g., laboratory test results and medications). Third, according to our experimental results, synthetic VUMC data are more resistant to the proposed attack than synthetic *All of Us* data. We suspect the primary reason for this difference is that the *All of Us* data benefits more episode per patient. However, there could be many other potential reasons for the difference, such as the fact that *All of Us* is composed of data from a wide variety of organizations' EMRs whereas the VUMC data is composed of a single organization. We expect further research with more datasets to investigate the driving factors for such differences and the generalizability of our findings.

CHAPTER 6

Conclusion

6.1 Summary of Results and Contributions

This dissertation focused on synthetic medical data simulation. The goal was to create a proof-of-concept data-driven pipeline for large-scale synthetic EMR simulation that is not dependent on clinical knowledge. We approached this goal from three aspects: 1) developing algorithms to model the distribution of EMR data, based on which synthetic data can be generated, 2) building an evaluation system to assess the utility of synthetic data, and 3) examining the privacy implications involved with sharing synthetic data. Our contribution to each of these aspects is summarized below.

First and foremost, we developed deep learning-based algorithms for modeling and generating both patient profiles and longitudinal medical records. We demonstrated that, when it came to patient profiles, the GAN-based generative framework could be improved in terms of either model architecture or training strategy. The improved framework generated synthetic data with comparable statistics and predictive capabilities to the original data upon which the simulation was based. With respect to longitudinal records, we presented a composite generative framework utilizing techniques including GANs, autoregressive models, and time point processes. The synthetic medical records generated by this framework demonstrated similar temporal dependencies and time-related statistics as real data. Further, we introduced a method for partially synthetic medical data generation based on a variation of the multiple imputation strategy. In comparison to fully synthetic data, partially synthetic data was found to more difficult to distinguish from real data. We proved the feasibility of generating synthetic data with a realistic feel with the proposed methodologies.

Second, we provided a family of methods for utility assessment that account for both the fidelity and diversity of synthetic data compared to real data. The evaluation was accomplished by examining the prediction capability and the distribution of latent features of synthetic data. Additionally, we proposed to evaluate synthetic data by discriminating it from real data. We developed an implementation that utilized knowledge of the generative model to aid in the discrimination, resulting in substantially more reliable evaluation results than training a classifier naively.

For the purpose of assessing privacy risks, we empirically demonstrated that both fully synthetic patient profiles and longitudinal records generated by the proposed methods were resistant to membership inference and attribute inference attacks. We further showed that partially synthetic longitudinal medical records were subject to membership inference using the state-of-the-art machine learning methods, when the adversary has complete knowledge of the medical record of target individuals. The above findings suggest fully synthetic data can be deemed sufficiently protected, while the use of partially synthetic data requires caution.

6.2 Discussion

In this section, we discuss the limitation of the current work, as well as open issues and future opportunities related to synthetic data simulation.

6.2.1 Limitation in Data Utility

In many medical data simulation scenarios, the subset of patients assigned a positive status for a particular diagnosis code may be extremely small due to the rarity of the related disease. As a result, only a limited number of training instances of the corresponding subpopulation are available to the model training process, which compose an insufficient representation of the subpopulation's data distribution. In this case, the current generative models based on machine learning may be incapable of extracting sufficiently strong signals from data to reflect unbiased knowledge of the target distribution. Recognizing this issue, we perform attribute curation in both the patient profile and longitudinal record simulation settings prior to training generative models to ensure that the attribute space is well-specified by the training data. This is accomplished by aggregating medical concepts into higher-level categories and removing the resulting categories with a low prevalence. For example, we group ICD-9 diagnosis billing codes by their first three digits and retain only 854 three-digit codes for the purpose of simulating patient profiles.

This practice enables us to provide a proof of concept of generating synthetic medical data with high statistical similarity to its real world counterpart. However, there is still room to improve the utility of synthetic data in general and in terms of fairness. From a broader perspective, the extent to which the attribute space of original data is preserved in synthetic data (which we refer to as the completeness of synthetic data) is critical because the underlying application of synthetic data (e.g., hypothesis formulation or testing) may require

the retention of specific attributes. As such, the inability of current modeling to accurately capture the distribution of phenotypes with a small population inevitably lead to data utility loss. From the fairness standpoint, synthetic data may be unable to represent the minority population with rare diseases in certain applications, as the minority population is either neglected in the simulation due to attribute curation, or is likely to be biasedly represented in the synthetic data because of the generative model failing to accurately capture its distribution. We believe that future efforts to improve synthetic medical data simulation should focus on the generative model's ability to elevate data completeness while maintaining data quality.

6.2.2 Differential Privacy in Synthetic Data Simulation

In this dissertation, we did not use statistical disclosure limitation models, including differential privacy (DP), in the design of generative algorithms for medical data simulation. Nonetheless, we are aware of the growing trend toward incorporating DP into synthetic data simulations in a broader scope of scenarios.

If properly realized in the data simulation process, DP could provide theoretically-guaranteed privacy protection to the individuals whose information is used to generate the synthetic data. Further, a general concept of privacy loss is inherent in differentially private algorithms. This concept could be used to track the upper bound of privacy risks associated with sharing synthetic data and formulate privacy policies governing the release of any synthetic dataset. Additionally, it is worth noting that techniques for incorporating DP into the training of deep learning-based generative models already exist. Abadi and colleagues [117] proposed the Differentially-Private Stochastic Gradient Descent (DP-SGD) algorithm for training differentially private deep learning models. Xie and colleagues [26] proposed differentially private generative adversarial networks based on DP-SGD, which were evaluated in the context of patient profile generation. Multiple prior works have also adopted DP in training autoregressive models to generate synthetic sequential data (e.g., natural language text) that is not subject to unintended memorization, a phenomenon in which the synthetic instances replicate fine details from a specific training sample.

However, DP is a double-edged sword in EMR simulation. The privacy guarantee,in terms of acceptable privacy loss, is achieved at a significant cost of the utility of the resulting synthetic data [26, 27]. This is because generative models aim at capturing the complete data distribution, with all details preserved, rather than merely identifying distinguishable

patterns in data. As a result, their training typically requires a large number of steps of gradient descent. Under this premise, the mechanism to realize DP demands a large amount of noise to be added in the training process, which may overwhelm meaningful signals extracted from the training data.

As such, there is a quandary over whether or not to adopt DP in the simulation of synthetic medical data. Without DP, privacy assessments prior to sharing synthetic data can only rely on knowledge about the types of privacy threats that have been previously recognized, while the technology that can be leveraged for privacy intrusion against synthetic data is still evolving. As a result, privacy risks persist and what seems protected today may not be tomorrow. On the other hand, there is no evidence that sharing synthetic data generated by a model that is not differentially private or that endures high privacy loss (defined by DP) leads to considerable privacy risks, as there may be no existing technique for deriving meaningful knowledge about any real individual from the synthetic data. Under this premise, it may not be necessary to sacrifice utility in exchange for the privacy guarantee of DP.

However, the fact that the current generative models that embrace DP demonstrating low utility do not necessarily indicate utility and theoretically-guaranteed privacy cannot be both realized. The current algorithm to realize DP for deep generative models loosely estimates the upper bound of privacy loss. Future work may enable a much tighter estimation of privacy loss, thereby mitigating the privacy-utility tradeoff in adopting DP.

6.2.3 An Open Question Regarding Synthetic Data's Application

Moving steps forward, there is an open question worth discussing. *What could be the real world application of synthetic medical data?*

Ideally, synthetic datasets would share the same underlying distribution as the original real datasets. However, this expectation can hardly be achieved in real world practice due to the imperfectness of generative methods. Recognizing this, anyone who intends to use synthetic data may have concerns about its reliability. To date, there is no evidence that synthetic data can serve as a substitute for real data in an application-agnostic context. Various programs have embraced synthetic data to help support dissemination and outreach activities. For instance, the U.S. National Institutes of Health-sponsored National COVID Cohort Collaborative (N3C) [118] and U.K. Medicines and Healthcare Regulatory

Agency-sponsored Clinical Practice Research Datalink[1] are providing access to synthetic versions of their datasets. Additionally, NIH's AIM-AHEAD program[2] also incorporated synthetic data into their research goals. However, none of these programs asserts that synthetic data will be useful for applications including training machine learning algorithms for hypothesis formulation and testing for precision medicine. At the moment, it is believed that synthetic data can assist people in generating insight into real data without actually accessing them, but it is under the exploration of what the insight could be like. As such, a reasonable application of synthetic medical data, as of now, is to use them not as an alternative to the real data but as a supplement. For example, synthetic data can be used as a demonstration of real data to help people determine whether they want to go through restricted censors and a possibly lengthy process to apply access to real data.

[1] https://www.cprd.com/content/synthetic-data

[2] https://datascience.nih.gov/artificial-intelligence/aim-ahead

References

[1] Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372.

[2] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. Proceedings of the IEEE Symposium on Security and Privacy. 2008:111-25.

[3] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems. 2002;10.

[4] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. -diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data. 2007;1.

[5] Ninghui L, Tiancheng L, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and -diversity. Proceedings of the International Conference on Data Engineering. 2007:106-15.

[6] Xiao X, Tao Y. M-invariance: Towards privacy preserving re-publication of dynamic datasets. Proceedings of the ACM SIGMOD International Conference on Management of Data. 2007:689-700.

[7] Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: Theory meets practice on the map. Proceedings of the International Conference on Data Engineering. 2008:277-86.

[8] Rubun DB. Discussion statistical disclosure limitation. Journal of Official Statistics. 1993;9:461-8.

[9] Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering. 2021;5.

[10] Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment. 2018;11:1071-83.

[11] Reiter JP. New Approaches to Data Dissemination: A Glimpse into the Future (?). CHANCE. 2004;17:11-5.

[12] Sweeney L. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine and Ethics. 1997;25.

[13] Liu KS, Xiao C, Li B, Gao J. Performing co-membership attacks against deep generative models. Proceedings of the IEEE International Conference on Data Mining. 2019:459-67.

[14] Hayes J, Melis L, Danezis G, Cristofaro ED. LOGAN: Membership Inference Attacks Against Generative Models. Proceedings on Privacy Enhancing Technologies. 2019:133-52.

[15] Chen D, Yu N, Zhang Y, Fritz M. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. Proceedings of the ACM Conference on Computer and Communications Security. 2020:343-62.

[16] Hilprecht B, Härterich M, Bernau D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. Proceedings on Privacy Enhancing Technologies. 2019:232-49.

[17] Mukherjee S, Xu Y, Trivedi A, Patowary N, Ferres JL. privGAN: Protecting GANs from membership inference attacks at low cost to utility. Proceedings on Privacy Enhancing Technologies. 2021:142-63.

[18] Carlini N, Liu C, Úlfar Erlingsson, Kos J, Song D. The secret Sharer: Evaluating and testing unintended memorization in neural networks. Proceedings of the 28th USENIX Security Symposium. 2019:267-84.

[19] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics. 2008;4.

[20] Backes M, Berrang P, Humbert M, Manoharan P. Proceedings of the ACM Conference on Computer and Communications Security. 2016:319-30.

[21] Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks Against Machine Learning Models. Proceedings of the IEEE Symposium on Security and Privacy. 2017:3-18.

[22] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. Proceedings of the 23rd USENIX Security Symposium. 2014:17-32.

[23] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the ACM Conference on Computer and Communications Security. 2015;2015-October:1322-33.

[24] Dwork C. Differential privacy: A survey of results. International conference on theory and applications of models of computation. 2008:1-19.

[25] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science. 2013;9.

[26] Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially Private Generative Adversarial Network. arXiv preprint arXiv:180206739. 2018 2.

[27] Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes. 2019;12.

[28] Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. Journal of the American Medical Informatics Association. 2020;27:99-108.

[29] Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. SynTEG: A framework for temporal structured electronic health data simulation. Journal of the American Medical Informatics Association. 2021;28:596-604.

[30] Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. Journal of biomedical informatics. 2022;125:103977.

[31] McLachlan S, Dube K, Gallagher T. Using the CareMap with Health Incidents Statistics for Generating the Realistic Synthetic Electronic Healthcare Record. Proceedings of the 2016 IEEE International Conference on Healthcare Informatics. 2016:439-48.

[32] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association. 2018;25.

[33] Zeng Q, Cimino JJ. A knowledge-based, concept-oriented view generation system for clinical data. Journal of Biomedical Informatics. 2001;34.

[34] den Bulcke TV, Leemput KV, Naudts B, van Remortel P, Ma H, Verschoren A, et al. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics. 2006;7.

[35] Buczak AL, Babin S, Moniz L. Data-driven approach for creating synthetic electronic medical records. BMC Medical Informatics and Decision Making. 2010;10.

[36] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in Neural Information Processing Systems. 2014;3:2672-80.

[37] Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. Proceedings of the 2nd Machine Learning for Healthcare Conference. 2017.

[38] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. The 7th International Conference on Learning Representations. 2019.

[39] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. The 6th International Conference on Learning Representations. 2018.

[40] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. The 4th International Conference on Learning Representations. 2016.

[41] Fedus W, Goodfellow I, Dai AM. MaskGAN: Better text generation via filling in the _. The 6th International Conference on Learning Representations. 2018.

[42] Yu L, Zhang W, Wang J, Yu Y. SeqGAN: Sequence generative adversarial nets with policy gradient. 31st AAAI Conference on Artificial Intelligence. 2017.

[43] Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D. Adversarial learning for neural dialogue generation. The 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[44] Donahue C, McAuley J, Puckette M. Adversarial audio synthesis. 7th International Conference on Learning Representations. 2019.

[45] Engel J, Agrawal KK, Chen S, Gulrajani I, Donahue C, Roberts A. Gansynth: Adversarial neural audio synthesis. The 7th International Conference on Learning Representations. 2019.

[46] Baowaly MK, Lin CC, Liu CL, Chen KT. Synthesizing electronic health records using improved generative adversarial networks. Journal of the American Medical Informatics Association. 2019;26.

[47] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd International Conference on Machine Learning. 2015;1:448-56.

[48] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016:770-8.

[49] Borji A. Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding. 2019;179.

[50] Hjelm RD, Jacob AP, Che T, Trischler A, Cho K, Bengio Y. Boundary-seeking generative adversarial networks. The 6th International Conference on Learning Representations. 2018.

[51] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. The 34th International Conference on Machine Learning. 2017;1.

[52] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs. Advances in Neural Information Processing Systems. 2017;30.

[53] Ba JL, Kiros JR, Hinton GE. Layer Normalization. arXiv:160706450v1. 2015.

[54] Esteban C, Hyland SL, Rätsch G. Real-valued (medical) time series generation with recurrent conditional gans. arXiv preprint arXiv:170602633. 2017.

[55] Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. The 5th International Conference on Learning Representation. 2017.

[56] Kingma DP, Welling M. Auto-encoding variational bayes. 2nd International Conference on Learning Representations. 2014.

[57] Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. LINE: Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web. 2015:1067-77.

[58] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. SIGMOD Record (ACM Special Interest Group on Management of Data). 2000;29.

[59] Yan C, Zhang Z, Nyemba S, Malin BA. Generating Electronic Health Records with Multiple Data Types and Constraints. AMIA Annual Symposium proceedings AMIA Symposium. 2020;2020.

[60] Ma F, Wang Y, Gao J, Xiao H, Zhou J. Rare disease prediction by generating quality-assured electronic health records*. Proceedings of the 2020 SIAM International Conference on Data Mining. 2020:514-22.

[61] Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A deep dynamic memory model for predictive medicine. Pacific-Asia conference on knowledge discovery and data mining. 2016:30-41.

[62] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports. 2016;6.

[63] Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. The 4th International Conference on Learning Representations. 2016.

[64] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR workshop and conference proceedings. 2016;56.

[65] Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: A deep learning approach. 16th SIAM International Conference on Data Mining 2016. 2016:432-40.

[66] Metz L, Sohl-Dickstein J, Poole B, Pfau D. Unrolled generative adversarial networks. 5th International Conference on Learning Representations. 2017.

[67] Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, et al. Adversarially learned inference. 5th International Conference on Learning Representations. 2017.

[68] Berthelot D, Schumm T, Metz L. BEGAN: Boundary Equilibrium Generative Adversarial Networks. arXiv preprint arXiv:170310717. 2017.

[69] Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least Squares Generative Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision. 2017:2794-802.

[70] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems. 2017;30.

[71] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems. 2014;4.

[72] Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets. The 4th International Conference on Learning Representationss. 2016.

[73] Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. CEUR Workshop Proceedings. 2019;2540.

[74] Nagarajan V, Kolter JZ. Gradient descent GAN optimization is locally stable. Advances in Neural Information Processing Systems. 2017;30.

[75] Mescheder L, Geiger A, Nowozin S. Which training methods for GANs do actually converge? The 35th International Conference on Machine Learning. 2018;8:3481-90.

[76] Roth K, Lucchi A, Nowozin S, Hofmann T. Stabilizing training of generative adversarial networks through regularization. Advances in Neural Information Processing Systems. 2017;30.

[77] Omi T, Ueda N, Aihara K. Fully neural network based model for general temporal point processes. Advances in Neural Information Processing Systems. 2019;32.

[78] Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS ONE. 2017;12.

[79] Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature Biotechnology. 2013;31.

[80] Wang Y, Ni Z, Song S, Yang L, Huang G. Revisiting locally supervised learning: An alternative to end-to-end training. The 9th International Conference on Learning Representations. 2021.

[81] The "All of Us" Research Program. New England Journal of Medicine. 2019;381:1883-5.

[82] Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical Pharmacology and Therapeutics. 2008;84:362-9.

[83] Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduled sampling for sequence prediction with recurrent neural networks. Advances in Neural Information Processing Systems. 2015;28.

[84] Ranzato M, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. The 4th International Conference on Learning Representations. 2016.

[85] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748. 2018.

[86] Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. The 37th International Conference on Machine Learning. 2020:9929-39.

[87] Wang F, Xiang X, Cheng J, Yuille AL. NormFace: L2 hypersphere embedding for face verification. Proceedings of the 2017 ACM Multimedia Conference. 2017:1041-9.

[88] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. The 37th International Conference on Machine Learning. 2020:1575-85.

[89] Wells MT, Casella G, Robert CP, et al. Generalized accept-reject sampling schemes. In: A festschrift for herman rubin. vol. 45; 2004. p. 342-8.

[90] Azadi S, Odena A, Olsson C, Darrell T, Goodfellow I. Discriminator rejection sampling. The 7th International Conference on Learning Representations. 2019.

[91] Lee D, Yu H, Jiang X, Rogith D, Gudala M, Tejani M, et al. Generating sequential electronic health records using dual adversarial autoencoder. Journal of the American Medical Informatics Association. 2020 9;27:1411-9.

[92] El Emam K, Mosquera L, Fang X, El-Hussuna A, et al. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. JMIR Medical Informatics. 2022;10(4):e35734.

[93] El Emam K, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. Journal of the American Medical Informatics Association. 2021;28:3-13.

[94] Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ open. 2021;11(4):e043497.

[95] Woo MJ, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality. 2009;1(1).

[96] Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A kernel method for the two-sample-problem. Advances in Neural Information Processing Systems. 2006;19.

[97] Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ digital medicine. 2020;3(1):1-13.

[98] Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, et al. Defending against neural fake news. Advances in Neural Information Processing Systems. 2019;32:770-8.

[99] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26.

[100] Reiter JP. Using CART to generate partially synthetic public use microdata. Journal of Official Statistics. 2005;21:441-62.

[101] Reiter JP. Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology. 2003;29.

[102] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. MixUp: Beyond empirical risk minimization. The 6th International Conference on Learning Representations. 2018.

[103] Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. Proceedings of the IEEE Computer Security Foundations Symposium. 2018:268-82.

[104] Long Y, Bindschaedler V, Wang L, Bu D, Wang X, Tang H, et al. arXiv preprint arXiv:180204889. 2018.

[105] Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. arXiv preprint arXiv:180601246. 2018.

[106] Jayaraman B, Wang L, Knipmeyer K, Gu Q, Evans D. Revisiting Membership Inference Under Realistic Assumptions. Proceedings on Privacy Enhancing Technologies. 2021:348-68.

[107] Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. Journal of official statistics. 2003;19(1):1.

[108] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:7871-80.

[109] Kobayashi S. Contextual augmentation: Data augmentation bywords with paradigmatic relations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018;2:452-7.

[110] Donahue J, Darrell T, Krähenbühl P. Adversarial feature learning. The 5th International Conference on Learning Representations. 2017.

[111] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. Advances in Neural Information Processing Systems. 2016:2180-8.

[112] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research. 2020;21.

[113] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877-901.

[114] Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A Survey on Contrastive Self-Supervised Learning. Technologies. 2020;9:2.

[115] Le-Khac PH, Healy G, Smeaton AF. Contrastive Representation Learning: A Framework and Review. IEEE Access. 2020;8:193907-34.

[116] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019;1:4171-86.

[117] Abadi M, McMahan HB, Chu A, Mironov I, Zhang L, Goodfellow I, et al. Deep learning with differential privacy. Proceedings of the ACM Conference on Computer and Communications Security. 2016:308-18.

[118] Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association. 2021;28:427-43.

Appendix A

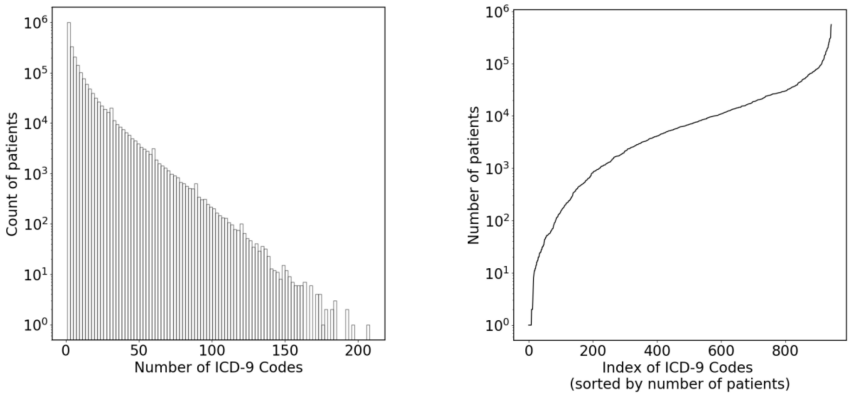
Data Triage Strategy in Chapter 2

Figures A.1 presents the number of patients who exhibited a certain number of distinct ICD-9 codes. It can be seen that many patients have only a few ICD-9 codes. This happens for a number of reasons. Some of these patients, for instance, are healthy or are diagnosed with minor issues in an outpatient setting (e.g., influenza). In other situations, the patients are observed for only a short period of time either because they were treated in an emergency and never seen again or they moved on to a different healthcare organization. As a consequence, these EMRs are somewhat incomplete or lacking in information and thus cannot support the generation of meaningful synthetic ICD-9 code lists of patients.

At the same time, it can be seen that some patients have an abnormally large number of ICD-9 codes. This phenomenon was rare and may correspond to anomalous events (e.g., test records) or patients that have been observed for long periods of time.

Additionally, from the perspective of the individual ICD-9 codes, Figure A.1b presents the number of patients assigned each ICD-9 code. It can be seen that there is a subset which is rare in SD, as shown by the left-most region of Figure A.1b. The primary reason lies in the rareness of the corresponding diseases among the population visiting VUMC.

Based on this evidence, we remove the EMRs with less than 5 or more than 100 distinct ICD-9 codes. We further remove the ICD-9 codes appearing in less than 100 EMRs.



(a) Patients per ICD-9 code set size.

(b) Patients per ICD-9 code.

Figure A.1: Summary statistics about EMRs and billing codes in the VUMC SD.

Appendix B

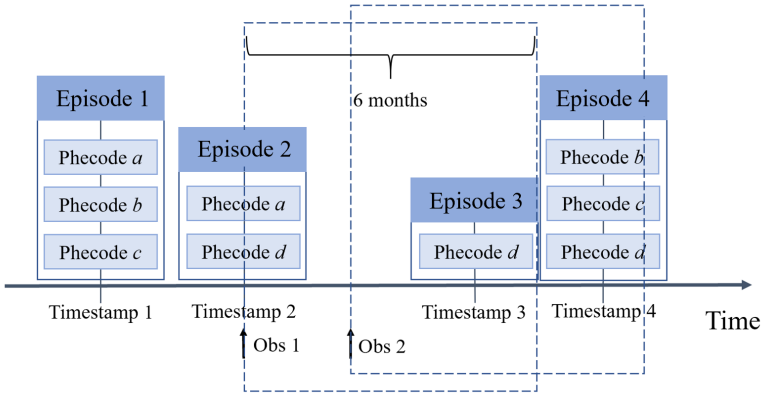
Chronic Disease Subpopulations in Chapter 3

This appendix provides details on how the chronic disease subpopulations were composed for this study. Specifically, we build the chronic disease subpopulations from both the real and synthetic datasets through a three steps process. First, we select all real and synthetic records that satisfy the definition of a disease of interest. Each record must contain at least one indicative phecode 250.2X, 428.X, 401(401.1), and 496.21 for Type-2 diabetes, heart failure, hypertension, and chronic obstructive pulmonary disease (COPD), respectively. Second, we retain only records that 1) have at least one visit prior to the first occurrence of the indicative phecode and 2) have at least ten visits within the five-year period following the indication of the disease, as they contain relatively complete temporal patterns of the phenotype. Third, to focus on the temporal patterns of the target phenotype, for all selected records we retain only the phecodes that are positively correlated with the disease of interest. Here, the correlation of a diagnosis code with the disease of interest is represented as the risk ratio of the diagnosis and the disease (on a log scale). Phecodes with ratios beyond two standard deviations are considered positively correlated. Table B.1 provides summary statistics of the selected subpopulations.

We sample 28 observation points uniformly cover the 5-year period following the first presence of the disease of interest, each of which contains all unique phecodes assigned to the patient in the following 6-month window. An example is illustrated in Figure. An example is illustrated in Figure B.1. Thus, we build ml feature matrices M_r and M_s for the real and synthetic data, respectively, where m is the number of records and l is the number of phecodes times 28.

Table B.1: A summary of the selected diseases and their positively correlated phecodes (PCP).

Disease	**Real Records**	**PCPs**	**Episodes with PCP(s) in the** Observation Windos
Type-2 diabetes	4,949	20	19.93
Heart failure	4,161	42	18.52
Hypertension	8.836	53	23.46
COPD	611	9	22.78



Observation 1				Observation 2				
Phecode *a*	Phecode *b*	Phecode *c*	Phecode *d*	Phecode *a*	Phecode *b*	Phecode *c*	Phecode *d*	
1	0	0	1	0	1	1	1	A row in feature matrix

Figure B.1: An example of how one row of the feature matrix is composed. Phecode *d* indicates the target disease.