Enhancing Collaborative Medical Outcomes through Private Synthetic Hypercube Augmentation: PriSHA

Shinpei Nakamura-Sakai S.NAKAMURA.SAKAI@YALE.EDU
Department of Statistics and Data Science, Yale University, CT, USA

Dennis Shung*[*] DENNIS.SHUNG@YALE.EDU
Yale School of Medicine, CT, USA

Jasjeet Sekhon[*] JASJEET.SEKHON@YALE.EDU
Department of Statistics and Data Science, Yale University, CT, USA

Abstract

Data sharing across multiple health systems has the significant challenge of maintaining data privacy. Access to detailed, high-quality data is important for machine learning models trained to predict clinically relevant outcomes to generalize across different patient populations. However, health systems often are limited to patient data within their networks, which may not adequately represent the breadth of patient populations. This limitation is especially pronounced in the case of patients with rare or unique characteristics, resulting in decreased accuracy for this minority group. To address these challenges, our work introduces a framework designed to enhance existing clinical models, Private Synthetic Hypercube Augmentation (PriSHA). We use generative models to produce synthetic data as a means to augment these models while adhering to strict privacy standards. This approach has the potential to improve model performance without compromising patient confidentiality. To our knowledge, our framework is the first synthetic data augmentation framework that merges privacy-preserving tabular data and real data from multiple sources.

Data and Code Availability We applied our methodology to two electronie health record (EHR) datasets. The first dataset includes EHR data from patients admitted to a tertiary academic health center with acute gastrointestinal bleeding, a common gastroenterological condition that requires hospital-based care, from 2014 to 2019. This dataset is not publicly accessible. The second dataset is extracted from the eICU Collaborative Research Database, a publicly available ICU dataset pooled across multiple institutions Pollard et al. (2018). We have included the relevant code and data as a zip file in the supplemental materials.

Institutional Review Board (IRB) Our research requires IRB approval. This IRB information will be provided if the paper is accepted.

1. Introduction

A primary challenge for machine learning models for clinical decision support is ensuring data privacy. Clinical models use EHR data as input variables, which may contain both direct and indirect identifiers that can be used to link the patient to their protected health information (PHI). The Health Insurance Portability and Accountability (HIPAA) law Fitzgerald (2015) specifically defines 18 identifiers that must be removed, and the nationally accepted standard is no greater than 0.04% reidentification risk Emam (2013). Synthetic data is a promising solution to maintain patient privacy while enabling wider use of previously sensitive data. Synthetic data has been used for a wide array of applications across finance, satellite images, and healthcare Jordon et al. (2022) Giuffrè and Shung (2023). Synthetic data effectively combats data scarcity by creating datasets imbued with characteristics that are useful for downstream analysis; these characteristics are fidelity, which is the degree of resemblance in distribution between the synthetic and original datasets and utility, representing the applicability and effectiveness of the data in specific tasks. Most importantly, synthetic data maintains privacy, protecting sensitive information in the original dataset from being exposed.

[*] Denotes co-senior authorship.

While synthetic data presents a valuable solution for research and analysis when privacy concerns restrict access to real data, several drawbacks exist. As highlighted by Hittmeir et al. (2019), synthetic data can result in reduced downstream performance Manousakas and Aydore (2023) further notes a lack of substantial evidence supporting the usefulness of synthetic tabular data for augmentation. A key issue might be that not all synthetic data generated is beneficial, some of it could be irrelevant or noisy, with its utility varying significantly across different downstream tasks. This underscores the necessity of implementing task-specific supervision and managing the relationship among multiple covariates. The challenge often lies not in utilizing the entirety of the synthetic data but in identifying and leveraging specific segments, such as data about male patients or, more narrowly, male patients of a certain ethnicity. This example underscores the potential value of selectively navigating the patient characteristics of the synthetic data to enhance model performance. Additionally, in a systematic review by Hernandez et al. (2022), various studies focus on data augmentation or privacy preservation in medical tabular data, yet none tackle the issue of synthetic data augmentation using privacy-preserving synthetic data from multiple data sources.

Moreover, different hospitals or health systems may have different distributions, which could be mitigated through data sharing. However, the imperative need to protect patient privacy makes this infeasible. Therefore, it is crucial to develop and implement frameworks aimed at maximizing performance augmentation using synthetic data while maintaining privacy guarantees, ensuring the optimal utilization of synthetic data in healthcare contexts.

Our contributions

1. Our approach effectively adapts synthetic data to datasets with varying distributions, taking advantage of distribution shifts.

2. Our approach employs supervised learning to select synthetic data subsets to enhance downstream task performance. This method outperforms traditional data augmentation by leveraging this diversity for more effective results.

3. Our methodology demonstrates improved performance over standard augmentation approaches while maintaining a privacy guarantee. This approach facilitates collaborative medical outcomes by enabling the sharing of private synthetic data to bolster machine-learning models.

2. Related Work

2.1. Synthetic Tabular Data Generation

Previous methods to generate synthetic tabular data include traditional statistical models Li et al. (2020), random-forest-based methods Caiola and Reiter (2010), and GANs (Generative Adversarial Networks) Park et al. (2018), Xu et al. (2019) Zhao et al. (2022), and diffusion-based model Kotelnikov et al. (2023).

2.2. Application to EHR data

Synthetic data generation for Electronic Health Records (EHR) offers significant benefits Hernandez et al. (2022) by addressing two critical challenges: data privacy Choi et al. (2017), Norgaard et al. (2018), Yoon et al. (2023) and dataset enhancement Che et al. (2017), Fowler et al. (2020), Koivu et al. (2020).

We propose a framework that guarantees differential privacy while specifically addressing disparities in data distribution by providing a variety of synthetic data with unique distributions. We demonstrate the value of our framework in improving the generalizability and robustness of predictive models.

While augmentation techniques have achieved significant success in the realm of image data Frid-Adar et al. (2018) Sandfort et al. (2019), the exploration of similar approaches for tabular data is relatively underdeveloped. Additionally, there is growing evidence to suggest that the augmentation of tabular data is fundamentally more difficult than for language or text data Manousakas and Aydore (2023). As a result, refining and enhancing existing augmentation methods is essential for the effective generation of synthetic tabular data. In this paper, we define traditional augmentation as the process of combining real and synthetic data in any proportion.

2.3. Distribution Shift

Distribution shifts in Electronic Health Record (EHR) data can arise from various factors, including changes in patient demographics, evolution in medical practices, and the transferability of models across health systems Avati et al. (2021). Our research will primarily address scenarios where multiple hospitals serve distinct patient populations. Specifically, we aim to enhance model performance in situations where a particular hospital's predictive models underperform for certain patient subgroups, while other institutions hold relevant data for these minority groups. Despite the critical nature of addressing such distributional shifts, research on frameworks designed to tackle this specific challenge remains scarce. For instance, Nestor et al. (2019) evaluated various advanced prediction models, revealing a decline in predictive accuracy when models trained on historical data were applied to future datasets. They suggested a mitigation strategy that involved organizing features into clinical concept groups. While preprocessing strategies prove beneficial, there is an evident need for more sophisticated model-based approaches for OOD detection and mitigation. This need underpins our current research. Although several methods have successfully managed OOD data in the context of image processing Hendrycks and Dietterich (2019), the exploration of handling OOD synthetic data in tabular formats is still in its infancy. To bridge this gap, we propose a novel approach that leverages Bayesian optimization to identify and utilize the most informative data subsets for augmentation, aiming to effectively counter the challenges posed by distributional shifts in EHR data.

2.4. Federated Learning

Federated Learning (FL) represents an approach in machine learning, that emphasizes decentralized training across a variety of devices or servers, each managing its unique local dataset. This approach notably bolsters privacy, as it circumvents the need to exchange sensitive data, with entities like smartphones or hospital servers transmitting only model updates to a central aggregator. This not only enhances privacy but also optimizes bandwidth usage Konecný et al. (2016); McMahan et al. (2016); Bonawitz et al. (2019); Rieke et al. (2020). Nevertheless, FL is not without challenges. Variations in data distributions across nodes can negatively impact model efficacy Kairouz et al. (2019). Furthermore, a significant privacy concern has been raised by Hitaj et al. (2017), revealing that private data could potentially be reconstructed from model updates using GANs, thus challenging the perceived security of FL. These issues underline the imperative for meticulous analysis of FL's limitations, especially in contexts demanding secure and precise data management. In response to these challenges, particularly the non-identical distribution issue, our methodology aims to enhance model performance through private data augmentation from diverse sources, addressing the distribution shift while preserving privacy.

3. Private Synthetic Data Generation

We use differential privacy to ensure the privacy of the generated synthetic data introduced in Dwork et al. (2014). Define an algorithm $\mathcal{M}$ which takes input dataset $\mathcal{D}$ and outputs values to an output space $\mathcal{O}$. We can define the Neighboring dataset as

Definition 1 (Neighboring Datasets) *Datasets $\mathcal{D}$ and $\mathcal{D}'$ are neighboring if*

$$\exists x \in \mathcal{D} \; s.t \; \mathcal{D} \setminus \{x\} = \mathcal{D}'$$

Definition 2 (Differential Privacy) *A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differential private if for all $\mathcal{S} \subset \mathcal{O}$ and for all $\mathcal{D}$ and $\mathcal{D}'$ which differs only on a single observation:*

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in \mathcal{S}) \leq e^{\epsilon} \, \mathbb{P}(\mathcal{M}(\mathcal{D}') \in \mathcal{S}) + \delta$$

where $\mathcal{O}$ is the output space, ϵ and δ are parameters in differential privacy, where ϵ represents the privacy loss, with smaller values indicating stronger privacy and δ quantifies the probability that the privacy guarantee might not hold, aiming for it to be close to zero.

The comprehensive survey by Bauer et al. (2024) reviews 417 models for generating synthetic data, highlighting the Differentially Private GAN (DP-GAN) Xie et al. (2018), Private Aggregation of Teacher Ensembles-GAN (PATE-GAN) Jordon et al. (2018), and PrivBayes Zhang et al. (2017) for their differential privacy guarantees tailored to tabular data. We concentrate on PATE-GAN and DP-GAN, given that EHR data is often high-dimensional, whereas PrivBayes tends to underperform or becomes computationally infeasible Ganev et al. (2021). Although Kotelnikov et al. (2023) signifies a recent high-utility innovation through a diffusion-based model, diffusion-based models with differential privacy for tabular data remain unidentified.

DPGAN and PATE-GAN are both machine learning frameworks designed to generate synthetic data while ensuring the privacy of individuals in the training dataset. DPGAN integrates differential privacy into the traditional GAN structure by adding noise to the gradients, ensuring that the final model does not reveal sensitive information about the training data. This approach, however, requires a careful balance between data utility and privacy. On the other hand, PATE-GAN employs a different strategy based on the PATE framework Papernot et al. (2018). It uses an ensemble of teacher models, each trained on a disjoint subset of the original data, to generate labels for a student model. The student model, which is a GAN in the case of PATE-GAN, then learns to generate synthetic data based on these labels. The PATE framework ensures that the student model's learning process is differentially private, as it only has access to aggregated information from the teacher models, significantly reducing the risk of exposing sensitive information from the training dataset.

4. Synthetic Data Augmentation from Multiple Data Sources

In healthcare analytics, recognizing and adjusting for distribution shifts in EHR is paramount, as these shifts profoundly influence the precision and dependability of machine learning models used in clinical decision-making. Such shifts can be attributed to a variety of factors, including evolving patient demographics, changes in clinical practices, or alterations in how data is documented. These changes can substantially compromise the efficacy of models initially trained on EHR data, making it crucial to continually adapt these models. Ensuring models remain generalizable across diverse healthcare contexts, upholding high standards of data integrity, and adhering to the stringent ethical and regulatory demands of the healthcare sector are essential. Additionally, this approach supports the advancement of personalized medicine by accommodating the dynamic and evolving nature of healthcare data, which is essential for optimizing patient care and the efficient allocation of medical resources.

To conceptualize a specific scenario, envision two datasets: one from Hospital A, denoted as $\mathcal{D}_A = \{\mathcal{Y}_A, \mathcal{X}_A\}$, and another from Hospital B, $\mathcal{D}_B = \{\mathcal{Y}_B, \mathcal{X}_B\}$, each with distinct distributions where $\mathcal{Y}_* \in \mathbb{R}^{M_*}$ and $\mathcal{X}_* \in \mathbb{R}^{M_* \times p}$ and M_* and p indicates number of observation and covariates. Suppose our objective is to predict outcomes for $\mathcal{D}_A^{pred}$, a dataset of patients more closely aligned in characteristics with $\mathcal{D}_B$ than with $\mathcal{D}_A$. Specifically, $\mathcal{D}_A^{pred}$ refers to a dataset within Hospital A where the model trained using only $\mathcal{D}_A$ struggles to make accurate predictions. This challenge arises because the data in $\mathcal{D}_A^{pred}$ represents a minority distribution within Hospital A, indicating that the model's ability to predict outcomes for this subset is not as strong as for the majority distribution. However, the predictive performance on $\mathcal{D}_A^{pred}$ could potentially be improved by incorporating data from $\mathcal{D}_B$, which may provide additional insights or represent similar minority distributions from another context, thereby enriching the model's training data and enhancing its accuracy for the challenging subset within Hospital A.

In an initial approach, we might consider utilizing solely $\mathcal{D}_A$ for our model, $\hat{\mu}_{\mathcal{D}_A} = f(\mathcal{Y}_A \sim \mathcal{X}_A)$, where $f(\mathcal{Y} \sim \mathcal{X})$ symbolizes a regression motivated estimator (though f could represent any machine learning estimator), and $\hat{\mu}$ is the resultant learned function.

However, in an ideal scenario devoid of privacy constraints, we could incorporate $\mathcal{D}_B$, forming a combined dataset

$$\mathcal{D}_{AB} = \begin{pmatrix} \mathcal{D}_A \\ \mathcal{D}_B \end{pmatrix} = (\mathcal{Y}_{AB}, \mathcal{X}_{AB})$$

and then fit $\hat{\mu}_{\mathcal{D}_{AB}} = f(\mathcal{Y}_{AB} \sim \mathcal{X}_{AB})$. This model is expected to perform better, benefiting from the generalizability gained from $\mathcal{D}_B$.

The focus of this manuscript, however, is on an alternative scenario where, due to data privacy concerns, we cannot directly use $\mathcal{D}_B$. Instead, we generate a differentially private synthetic dataset $S_B(N; \epsilon, \delta)$ from $\mathcal{D}_B$, where S_B includes N observations, formulated under privacy constraints ϵ and δ defined in definition 2. Typically, this N is the number of observation of $\mathcal{D}_A^{pred}$. The dataset for model training then becomes

$$\mathcal{D}_{AB'} = \begin{pmatrix} \mathcal{D}_A \\ \mathcal{S}_B(N; \epsilon, \delta) \end{pmatrix} = (\mathcal{Y}_{AB'}, \mathcal{X}_{AB'}),$$

and we fit $\mu(\mathcal{D}_{AB'}) = f(\mathcal{Y}_{AB'} \sim \mathcal{X}_{AB'})$. This approach leverages data from Hospital B with a defined level of privacy assurance. The underlying hypothesis is that this model, while potentially less informative than $\mu(\mathcal{D}_{AB})$ using real data, should still surpass the performance of $\mu(\mathcal{D}_A)$, as it incorporates information from the patient group of interest, albeit with data that has been modified for privacy considerations.

5. Private Synthetic Hypercube Augmentation (PriSHA)

We define the approach of Standard Augmentation using the findings of Manousakas and Aydore (2023),

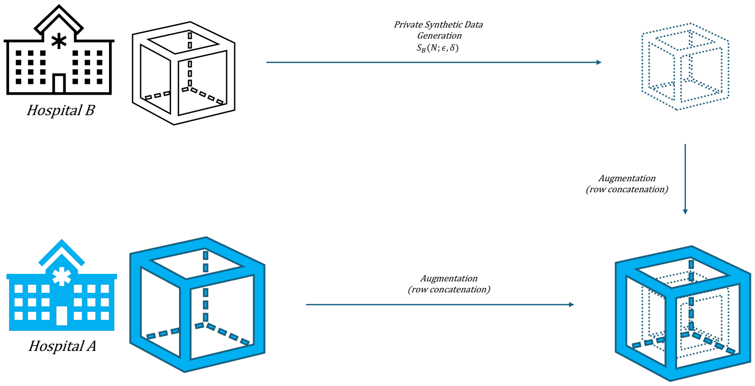

Figure 1: Overview of Standard Augmentation

which reveal the limited benefits of synthetic data augmentation in model training enhancement. Unlike their method, which solely uses synthetic data generated from the same source, our strategy involves the incorporation of synthetic data from diverse sources, each with distinct distributions, denoted as $\mathcal{D}_B$, and combining this with our primary data source, $\mathcal{D}_A$, to which we have full access. Moreover, our approach integrates considerations for differential privacy, ensuring enhanced data protection. The approach is visualized in Figure 1.

We propose a refinement to this augmentation process by introducing a selective augmentation strategy. We conceptualize the dataset $\mathcal{D}_B \in \mathbb{R}^{M_B \times p}$ as a p-dimensional hypercube, with each dimension representing a covariate, and selectively augment the dataset with only pertinent segments of synthetic data for row concatenation with real data. To clarify this point, Figure 2 can be compared to Figure 1. Here, the selection of the red cube from the private synthetic data is performed through Bayesian optimization in a supervised manner, ensuring only "useful" data that enhances the model performance is utilized. We call this framework Private Synthetic Hypercube Augmentation (PriSHA).

Consider a practical scenario involving a two-dimensional dataset comprising patients' age and heart rate. Suppose your dataset $\mathcal{D}_A$ shows suboptimal performance in predicting outcomes for young patients with high heart rates, a subset of data available at Hospital B. Due to privacy concerns, Hospital B cannot directly share this data. In this context, our enhanced model, through the PriSHA framework, can acquire differentially private data on young patients with high heart rates from Hospital B, thereby improving model accuracy of our model predicting patients in Hospital A.

The PriSHA framework begins by computing the feature importance for each covariate in the dataset, employing any established metric from the literature to identify the most significant contributors to the model's predictive power. Based on this analysis, it selects a specific number of covariates or dimensions of the hypercube, denoted by K, for slicing; this selection is flexible and can vary as shown in Figure 2, where slicing occurs on three dimensions, but K can be any natural number. The final step involves optimizing the intervals for each chosen dimension using Bayesian optimization to determine the most relevant synthetic data for augmentation, thereby enhancing the dataset with precision-targeted information for improved model performance. Further details on each of these steps will be discussed in the remainder of this section, providing a deeper understanding of the methodology and its application.

The initial phase of our methodology involves estimating the feature importance for each covariate within the dataset. In machine learning, the significance of features can be evaluated using a variety of techniques, we can utilize any feature importance metric identified in the literature, a choice that will be further discussed in the subsequent section. The selection of the K parameter is crucial in our analysis, significantly influencing the segmentation and

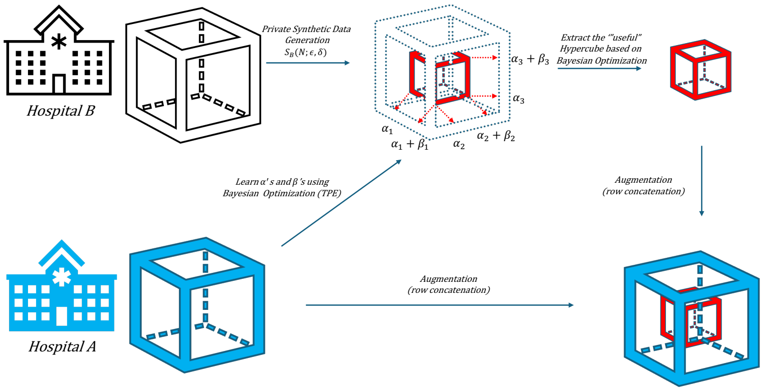

Figure 2: Overview of PriSHA

control over the synthetic data. A smaller K yields a more constrained data segmentation, thereby reducing the level of granularity control. In contrast, a larger K complicates conditional sampling due to the increased computational demand for acquiring specific data samples. The probability of sampling a constrained observation for the i^{th} variable is denoted as $\mathbb{P}(\mathcal{X}_i \in [\alpha_i, \alpha_i + \beta_i])$. For K covariates, this probability is expressed as $\prod_{i=1}^{K} \mathbb{P}(\mathcal{X}_i \in [\alpha_i, \alpha_i + \beta_i])$, and this value decreases quickly as K increases.

Following feature importance and picking K, we proceed to selectively sample the synthetic data. This process is guided by the hyperparameters $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{\beta} \in \mathbb{R}_+^K$. Here, $\boldsymbol{\alpha}$ represents the starting point of the interval from which we sample. In contrast, $\boldsymbol{\beta}$ determines the length of this interval, focusing on the top K features as indicated by their feature importance. This approach results in the generation of a "sliced" K-dimensional hypercube of the synthetic data tailored by $[\alpha_i, \alpha_i + \beta_i] \quad \forall i \in \{1, ..., K\}$. Furthermore, we enhance our methodology with a supervised component during the selection of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. To achieve this, we divide the predicted dataset $\mathcal{D}_A^{pred}$ into two subsets: a validation set $\mathcal{D}_A^{val}$ and a test set $\mathcal{D}_A^{test}$. These subsets come paired with their respective data points and outcomes, denoted by $\mathcal{X}_A^{val}, \mathcal{Y}_A^{val}$ for the validation set and $\mathcal{X}_A^{test}, \mathcal{Y}_A^{test}$ for the test set. The purpose of the validation set, $\mathcal{D}_A^{val}$, is to facilitate the optimal selection of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This is accomplished by addressing the following optimization problem:

$$\begin{aligned}
\boldsymbol{\alpha}^*, \boldsymbol{\beta}^* = \quad \underset{\boldsymbol{\alpha}\in\mathbb{R}^k,,\boldsymbol{\beta}\in\mathbb{R}_+^k}{\arg\min} \quad & \mathcal{L}(\mathcal{Y}_A^{val}, \hat{\mathcal{Y}}_A^{val}) \\
\text{s.t} \quad & D_{B'} = S_B(N;\epsilon,\delta) \\
& \mathcal{D}_{AB'} = \begin{pmatrix} \mathcal{D}_A \\ \mathcal{D}_{B'} \end{pmatrix} \\
& \hat{\mu}_{AB'} = f(\mathcal{Y}_{AB'} \sim \mathcal{X}_{AB'}) \\
& \hat{\mathcal{Y}}_A^{val} = \hat{\mu}_{AB'}(\mathcal{X}_A^{val})
\end{aligned}$$

where $\mathcal{L}$ is the desired loss function of interest.

The hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ define a refined hypercube, selectively including observations considered useful for enhancing the predictive power of our model on $\mathcal{D}_A^{pred}$. This strategic selection aims at improving downstream model performance through the integration of actual data from $\mathcal{D}_A$ with privately generated synthetic data $\mathcal{D}_{B'}^{j^*}$, an outcome derived from the PriSHA algorithm as detailed in Algorithm 1. This concatenation of real and synthetic data results in a combined dataset, formally expressed as $\mathcal{D}_{AB'}^{j^*} = \begin{pmatrix} \mathcal{D}_A \\ \mathcal{D}_{B'}^{j^*} \end{pmatrix}$. Utilizing this dataset $\mathcal{D}_{AB'}^{j^*}$ for model training, our goal is to enhance the prediction accuracy for $\mathcal{D}_A^{pred}$. If our supervised approach proves effective and Hospital B's data contains valuable observationsd for predicting $\mathcal{D}_A^{pred}$, then we expect an uplift in model performance through this framework.

Algorithm 1: Private Synthetic Hypercube Augmentation (PriSHA)

```
Input: $N$, $K$, $\mathcal{D}_A, \mathcal{D}_A^{pred}, \mathcal{D}_B$, $\epsilon$, $\delta$
Split $\mathcal{D}_A^{pred} = \{\mathcal{D}_A^{val}, \mathcal{D}_A^{test}\}$
 Calculate feature importance scores from $f(\mathcal{Y}_\mathcal{A} \sim \mathcal{X}_\mathcal{A})$
 Initialize $\boldsymbol{\alpha}^0 = \{\alpha_i\}_{i \in \{1,\dots,k\}}$ and $\boldsymbol{\beta}^0 = \{\beta_i\}_{i \in \{1,\dots,k\}}$
 Train $S_B(N; \epsilon, \delta)$ using $\mathcal{D}_B$
 for $j = 1$ to $J$ do
     Conditionally sample $D_{B'}^j = S_m(N; \epsilon, \delta)$ with constraint $\mathcal{X}_i \in [\alpha_i, \alpha_i + \beta_i] \quad \forall i \in \{1, \dots, K\}$
     Create $\mathcal{D}_{AB'}^j = \begin{pmatrix} \mathcal{D}_A \\ \mathcal{D}_{B'}^j \end{pmatrix}$
     Train $\hat{\mu}^j = f(\mathcal{Y}_{AB'}^j \sim \mathcal{X}_{AB'}^j)$
     Estimate $\hat{Y}_{val}^j = \hat{\mu}^j(\mathcal{X}_A^{val})$
     Compute $l^j = \mathcal{L}(\hat{Y}_{val}^j, \mathcal{Y}_{val})$
     Suggest $\boldsymbol{\alpha}^{j+1}$ and $\boldsymbol{\beta}^{j+1}$ using Bayesian optimization based on $\{\boldsymbol{\alpha}^0, \dots, \boldsymbol{\alpha}^j\}$, $\{\boldsymbol{\beta}^0, \dots, \boldsymbol{\beta}^j\}$, and $\{l^0, \dots, l^j\}$
end
return $D_{B'}^{j^*}$ where $j^* = \arg\min_j l^j$
```

Table 1: Observation counts, number of features, and outcome distribution across demographic groups in EHR and eICU datasets.

Dataset	Dem Grp	Outcome	Obs	Feat	Label=1
EHR	Hispanic	Composite	599	38	31.55%
	Non-Hisp.		3723		39.70%
eICU	Young	Died	2311	43	9.17%
	Old		7689		12.36%

6. Experimental Results

6.1. Data Description

To evaluate the performance of PriSHA, we conducted a series of analyses using two different datasets: the first, EHR data for patients presenting to a tertiary academic health center with acute GIB, and the second, a dataset of ICU patients from the eICU collaborative research database.

The first dataset comprises EHR data of patients who presented to a tertiary academic health center with acute gastrointestinal bleeding from 2014 to 2019, the most common gastrointestinal condition requiring hospitalization. Input variables include demographics, initially measured vital signs, initially measured laboratory values within 4 hours of presentation, and nursing assessments. The outcome is a binary composite outcome of whether or not the patient received a hospital-based intervention (red blood cell transfusion, intervention to stop bleeding, or all-cause 30-day mortality). From now on, we will refer to this data as GIB EHR data.

The second dataset is extracted from the eICU Collaborative Research Database, and is a collection of patients admitted to intensive care units with more than 200,000 ICU admissions from 2014 to 2015 with collated information collected as part of their hospital stay. The variables included in our dataset include demographics, initial laboratory test results within 24 hours of ICU admission, medications administered, and need for ICU-specific treatments such as advanced respiratory support or vasopressor treatment.

For the GIB EHR data, we used the metadata of race/ethnicity to create two subsets: patients who identified themselves as ethnically Hispanic and those who did not. These subsets were denoted as $\mathcal{D}_A$ and $\mathcal{D}_B$, respectively. In the analysis of the eICU data, we have segregated the dataset into two distinct groups: older and younger populations. For an in-depth explanation of this process including our approach to handling missing data, please refer to Appendix B. For additional details about the datasets, please see Table 1. Although various predictive models were available, we opted for XGBoost Chen and Guestrin (2016) due to its state-of-the-art performance on tabular data Borisov et al. (2022) Shwartz-Ziv and Armon (2021). The model performance was gauged using the Area Under the Receiver Operating Characteristic curve (AUC) as the loss function.

6.2. Baseline Scores

In our initial study, we assessed the impact of dataset enhancement on model performance. We established two AUC benchmarks: one using a model trained on $\mathcal{D}_A$ and tested on $\mathcal{D}_A^{pred}$, and the other with a model trained on the extended dataset $\mathcal{D}_{AB}$, also tested against $\mathcal{D}_A^{pred}$. Applying the DeLong test DeLong et al. (1988), we noted a significant AUC improvement with the augmented dataset, detailed in Table 2. DeLong's test is particularly effective for comparing the AUCs of two models because it accurately accounts for the correlation between these AUC from $\mathcal{D}_A^{pred}$. This makes it more appropriate than other tests for ensuring precise evaluation of statistical significance between correlated ROC curves.

This finding suggests that the integration of $\mathcal{D}_B$ enhances model performance. Augmentations resulting in an AUC below 85.30% and 76.19% are deemed suboptimal, while surpassing 88.87% and 82.91% AUC is a challenging yet attainable goal, especially when incorporating only private synthetic data in EHR and eICU datasets respectively.

Table 2: Comparison of AUC percentages and the associated DeLong's test p-values for different data types and datasets.

Dataset	Data Type	AUC (%)	DeLong pval
EHR	Hisp Only	85.30	0.0167
	Hisp + No-Hisp	88.87	
eICU	Young Only	76.19	0.0025
	Young + Old	82.91	

6.3. Feature Importance

In our study, we evaluated feature importance through metrics such as gain and weight. Specifically, for the eICU dataset, we relied on gain as the key importance metric. In the case of EHR, the analysis revealed that the most significant features based on gain are `lab_HGT` and `lab_HCT`, which represent hemoglobin and hematocrit levels, respectively. Notably, these two features exhibit a high degree of correlation, with a correlation coefficient of 95%, indicating their shared relevance in assessing red blood cell metrics. This high correlation suggests that selecting both features might lead to redundancy in feature representation, as gain tends to create repetitive divisions based on these two similar features. Furthermore, given the nature of boosting algorithms, features utilized after hemoglobin and hematocrit primarily serve to rectify errors from preceding trees, which predominantly involve these two covariates. In contrast, `lab_PLT` shows considerably lower correlations with hemoglobin and hematocrit (3% and 6%, respectively), thereby capturing an orthogonal component that the former two do not predict. Consequently, we opted to use weight over gain for determining feature importance, as it better represents diverse and distinct aspects of our dataset.

The accompanying Figure 3 presents a horizontal bar plot that visually contrasts two metrics: Normalized Gain and Normalized Weight, across various variables. The plot aligns variables on the y-axis against their normalized importance on the x-axis, ranging from 0.0 to 1.0. The blue bars depict Normalized Gain, illustrating the extent to which each variable amplifies the model's predictive accuracy. Conversely, the green bars denote Normalized Weight, indicating the frequency of each variable's usage in data splits across ensemble models like Random Forest or Gradient Boosting. The variables are ordered by Normalized Weight in descending order, signifying that those positioned higher are more frequently utilized in model decisions, while those lower are used less often. This graphical representation aids in comprehending the distinct roles and frequencies of variable utilization within our model. Features include demographic variables (age, biological sex), initial laboratory values (metabolic chemistries, blood counts, and liver function tests), nursing assessments (Glasgow Coma Score [GCS]), and vital sign measurements (blood pressure, respiratory rate, temperature, pulse). The top three features are laboratory values: platelet count, alkaline phosphatase, and hemoglobin level.

Figure 4 is organized by gain, the chosen metric for measurement. Notably, the top three features identified are blood lactate levels, bilirubin, and the partial pressure of oxygen in arterial blood (PaO2). Elevated lactate levels are significant markers in critical care, as they may signal insufficient tissue oxygenation and may reflect provider concern for sepsis, a condition that could lead to increased mortality risk and influencing patient treatment strategies. Bilirubin levels are equally important, offering insights into liver health, potential obstructions in bile ducts, and signs of hemolytic conditions, all pivotal for prompt diagnosis and treatment in the ICU setting. Finally, PaO2 is indispensable for gauging a patient's oxygenation status, playing a key role in managing respiratory support and determining the risk of respiratory failure.

Given our objective to examine various parameters such as epsilon, synthetic data generation methods, and different datasets, we chose $K = 3$. This creates a standard cube, striking a balance between the flexibility of PriSHA and the simplicity of conditional sampling in synthetic data generation. This K can be adjusted in practice using a separate validation set. To optimize $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we utilized the `hyperopt` package Bergstra et al. (2013), facilitating Bayesian optimization for optimal parameter suggestion.

Dataset	Method	ε	Standard AUC (%)	PriSHA AUC (%)	t-stat	p-val
EHR	DPGAN	0.1	85.45	85.52	0.3782	0.7095
EHR	DPGAN	1.0	85.31	85.34	0.3156	0.7557
EHR	DPGAN	2.0	85.14	85.70	2.8583	0.0101**
EHR	DPGAN	3.0	85.25	85.50	1.5105	0.1474
EHR	DPGAN	5.0	85.04	85.35	1.9779	0.0626*
EHR	DPGAN	10.0	85.26	85.50	1.1372	0.2696
EHR	DPGAN	20.0	85.45	85.70	1.6071	0.1245
EHR	DPGAN	100.0	85.44	85.47	0.1623	0.8728
EHR	PATE-GAN	0.1	85.60	85.33	-1.4450	0.1647
EHR	PATE-GAN	1.0	84.94	85.28	1.2849	0.2143
EHR	PATE-GAN	2.0	83.90	85.68	4.2384	0.0004***
EHR	PATE-GAN	3.0	82.79	85.64	5.6358	<0.0001***
EHR	PATE-GAN	5.0	85.53	86.27	2.5170	0.0210**
EHR	PATE-GAN	10.0	85.62	86.39	1.9700	0.0638*
EHR	PATE-GAN	20.0	85.51	86.39	3.7618	0.0013***
EHR	PATE-GAN	100.0	85.51	86.22	2.8335	0.0106**
eICU	DPGAN	0.1	74.96	76.34	2.38	0.0280**
eICU	DPGAN	1.0	75.32	76.33	1.36	0.1890
eICU	DPGAN	2.0	77.03	77.11	0.21	0.8372
eICU	DPGAN	3.0	76.07	77.13	2.10	0.0495**
eICU	DPGAN	5.0	75.95	77.00	2.32	0.0314**
eICU	DPGAN	10.0	75.00	76.33	2.00	0.0605*
eICU	DPGAN	20.0	75.28	75.97	1.73	0.1031
eICU	DPGAN	100.0	75.54	76.14	1.58	0.1304
eICU	PATEGAN	0.1	72.98	76.49	5.15	<0.0001***
eICU	PATEGAN	1.0	72.98	76.70	6.92	<0.0001***
eICU	PATEGAN	2.0	72.99	76.90	5.94	<0.0001***
eICU	PATEGAN	3.0	71.91	76.68	8.03	<0.0001***
eICU	PATEGAN	5.0	75.70	77.55	3.71	<0.0001***
eICU	PATEGAN	10.0	76.40	76.93	0.93	0.3655
eICU	PATEGAN	20.0	76.86	77.22	0.66	0.5176
eICU	PATEGAN	100.0	76.60	76.99	0.42	0.6852

Table 3: Standard Augmentation versus PriSHA using PATE-GAN and DPGAN across various ϵ values for EHR and eICU datasets. Significance levels are indicated as $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

6.4. Performance Evaluation

We proceeded to evaluate the performance of standard synthetic data augmentation by integrating $\mathcal{D}_A$ with $S(N;\epsilon,\delta)$ against that of PriSHA. To ensure robustness, 20 experiments were conducted, and a t-test was performed for each dataset, method, and ϵ. We focused on two methods, DPGAN and PATE-GAN where we used the `synthcity` and we fixed the δ to be $\frac{1}{M\sqrt{M}}$ as suggested by Qian et al. (2023) and explored a range of ϵ values: $0.1, 1, 2, 3, 5, 10, 20, 100$. The results are presented in Table 3.

Our findings reveal that, with the exception of PATE-GAN at $\epsilon = 0.1$ for EHR, all scenarios exhibited an average AUC enhancement. Specifically, PATE-GAN demonstrated superior performance in conjunction with PriSHA compared to DPGAN, yielding more significant improvements in various settings. For EHR data, we observed significant results at all $\epsilon \geq 2$. In contrast, for eICU data, significant improvements occurred at $\epsilon \leq 5$. This pattern implies that lower ϵ values lead to a decline in synthetic EHR data quality, producing results similar to noise. On the other hand, higher ϵ values in the eICU dataset seem to generate consistently high-quality synthetic data across the entire cube, avoiding the need for targeted hypercube segmentation.

7. Discussion and Future Work

We present a novel framework PriSHA for synthetic data augmentation in healthcare that improves performance while maintaining data privacy. PriSHA enhances model fairness across different subsets of patients by accounting for distribution shifts using Bayesian optimization. PriSHA also builds differential privacy guarantees, potentially decreasing the barrier to collaboration and data sharing across health systems.

The key strength of our approach lies in its ability to selectively utilize the most predictive hypercube of synthetic data, thereby enhancing the model training process. Our experimental results confirm that this targeted approach to data augmentation can significantly improve model performance, especially in contexts where privacy concerns restrict the use of real-world data.

PriSHA outperforms standard synthetic data augmentation in most comparisons, achieving better performance in 31 out of 32 scenarios. Notably, 17 of these scenarios showed statistically significant improvements. While the field of synthetic data augmentation for tabular data is still nascent, our research emphasizes the necessity for high-quality synthetic data rather than simply more synthetic data. Prior studies have shown that indiscriminate augmentation with all generated data can introduce noise to the downstream model, potentially degrading performance or failing to show improvement. We conclude that carefully selecting a portion of synthetic data—specifically, those parts that are supervised for a particular outcome—can significantly improve the efficacy of synthetic tabular data augmentation models.

We note that although PriSHA outperforms standard augmentation methods, its performance remains below that of an ideal scenario where data from both sources can be combined without privacy constraints. We believe that our results provide a baseline to build upon, encouraging future research to approach or exceed this utility threshold.

In future studies, we aim to refine our model using a meta-learning approach referenced in Hamad et al. (2023). This will enable us to harness the strengths of both DPGAN and PATE-GAN to improve the quality of private synthetic data. Additionally, if a diffusion-based method for tabular data synthesis with differential guarantee becomes practical, we plan to incorporate it to further increase the utility of our synthetic datasets.

An important future endeavor is to tailor our model for large language models (LLMs), particularly by harnessing LLMs trained across various medical institutions to mitigate distributional shifts. This adaptation aims to maintain the efficacy of LLMs amidst diverse clinical settings and ensure patient privacy while handling sensitive medical language data. Emphasizing privacy-preserving methods and the unique medical lexicon, this extension seeks to uphold the precision and dependability of LLMs in healthcare.

Acknowledgments The authors want to thank the anonymous reviewers for their constructive suggestions.

References

Anand Avati, Martin Seneviratne, Emily Xue, Zhen Xu, Balaji Lakshminarayanan, and Andrew Dai. Beds-bench: Behavior of ehr-models under distributional shift–a benchmark, 07 2021.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *ArXiv*, abs/2401.02524, 2024. URL https://api.semanticscholar.org/CorpusID:266818403.

James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, H. B. McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. *ArXiv*, abs/1902.01046, 2019. URL https://api.semanticscholar.org/CorpusID:59599820.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2022. doi: 10.1109/TNNLS.2022.3229161.

Gregory Caiola and Jerome P Reiter. Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.*, 3(1):27–42, 2010.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 787–792, 2017. doi: 10.1109/ICDM.2017.93.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/choi17a.html.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44 (3):837–845, 1988.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

K.E. Emam. *Risky Business: Sharing Health Data While Protecting Privacy*. Trafford Publishing, 2013. ISBN 978-1-4669-8050-1. URL https://books.google.com/books?id=D91RR3dDlr0C.

Bill Fitzgerald. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. In *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, 2015. URL https://api.semanticscholar.org/CorpusID:168483036.

Erin E. Fowler, Anders Berglund, Michael J. Schell, Thomas A. Sellers, Steven Eschrich, and John Heine. Empirically-derived synthetic populations to mitigate small sample sizes. *Journal of Biomedical Informatics*, 105:103408, 2020. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103408. URL https://www.sciencedirect.com/science/article/pii/S1532046420300368.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:237605416.

Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6(1):186, 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00927-3. URL https://doi.org/10.1038/s41746-023-00927-3.

Fadi Hamad, Shinpei Nakamura-Sakai, Saheed Obitayo, and Vamsi Potluru. A supervised generative optimization approach for tabular data. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 10–18, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702402. doi: 10.1145/3604237.3626907. URL https://doi.org/10.1145/3604237.3626907.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019. URL https://api.semanticscholar.org/CorpusID:56657912.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.04.053. URL https://www.sciencedirect.com/science/article/pii/S0925231222004349.

Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017. URL https://api.semanticscholar.org/CorpusID:5051282.

Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–6, 2019.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:53342261.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Peter Kairouz, H. B. McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2019. URL https://api.semanticscholar.org/CorpusID:209202606.

Aki Koivu, Mikko Sairanen, Antti Airola, and Tapio Pahikkala. Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association*, 27(11):1667–1674, Nov 2020.

Jakub Konecný, H. B. McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *ArXiv*, abs/1610.02527, 2016. URL https://api.semanticscholar.org/CorpusID:2549272.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm:

modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zheng Li, Yue Zhao, and Jialin Fu. Sync: A copula based framework for generating synthetic data from aggregated sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 571–578. IEEE, 2020.

Dionysis Manousakas and Sergul Aydore. On the usefulness of synthetic tabular data generation. In *ICML 2023 Workshop on Data-centric Machine Learning Research (DMLR)*, 2023. URL https://www.amazon.science/publications/on-the-usefulness-of-synthetic-tabular-data-generation.

H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016. URL https://api.semanticscholar.org/CorpusID:14955348.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 381–405. PMLR, October 2019. URL https://proceedings.mlr.press/v106/nestor19a.html. ISSN: 2640-3498.

Skyler Norgaard, Ramyar Saeedi, Keyvan Sasani, and Assefaw H. Gebremedhin. Synthetic sensor data generation for health applications: A supervised deep learning approach. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1164–1167, 2018. doi: 10.1109/EMBC.2018.8512470.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

TJ Pollard, AEW Johnson, JD Raffa, LA Celi, RG Mark, and O Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178. URL https://www.nature.com/articles/sdata2018178.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023. URL https://arxiv.org/abs/2301.07573.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu Galtier, Bennett A. Landman, Klaus H. Maier-Hein, Sébastien Ourselin, Micah J. Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and Manuel Jorge Cardoso. The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 2020. URL https://api.semanticscholar.org/CorpusID:212747909.

Veit Sandfort, Ke Yan, Perry J Pickhardt, et al. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9:16884, 2019. doi: 10.1038/s41598-019-52737-x. URL https://doi.org/10.1038/s41598-019-52737-x.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81(arXiv:2106.03253), 2021. URL http://arxiv.org/abs/2106.03253.

Liang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739, 2018.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

J. Yoon, M. Mizrahi, N.F. Ghalaty, et al. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine*, 6:141, 2023. doi: 10.1038/s41746-023-00888-7. URL https://doi.org/10.1038/s41746-023-00888-7.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), oct 2017. ISSN 0362-5915. doi: 10.1145/3134428. URL https://doi.org/10.1145/3134428.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan+: Enhancing tabular data synthesis. *arXiv preprint arXiv:2204.00401*, 2022.

Appendix A. Feature Importance

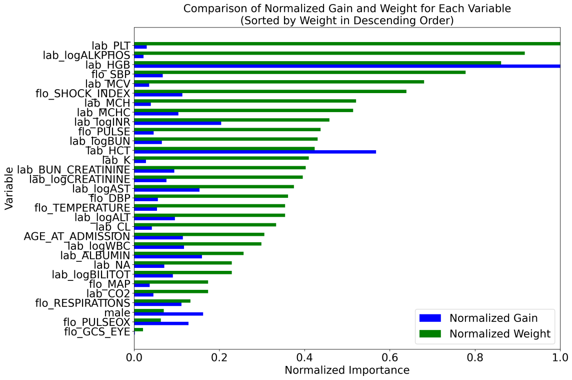

Figure 3: Feature importance of electronic health record data used for the predictive model.

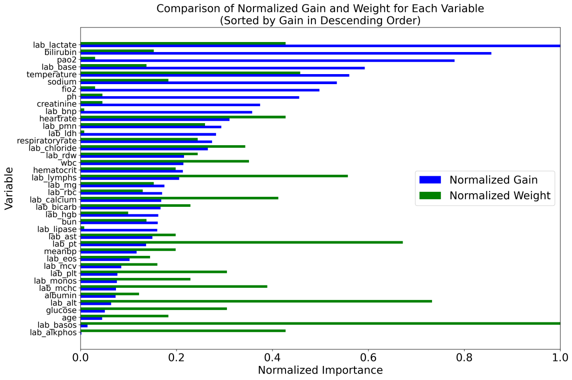

Figure 4: Feature importance of eICU data

Appendix B. eICU Preprocessing

B.1. Sampling

To preprocess the eICU data in $\mathcal{D}_A$ and $\mathcal{D}_B$ with distribution shift, we sampled the original eICU data using the age using a logistic function

$$f(x) = \frac{1}{1 + e^{-(x-a)/15}}$$

where $a = 40$ is the age threshold. The resulting histogram is observed in Figure 5 resulting in a datasets with different distribution.

B.2. Handling missing data

While EHR data does not contain missing data, to address the missing data in the eICU dataset, we employed Multivariate Imputation by Chained Equations (MICE), utilizing Bayesian Ridge Regression as the underlying imputation model. Bayesian Ridge Regression, which incorporates Bayesian inference into a linear regression framework, is particularly suitable at handling situations with limited data or high uncertainty, making it a suitable choice for imputation. However, we noted challenges with variables that had an excessive number of missing observations (more than 50% of observations missing), leading to unreliable results. To mitigate this, we first imputed such variables with the middle value from their reference ranges, based on the assumption that missing lab values might indicate a lack of necessity for tests, hence a likely healthy status. This approach posits that imputing a 'healthy' reference value is a reasonable estimate for these instances. Following this initial step, we conducted a more extensive MICE imputation using Bayesian Ridge Regression, thereby enhancing the completeness and accuracy of the dataset with a two-step imputation strategy.

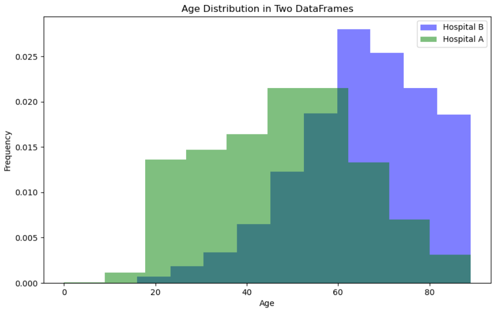

Figure 5: Comparative Histogram of Ages for Datasets $\mathcal{D}_A$ and $\mathcal{D}_B$ within the eICU data

Appendix C. Analyzing Performance with Different Values of ϵ

In the presented line graphs, we evaluate two distinct data augmentation methods and their influence on model performance, as measured by the Average

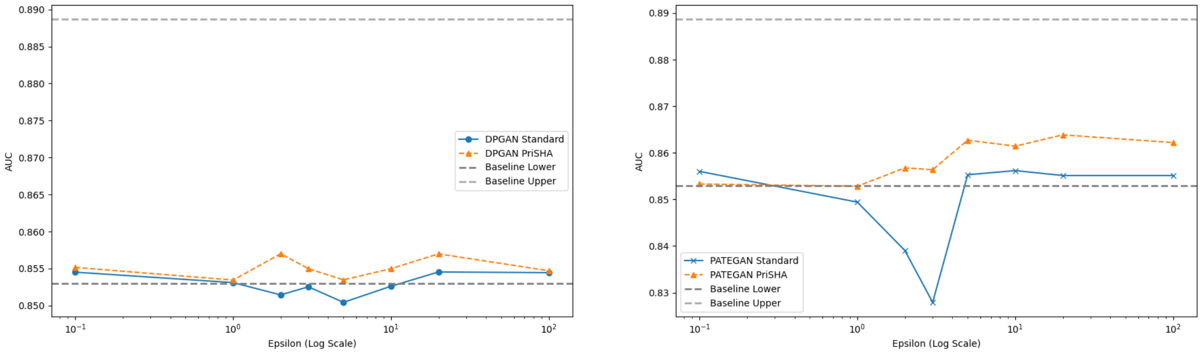

Figure 6: Average External Validation AUC Over 20 Simulations: Left - DPGAN in PriSHA vs. Standard Augmentation, Right - PATE-GAN in PriSHA vs. Standard Augmentation for EHR data across different values of ϵ

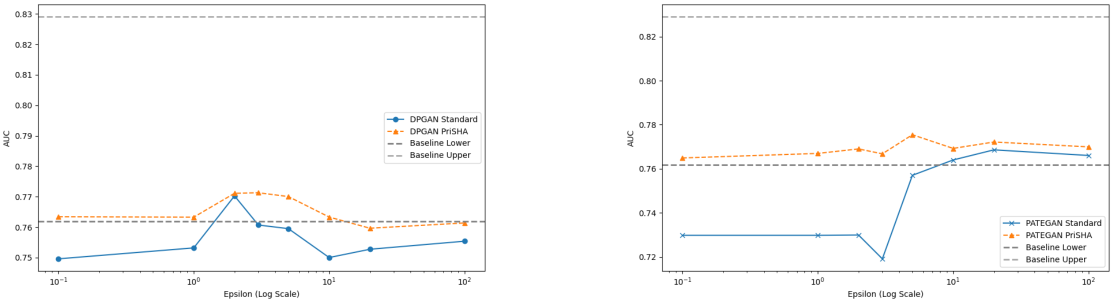

Figure 7: Average AUC Over 20 Simulations: Left - DPGAN in PriSHA vs. Standard Augmentation, Right - PATE-GAN in PriSHA vs. Standard Augmentation for eICU data across different values of ϵ

External Validation AUC across 20 simulations for both EHR (Figure 6) and eICU (Figure 7). The blue line indicates the model's performance with standard data augmentation techniques, whereas the orange line shows performance improvements when utilizing PriSHA. The lower horizontal lines in each graph serve as benchmarks, representing models trained solely on $\mathcal{D}_A$, thereby establishing a baseline. In contrast, the upper horizontal lines illustrate the enhanced performance of models trained on the combined dataset $\mathcal{D}_{AB}$. Our analysis reveals that for both DPGAN and PATE-GAN fall short of reaching the performance level achieved through augmentation with real data, suggesting that synthetic data, even at high ϵ values, may not provide sufficient signal to significantly boost the model. However, PATE-GAN consistently improves the baseline performance across a range of privacy levels (ϵ), highlighting that PriSHA can effectively enhance model performance by leveraging data from diverse distributions.

Appendix D. Analyzing Performance with Different Values of K

In Figures 8 and 9, we examine how the downstream performance (measured by AUC), the variable K, and the Maximum Sample (MS) parameter influence each other for $\epsilon = 5$. MS serves as a threshold for rejection sampling, ceasing the sampling process when Bayesian optimization suggests a range challenging to generate. Specifically, MS indicates that the system

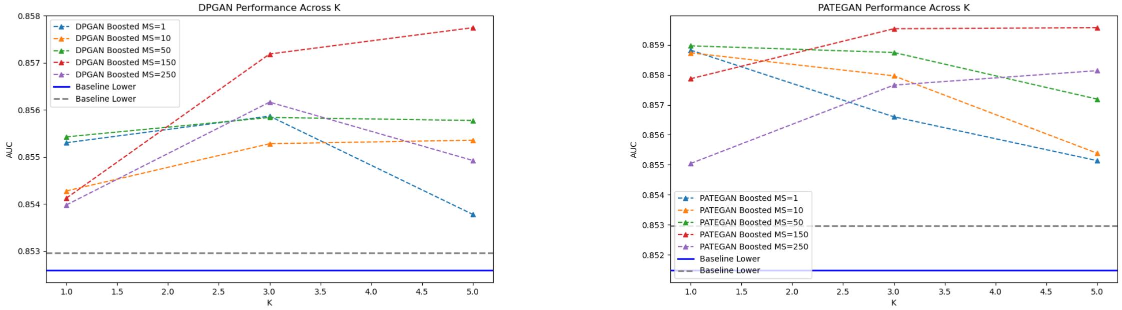

Figure 8: Average AUC Over 10 Simulations: Left - DPGAN in PriSHA vs. Standard Augmentation, Right - PATE-GAN in PriSHA vs. Standard Augmentation for EHR data across different values of K

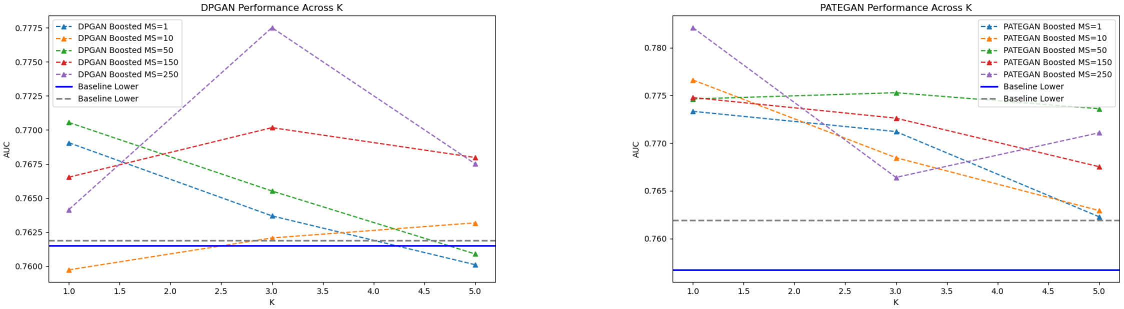

Figure 9: Average AUC Over 10 Simulations: Left - DPGAN in PriSHA vs. Standard Augmentation, Right - PATE-GAN in PriSHA vs. Standard Augmentation for eICU data across different values of K

attempts to sample N observations through rejection sampling in $MS * N$ draws. if it fails to obtain N samples after making $MS * N$ attempts, the process stops, and the gathered samples are used for augmentation. This mechanism acts similarly to early stopping, preventing the generation of data that significantly deviates from the target data $\mathcal{D}$ when applying the sampling strategy $S(N; \epsilon, \delta)$. The data indicates a rapid decline in model performance with increasing K values when MS is low, highlighting the difficulty in drawing samples that meet the specified constraints.