Individualized Data Generation for Electronic Health Records

Ghadeer Ghosheh
Prof. Tingting Zhu
Mansfield College
University of Oxford

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fullfillment of the requirements for the degree of Doctor of Philosophy

Doctor of Philosophy

Hilary 2025

Dedication

First and foremost, I thank God, the Almighty, to whom everything is dedicated, for His countless blessings and guidance throughout this journey.

With all my heart, I dedicate this thesis to my parents, Omar and Rawan, whose endless love, support, and sacrifices have been the foundation of everything I have achieved. To my husband, Mohammad, for his constant love and encouragement. To my siblings, Abdelkareem and Leen, for their support and belief in me. To my daughter, Lubna, whom I eagerly await as I write this thesis and who already fills my heart with joy.

This work is for you.

Acknowledgements

Although writing this DPhil thesis has sometimes felt like a lonely and long journey, this work would not have seen the light of day without the encouragement and contributions of many. I am deeply grateful to all who have supported and shaped this journey in countless ways.

First, my heartfelt gratitude goes to my supervisor, Professor Tingting Zhu, for her continual support, insightful feedback, and guidance. Our long and thoughtful meetings shaped not only the philosophy of this work but also my growth as a researcher, and I cannot be more thankful for her patience, generosity, and mentorship.

I also sincerely thank my collaborators and lab mates Jin Li, Zhiyao Luo, Munib Mesinovic, Soheila Molaei, Taha Certli, Professor James Sheppard, Professor Louise Thwaites, and many more whose shared enthusiasm have enriched this work. Thank you for the stimulating discussions and technical contributions that brought this research to life. To my transfer and confirmation examiners, Professor Tim Denison, Professor Lionel Tarassenko, and Professor Jens Rittscher, thank you for your valuable feedback and encouragement, which helped sharpen the focus of this thesis. I would also like to thank my thesis examiners, Professor Haiping Lu and Professor Xiaowen Dong, for their thoughtful feedback and suggestions, which strengthened the final version of this thesis.

Special thanks go to all friends from various labs at Oxford's Old Road Campus Research Building, whose uplifting conversations brought perspective and depth into this journey.

To my father, Dr. Omar Ghosheh, my role model and mentor, I am deeply grateful for your wisdom, support, immense faith in me, and the countless brainstorming sessions that always grounded me. To my mother Rawan, my husband Mohammad, and my siblings, Abdulkareem and Leen, you have always been my steady foundation through every step of this path.

My sincere gratitude goes to all of you.

Abstract

Personalized medicine is emerging as a new direction of research, in which patients are holistically examined as "individuals" rather than using symptoms or diagnoses based on the general population to determine the optimal treatment Electronic health records (EHRs) have opened the door to longitudinal data from real-world patients for a wide range of research, such as the creation of clinical decision support systems for different medical applications. Furthermore, the wealth of data collected in EHRs presents a great source for developing approaches in medicine that tailor the treatment to best suit each individual patient.

Data-driven approaches, specifically machine learning models, have shown promising results in various clinical applications, especially when using time-series EHR data. However, EHRs tend to be multivariate, highly missing, and irregularly sampled. Therefore, the success of machine learning models for personalized medicine is highly dependent on how the EHR data are represented, conditioning on the time-varying physiology, treatments, and missing values in the data. The existing body of work on personalized medicine mostly emphasizes the utilization of observable data. However, this approach tends to overlook the crucial aspect of missingness, which is essential for achieving true individualization and obtaining meaningful insights.

Another area of research in personalized medicine is estimating individualized treatment effects (ITE) from observational electronic health records (EHR) data, where various machine learning approaches estimate the effect of treatment at the individual level. ITE estimation aims to generate individualized potential patient responses to various treatments, allowing for better treatment allocation that best fits each patient. Despite the potential demonstrated by the various proposed methodologies, most of the works make unrealistic assumptions about the data or assume that the data have no missingness, which is not true in real-world EHRs.

This work proposes the use of deep generative models to generate individualized patient data to drive personalized medicine applications. Our work uses the individualized generative capabilities of deep generative models to generate data to impute missing values in time-series EHRs and generate new features that were never measured for the patient. We also work on generating counterfactual patient outcomes, indicating individualized potential patient responses to various treatments. To achieve this, our contributions include novel time-series data modeling techniques for representing individualized missingness and to generate missing-value imputations and completely missing features. Furthermore, we propose several contributions in the field of ITE estimation, where we're the first work to generate potential counterfactual estimates from missing time-series data. Furthermore, we propose the first work to estimate ITE for multiple outcomes from observational data. The proposed research aims to help utilize the wealth of data to produce applicable knowledge, paving the way for personalized medicine, that best fits each patient on the individualized level.

Contents

List of Figures

List of Tables

Table of Acronyms

Table 1: Table of Acronyms

Acronym	Definition
AKI	Acute Kidney Injury
ATE	Average Treatment Effect
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision-Recall Curve
BMI	Body Mass Index
CDSS	Clinical Decision Support System
CPRD	Clinical Practice Research Datalink
DBP	Diastolic Blood Pressure
EHR	Electronic Health Records
ELBO	Evidence Lower Bound
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPLVM	Gaussian Process Latent Variable Model
HiRID	High-Resolution ICU Dataset
HIPAA	Health Insurance Portability and Accountability Act
ICU	Intensive Care Unit
ITE	Individualized Treatment Effect
IGNITE	Individualized GeNeration of Imputations in Time-series EHRs
LSTM	Long Short-Term Memory
LR	Logistic Regression
MAE	Mean Absolute Error
MICE	Multiple Imputation by Chained Equations
MIMIC-III	Medical Information Mart for Intensive Care III
RNN	Recurrent Neural Network
RMSE	Root Mean Square Error
SSI	Sequential Strong Ignorability (causal assumption)
SSSI	Sequential Single Strong Ignorability
STRATIFY	Primary Care Dataset for Treatment Effects Estimation
TIA	Transient Ischemic Attack
TIUC	Time-Invariant Unobserved Confounding
VAE	Variational Autoencoder

List of Notations

Symbol	Description
A	Treatment assignment variable
A_i	Treatment assignment for individual i
ATE	Absolute Treatment Effect
D	Discriminator function in a Generative Adversarial Network
$Enc(x)$	Encoded representation of x
G	Generator function in a Generative Adversarial Network
$G_a(X;\theta_a)$	Treatment Model function
$G_y(X;\theta_y)$	Outcome Model function
$\mathcal{E}$	ITE estimator function
$\mathcal{L}_{\text{adv}}$	Adversarial loss function
$\mathcal{L}_{\text{contrast}}$	Multi-task contrastive loss for outcome consistency
$\mathcal{L}_{\text{ELBO}}$	Evidence Lower Bound loss
$\mathcal{L}_{\text{factual}}$	Factual prediction loss using binary cross-entropy
$\mathcal{L}_{\text{ITE}}$	Loss function for Individual Treatment Effect estimation
$\mathcal{L}_{\text{MMD}}$	Maximum Mean Discrepancy loss to balance treatment groups
N	Total number of patients in the dataset
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$P(X)$	Probability distribution of X
W_1	First weight matrix for attention transformation
W_2	Second weight matrix for attention transformation
$\mathbf{A}$	Treatment assignment matrix
$\mathbf{c}_t$	Context vector in the attention mechanism
$\mathbf{H}$	Hidden state matrix in sequential models
$\mathbf{M}$	Binary missingness mask matrix indicating observed and missing entries
$\mathbf{O}$	Observation mask indicating missing values
$\mathbf{T}$	Time steps in the dataset
$\mathbf{X}$	Matrix of input patient data
$\mathbf{X}_{i,t}$	Covariate matrix for patient i at time t
$\mathbf{Y}_{i,t}$	Outcome matrix for patient i at time t
$\mathbf{Z}$	Latent representation matrix for multiple samples
$\mathbf{x}$	Vector representation of input features
$\mathbf{y}$	Outcome vector for patients' labels
$\mathbf{z}$	Latent noise vector
$\hat{\mathbf{x}}$	Imputed or reconstructed data vector

Symbol	Description
ITE_i^m	Individual Treatment Effect for outcome m
∇	Gradient operator
θ_{shared}	Parameters for shared representation in multi-task learning

Chapter 1

Introduction and Background

The rapid growth of data-driven healthcare has opened new avenues for personalized medicine, where medical decisions and treatments are tailored to individual patient profiles. This shift marks a departure from generalized treatment guidelines towards an approach that considers unique characteristics such as genetics, medical history, and lifestyle. The ability to customize medical interventions not only improves patient outcomes, but also reduces unnecessary treatments and optimizes healthcare resources.

At the heart of personalized medicine lies the Electronic Health Record (EHR), a digital repository that captures longitudinal patient data, including diagnoses, treatments, lab results, and demographic information. EHRs have great potential to support personalized healthcare, allowing analysis of individual and population-level health trends. However, despite their promise, EHRs are often incomplete and irregular, with significant amounts of missing data due to non-standardized data entry, patients skipping appointments, or tests not being performed. This presents the following critical challenge.

1.1 Fundemental Topics

1.1.1 Electronic Health Records (EHRs) Fundamentals

In medical practice, healthcare professionals use EHRs to record and capture various forms of data about a patient during an encounter. Like paper records, EHRs store hospitalization and patient-level data such as demographics, comorbidities, medical history, vital signs, laboratory tests, prescribed medications, administered interventions, diagnosis, and clinical outcomes [22]. The nature of each of these types of data differs, resulting in multiple types of EHR data. Structured EHR data can

be presented in tabular or time-series formats. The tabular data store information that represents the patient's encounter, such as demographic features and aggregated mean vital signs, where each sample has one value for each feature. In contrast, time-series data presents a record of data points indexed in time order, which could be used to present disease progression over time as seen in longitudinal data [106] or even short-term records as seen in vital signs [233].

The variables recorded in each data type can be discrete, categorical, or continuous. Discrete variables represent values that can be obtained by counting and stored as integers such as age or number of visits per month. Categorical variables, on the other hand, are used when there is a finite number of categories, such as sex or ethnicity. Lastly, continuous variables are variables with values obtained by measurement and are not limited to whole numbers. Examples of continuous variables can be seen in many laboratory tests and vital signs such as albumin, body temperature, and total cholesterol. It should be noted that different types of EHR data usually coexist in the same patient record. For example, a patient might have both tabular and time-series data recorded for the same visit. This heterogeneous nature of EHRs often results in complexity in terms of their analysis, modeling, and use for machine learning purposes [51, 263]. EHR data can be recorded in different settings and stages

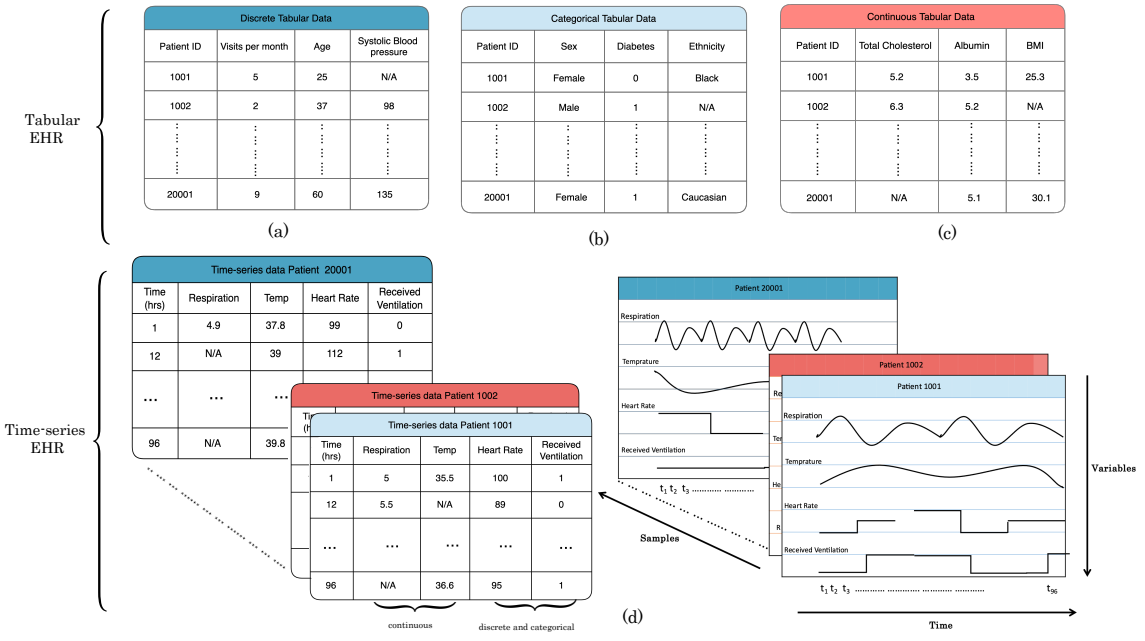


Figure 1.1: The two main types of EHR data, tabular and time-series are shown in their various forms. Discrete, categorical and continuous tabular data are shown in (a), (b) and (c), respectively. Time series data is shown in (d), where the record is shown on the left and a corresponding plot of the data is shown on the right.

of patient encounters or observations. During a hospital visit, a patient encounter can be classified as inpatient or outpatient, where the first requires hospitalization and admission, while the latter does not. For an inpatient encounter, a patient could go through various units within the same facility, depending on clinical status [278], the availability of human and material resources [60], or hospital capacity [72]. At the beginning of a hospital presentation, patients can be referred to emergency departments where an initial diagnosis and interventions are performed [241], where the focus is to admit the patient and then assess the patient according to medical need. In general inpatient units, patients receive regular laboratory tests, vital signs checks, treatment administration, and other necessary procedures as requested by the doctor. Patients who deteriorate or those whose cases require greater care are admitted to the Intensive Care Unit (ICU), where the data tend to be collected frequently because the patient is under close monitoring. Data collected in ICUs are usually referred to as critical care data [114]. The other type of EHR data is that of outpatient encounters, where the data collected is for patients who were not admitted to the hospital, as seen in the case of specialist consultations [103] and visits to general practitioners. The nature of outpatient data varies across countries, depending on the availability of primary care and the need for referrals to get a specialist consultation.

1.1.2 Understanding missingness in EHRs

Despite the great promise shown in many applications, the current literature on personalized medicine describes methods that focus only on observed data [37, 257] and often overlook the "individualized" and informative nature of unobserved or "missing" values in the data. In clinical practice, EHRs tend to be highly missing and sampled irregularly [166], where missingness can reach 80% in ICU settings (see MIMIC [121] and eICU [197]), despite continuous monitoring of patients. Therefore, there is a pressing need to learn proper patient representation and analysis of data missingness to improve ML models for better patient outcomes.

Although there are existing works for handling missingness in time-series EHRs, most of them assumed missingness was in the context of Missing at Completely Random (MCAR) [275, 31, 75], which may not be true in reality. In the case of MCAR, it is assumed that the missingness of the data is not related to the observed or unobserved data [165]. Such a strong assumption disregards any information related to the frequency and the reasons for recording, which might be related to the patient's health status. For example, missing values in time-series EHRs can occur due to various reasons such as machine and recording errors, irregular sampling and inconsistent

medical visits [139], or even high-cost and dangerous acquisition of information such as invasive or radiology procedures [27, 135]. In addition, the measurement frequency could also be related to factors such as patient severity and deterioration [7, 85], lack of medical need [65], bias, and quality of care [258, 260], all of which vary from one patient to another. This highlights the importance of representing the individualized missingness in personalized models. Recent ML models for healthcare have shown that the use of binary missingness masks to indicate the presence of observations as input features can result in equivalent performance to those built on the observed values of clinical features [225, 36], indicating the predictive value of missingness patterns. However, despite the utility and predictiveness of binary missingness masks, they are simplistic and lack indications of the patterns and frequencies of individual-level missingness for each patient.

In this thesis, we highlight new insights into missingness patterns in EHRs, which require a new perspective on representing missingness beyond binary masks. In time-series EHRs collected in hospitals, the measurement frequency of a variable over time often indicates the patient's underlying state. We refer to the missingness of observations of a variable in a patient record as feature-wise missingness. Specifically, some lab values are only measured for severe patients to differentiate clinical complications, such that a single measurement of a feature is enough to diagnose the patient. For example, cardiac troponins, sensitive biomarkers of myocardial injury, are not ordered for all patients in the ICU but are ordered for those suspected of having a cardiac-related diagnosis or complication [59]. Hence, a missing cardiac troponin recording can indicate that the patient is not suspected of developing a cardiac complication.In such cases, representing feature-wise missingness over timesteps via a binary mask would miss the opportunity to represent patterns in an individualized way.

Understanding the difference between general missingness in a patient record and feature-wise missingness for some variables is crucial to recognizing the implications of imputation. For example, filling in the missing cardiac troponin values for healthy patients by the population mean will result in an imputation value above the healthy range, since all available measurements are for patients suspected of having MI. Such imputations would result in a wrong patient representation and falsely indicate patient severity for healthy patients, resulting in a flawed data distribution. In Figure 1.2, we present an example in which population-based imputation methods do not impute data to represent a patient on a personalized level. Due to the limitations of such simple imputation techniques in handling individualized patterns in missingness

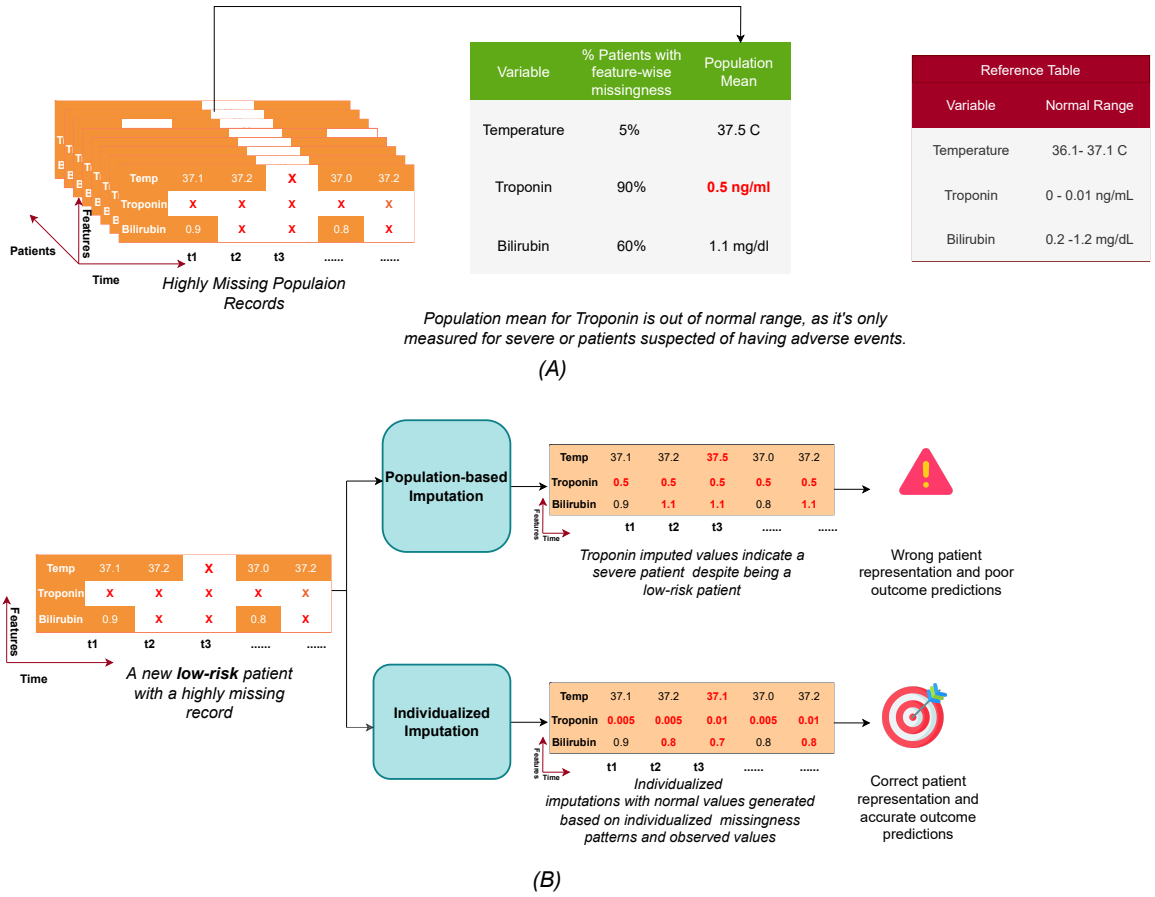


Figure 1.2: An example showing the impact of different imputation approaches for time-series EHRs with feature-wise missingness. In (A), highly missing EHRs are shown with population-level statistics and reference tables. In (B), the impact of individualized imputation compared to population-based imputation is shown to affect patient representation and predictive modeling, respectively.

and introducing bias in the data distribution, we believe that new approaches to representing missingness in predictive models can help generate high-quality imputations and build robust and high-performance predictive models. Conditional and generative imputation models [275, 75, 63] have advanced beyond traditional population-level approaches by generating more realistic imputations by learning the underlying population distribution in the observed data. However, these models often fail to take full advantage of the informative patterns inherent in missing data. Recent innovations in imputation focus on individualized approaches that consider both observed and missing data patterns specific to each patient. For example, in our work IGNITE [86] we introduce a frequency-based individualized Missingness Mask (IMM) that transforms a binary mask into a patient-specific representation, capturing the frequency of feature measurements over time. This allows the model to effectively account for personalized missing patterns. Similarly, Si et al. [229] proposed a Bayesian profiling method designed to capture patient-specific trends, particularly useful in datasets where trajectories, such as those for diabetes progression, vary significantly between individuals. Another notable framework, PRISM [289], leverages prototype patient representations for imputations and incorporates a feature confidence learner module to evaluate the reliability of each feature based on its missing status, enhancing imputation accuracy in sparse datasets. These approaches collectively highlight the shift toward personalized imputation methods, demonstrating the potential to improve data quality and predictive modeling in settings characterized by sparse and irregularly sampled EHRs.

We believe that insights related to feature-wise missingness and other real-world implications of missingness would open new doors to explore directions for better representation in predictive modeling. To this end, the imputation and learning missingness representations in the EHRs are beyond a preprocessing step, making the field incomplete without proper handling and representation of missingness in the data.

1.1.3 The Role of Deep Generative Models in Enhancing EHR Data Quality

Deep generative models have emerged as a disruptive tool to improve the quality and completeness of EHRs, addressing critical issues such as missing data, small datasets, and imbalanced records [85]. These models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), learn the underlying

structure of medical data and generate new, realistic records that capture the complexity and variability of real-world patient profiles.

In addition to handling missingness, it is worth emphasizing that deep generative models are also motivated by other challenges introduced earlier, including small sample sizes and privacy constraints, which are particularly relevant in low-resource settings and sensitive health domains.

One of the key applications of deep generative models is imputing missing patient data. In many cases, EHRs suffer from incomplete information due to missed appointments, incomplete tests, or incomplete patient records. When dealing with small datasets, traditional imputation techniques, such as statistical or simple interpolation methods, often fail to provide accurate and comprehensive solutions. However, deep generative models offer a more nuanced approach by estimating missing patient details based on patterns observed in similar patient profiles. These models ensure that imputed data is not only plausible but also maintains the internal consistency and variability found in actual medical records, which is especially crucial when working with small or sparse datasets.

In addition to filling in missing data, deep generative models also play a pivotal role in generating individualized patient data, which is essential for personalized medicine. These models can create synthetic patient records that best represent a specific patient's characteristics, such as their medical history, demographics, and treatment responses. For example, using GANs, it is possible to simulate patient profiles that mirror real patients with similar health conditions, enabling the generation of individualized synthetic data that can be used to predict patient trajectories, model treatment responses, or test potential interventions. This capability is particularly valuable in personalized medicine, where understanding how a treatment will affect a specific individual is critical. By generating realistic and individualized data, these models support a more precise estimate of individualized treatment effects (ITE), helping to tailor interventions to the unique needs of each patient, even when real-world data are incomplete.

Deep generative models also address the common problem of data scarcity and imbalanced datasets in EHRs. In many cases, certain groups of patients or medical conditions are underrepresented in the data, which can bias the predictive models and lead to inaccurate conclusions. By generating new synthetic samples that reflect the characteristics of these underrepresented groups, deep generative models help balance the dataset and improve the fairness and accuracy of downstream predic-

tions. This not only improves predictive modeling, but also ensures that treatment recommendations are equitable and applicable in diverse patient populations.

Furthermore, the ability of deep generative models to improve the quality of EHR data has far-reaching implications for downstream applications in personalized medicine. For example, these models can generate complete patient trajectories in time-series data, filling in gaps where real data is irregular or sparse. This allows for a more comprehensive analysis of the health of a patient over time, improving the accuracy of predictive models and treatment recommendations. In addition, these models provide a privacy-preserving solution by generating synthetic datasets that are statistically similar to real patient data but do not contain actual patient information. This enables healthcare providers and researchers to share data more freely for analysis and collaboration, without compromising patient privacy.

In summary, deep generative models are essential for improving the quality and completeness of EHRs. They offer solutions for imputing missing patient data, generating individualized patient profiles, and addressing data scarcity, all of which are critical to the advancement of personalized medicine. By allowing for more accurate, data-driven healthcare decisions tailored to individual patients, these models support the larger goal of transforming healthcare through personalized treatment and care.

1.1.4 Individualized Treatment Effects Estimation (ITE) from EHRs

In traditional medical research, Randomized Controlled Trials (RCTs) have long been the gold standard for estimating Average Treatment Effects (ATE)—the average impact of a treatment across a population[214]. Although RCTs provide valuable information, they are often limited in scope, time consuming [101], and expensive [214]. Furthermore, the results derived from RCTs may not always generalize well to the broader population, particularly in real-world clinical settings where patient profiles and treatment environments vary widely. This is where the use of EHRs becomes transformative, offering a rich source of real-world evidence (RWE) that can be used to estimate ITE on a scale and at a lower cost than RCTs.

ITE estimation focuses on predicting how a specific patient will respond to a treatment, given their unique characteristics and medical history, rather than determining a population-wide average effect. By leveraging EHRs, we can depart from generalized one-size-fits-all conclusions (as seen with ATEs) and move toward personalized healthcare that optimizes treatments for individual patients. EHRs allow us to observe large and diverse patient populations in real-world settings, capturing the

complexity and heterogeneity of medical data that RCTs often lack. However, EHRs come with inherent challenges, particularly due to missingness and irregularity in the data. In real-world practice, patients do not follow the highly controlled conditions of an RCT; they miss appointments, undergo different tests, or switch treatments, resulting in incomplete data records. Missing data can obscure the true effect of a treatment, leading to biased or incorrect estimates if not properly handled. This directly links back to the central research question: *"How can we improve the quality and completeness of EHRs to support personalized medicine, despite the challenges of highly missing and irregular data?"*

To overcome these challenges, modern machine learning techniques—such as causal inference models and counterfactual estimation—have been developed to estimate ITE from EHRs. These models can assess what would have happened if a patient had received a different treatment, effectively enabling counterfactual reasoning. Moreover, temporal models such as Recurring Neural Networks (RNNs) and transformers help capture the dynamic, time-dependent nature of patient records. By understanding how a patient's condition evolves over time, these models can provide more accurate, individualized predictions about the likely outcomes of various treatment options.

One major advantage of using EHRs to estimate ITE is the ability to generate real-world evidence at a fraction of the cost of running traditional RCTs. EHR data are collected during routine clinical practice, reflecting the actual conditions under which treatments are administered. These data can be retrospectively analyzed or used in prospective studies to validate treatment effects, providing clinicians with actionable insights based on real-world patient experiences. However, to fully leverage EHRs for ITE, it is crucial to address the issue of missing and incomplete data, as missingness data can distort treatment effect estimates and undermine the quality of real-world evidence. Solutions to handle missing data, such as deep generative models, become indispensable in this context.

Throughout this thesis, the term treatment effect" is used in the context of causal inference, specifically referring to the difference between potential outcomes under different treatments for the same individual. This follows the econometric framework of estimating counterfactuals. In contrast, the term 'treatment effect' in clinical practice can sometimes refer to predictive outcomes under a given treatment, without inferring causality. To avoid ambiguity, we consistently adopt the causal interpretation throughout this work. The Individualized Treatment Effect (ITE) refers to the difference in outcomes for an individual patient if they were treated versus if they were

not treated. Unless otherwise specified, all references to ITEs in this thesis pertain to causal estimands and are estimated using counterfactual modeling techniques.

1.2 Summary

The research question 'How can we improve the quality and completeness of EHRs to support personalized medicine, despite the challenges of highly missing and irregular data', centers on the concept of individualization. Personalized medicine relies on accurate patient-specific data, but EHRs often suffer from missing samples, missing variables, and missing patient records. To overcome these issues, novel machine learning methods, particularly deep generative models, provide promising solutions via missing value imputation, data synthesis, and estimation of treatment effect. A summary of the contribution areas as part of this thesis is presented below:

- **Imputing Missing Values**: Deep generative models help address the incomplete nature of EHRs by learning patterns from available data and imputing missing values. Whether individual observations are missing or entire patient records, these models enhance the completeness of EHRs, allowing more reliable, individualized predictions.

- **Generating Synthetic Patient Data**: Deep generative models can also create synthetic patient data, preserving privacy while enabling the exploration of personalized treatments. This supports individualization by simulating realistic patient scenarios and expanding the availability of diverse patient data for research.

- **Estimating Individualized Treatment Effects**: Estimating ITE from incomplete time-series data is vital for personalized medicine. By combining deep generative models with causal inference techniques, it is possible to predict how individual patients will respond to treatments, even when their records are incomplete or irregular. This ensures that healthcare decisions are tailored to each patient's unique history and conditions.

In summary, the methods proposed in the thesis aim to improve the completeness of the EHR, improve individual predictions, and support the transition to personalized, real-world healthcare that addresses the challenges of missing and irregular data in the EHR.

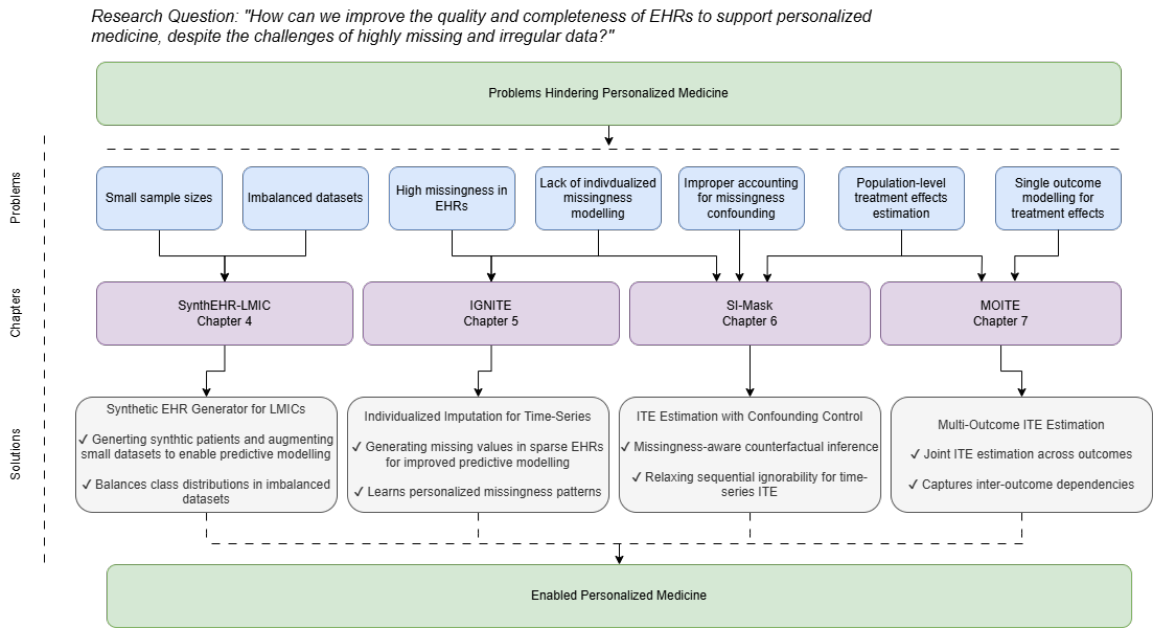


Figure 1.3: A Conceptual knowledge graph illustrating how specific methodological contributions address key challenges in Electronic Health Record (EHR) data to enable personalized medicine. The diagram is structured into four layers: (1) Problems hindering personalized medicine (top), including issues such as small sample sizes and high missingness in EHRs; (2) Methodological solutions developed in four dissertation chapters (middle layer in purple); (3) Specific solution strategies and innovations (below each chapter); and (4) The final outcome—enabled personalized medicine. Arrows denote the logical and methodological flow from problem identification through to solution implementation.

1.3 Thesis Structure

The structure of the thesis is organized as follows to systematically address the research question: *"How can we improve the quality and completeness of EHRs to support personalized medicine, despite the challenges of highly missing and irregular data?"* . The objectives of this thesis are as follows: (1) Review and synthesize existing deep generative modeling methods and their applications to EHRs for missing data imputation, ITE estimation, and synthetic data generation. (2) Introduce and analyze the structure and preprocessing of various EHR datasets. (3) Propose novel generative models to improve the quality and completeness of the EHR data. (4) Develop a new machine learning technique for estimating individualized treatment effects from sparse time-series data. (5) Provide a flexible framework for addressing missingness-related confounding in ITE estimation. We provide a summary figure showcasing how the various chapters contribute to the overall thesis goal.

Specifically, Chapter 2 provides a review of the literature summarizing current research on deep generative models, EHR data imputation, individualized treatment effect estimation, and synthetic data generation. Chapter 3 focuses on Data Description, detailing the datasets used, their clinical contexts, and data preprocessing methodologies. Chapter 4 explores Generative Models for Predictive Modeling, proposing novel methods to handle imbalanced and incomplete tabular EHR data to improve predictive accuracy. Chapter 5 introduces a new approach for Individualized Generation of Imputations in Time-series EHRs, enhancing data quality and downstream tasks such as adverse event prediction. Chapter 6 presents a Flexible Framework for Missingness-Related Confounding in ITE Estimation, ensuring robust treatment effect predictions even with significant data missingness. Chapter 7 tackles Multi-objective Individualized Treatment Effect (ITE) Estimation, offering a framework that considers multiple clinical outcomes simultaneously. Finally, Chapter 8 provides the Conclusion and Future Work, summarizing the contributions and discussing potential directions for further research.

1.4 Relevant Publications to the Thesis

- **Ghosheh, G.**, Gogl. M., Zhu, T. (2025). A Perspective on Individualized Treatment Effects Estimation from Time-Series Data. Journal of the American Medical Informatics Association, 2025; https://doi.org/10.1093/jamia/ocae323 [Chapter 6 & 7]

- **Ghosheh, G.**, Sheppard, J, Zhu, T. (2025) Multi-Objective Individualized Treatment Effects Estimation (ITE) for antihypertensive medications in five million patients longitudinal data. [In Submission]. [Chapter 7]

- **Ghosheh, G.**, Zhu, T. (2025). Relaxing Sequential Ignorability: A Flexible Framework for Accounting for Missingness-Related Confounding in Time-Series ITE Estimation. KDD 2025 [Under Review].[Chapter 6]

- **Ghosheh, G. O**., Li, J., & Zhu, T. (2025). Understanding Missingness in Time-series Electronic Health Records for Individualized Representation. arXiv preprint arXiv:2402.15730. NPJ Artificial Intelligence [Under Review] [Chapter 5]

- **Ghosheh, G. O.**, Li, J., & Zhu, T. (2024). IGNITE: Individualized GeNeration of Imputations in Time-series Electronic Health Records. arXiv preprint arXiv:2401.04402. NPJ Digital Medicine [Under Review] [Chapter 5]

- **Ghosheh, G. O.**, Thwaites C.L., Zhu, T. (2023). Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs). *MDPI Biomedicines*, 11(6), 1749. [Chapter 4]

- **Ghosheh, G.**, Li, J., Zhu, T. (2023). A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Computing Surveys*, 56(4), Article 63. [Chapter 4 & 5]

- **Ghosheh, G.**, Zhu, T. (2023). Synthesizing Electronic Health Records for Predictive Models in LMICs. *Practical ML for Developing Countries Workshop at the International Conference on Learning Representations (ICLR)*. [Chapter 4]

Other Publications via Collaboraions

- Molaei, S., Niknam, G., **Ghosheh, G. O.**, Chauhan, V. K., Zare, H., Zhu, T.,& Clifton, D. A. (2024). Temporal Dynamics Unleashed: Elevating Variational Graph Attention. Knowledge-Based Systems, 299, 112110.

- Molaei, S., Bousejin, N. G., **Ghosheh, G. O.**, Thakur, A., Chauhan, V. K., Zhu, T., & Clifton, D. A. (2024). CliqueFluxNet: Unveiling EHR Insights with Stochastic Edge Fluxing and Maximal Clique Utilisation using Graph Neural Networks. Journal of Healthcare Informatics Research, 8(3), 555-575.

- Ceritli, T., **Ghosheh, G. O**., Chauhan, V. K., Zhu, T., Creagh, A. P., & Clifton, D. A. (2023). Synthesizing Mixed-type Electronic Health Records using Diffusion Models. arXiv preprint arXiv:2302.14679.

- Chauhan, V. K., Zhou, J., Molaei, S., **Ghosheh, G**., & Clifton, D. A. (2023). Dynamic inter-treatment information sharing for heterogeneous treatment effects estimation. arXiv preprint arXiv:2305.15984.

Chapter 2

Literature Review

2.1 Overview

This chapter provides a comprehensive review of the existing literature on the application of advanced machine learning methods, particularly deep generative models, to improving the quality and usability of electronic health records (EHRs). As the healthcare industry moves toward personalized medicine, the ability to accurately analyze and interpret patient data becomes increasingly critical. However, real-world EHR data often suffer from challenges such as missing values, irregular data collection, and small sample sizes, which can hinder the development of effective predictive models. Addressing these challenges requires sophisticated approaches, and deep generative models have shown immense promise in this area.

The primary focus of this chapter is to explore and critically assess current methodologies used to improve the quality of EHR data. The review will focus on three main themes: (1) the use of deep generative models for synthesizing medical records, which helps augment data and preserve privacy, (2) approaches to imputing missing data, which is a pervasive issue in healthcare datasets and (3) methods for estimating treatment effects, with a particular focus on the individualized treatment effect (ITE) in the context of real-world EHRs. The chapter will provide an overview of foundational concepts, followed by a detailed discussion of state-of-the-art models, their applications, and limitations. This review not only sets the stage for the novel methodologies introduced in subsequent chapters but also highlights the gaps in current research that this thesis aims to address.

2.2 Deep Generative Models

2.2.1 Motivation for the Use of Deep Generative Models in Healthcare

Over the past decade, machine learning models have been shown to have great potential to support medical applications by using data collected in EHRs [263, 213]. Hospitals and medical providers are increasingly adopting and deploying EHR systems. In the US alone, 84% of hospitals adopted EHR systems in 2015, which is a 9-fold increase since 2008 [104]. The widespread recording of structured EHRs is paving the way for research opportunities in healthcare applications, such as patient stratification [216], drug repurposing [40], public health surveillance [22], as well as the novel discovery of disease mechanisms and correlations as seen in the early applications of COVID-19 [56]. EHRs also provide a valuable asset in the development of patient-specific and data-driven clinical decision support systems (CDSS) for diagnostic, prognostic, healthcare cost containment and workflow improvement applications [239, 145, 223]. However, full utilization of the wealth of EHR data in such applications is impeded by several challenges, including data sharing and privacy concerns [129], where data protection guidelines and regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [74] in the United States, and the General Data Protection Regulation (GDPR) [251] in Europe have detailed controlling measures that prevent direct access to much of the data for patient privacy purposes. Other data-specific challenges that make EHR processing burdensome include class imbalance [215], data missingness [166], noise [133], heterogeneity [51] and irregular sampling [237]. To mitigate these challenges, deep generative models have been proposed to generate synthetic data [38], notably variational autoencoders (VAE) [137], and Generative Adversarial Networks [92].

2.2.2 Generative Adversarial Networks (GANs)

Principles and Architecture

Since the introduction of GANs in 2014 [92], they have shown great potential to generate realistic data for various applications. The working principle of GANs essentially involves the training of a pair of deep neural networks in competition with each other [53]. The first neural network, the *generator* G, takes a noise vector $\mathbf{z}$ from latent space as input and generates synthetic samples $G(\mathbf{z})$ [53]. The other neural network, the *discriminator*, D is given both real $\mathbf{x}$ and generated samples $G(\mathbf{z})$, and

trained to discriminate between real and synthetic [92]. The discriminator outputs a vector of probability predictions of whether the input samples are real or synthetic. Both the generator and the discriminator are fine-tuned using the discriminator's output through backpropagation, as shown in Figure 2.1. Training involves both finding the parameters of a discriminator that maximize its classification accuracy and finding the parameters of a generator that minimize the discriminator's ability to tell the real and synthetic samples apart [92]. In other words, the objective loss function for the original GANs is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))]$$

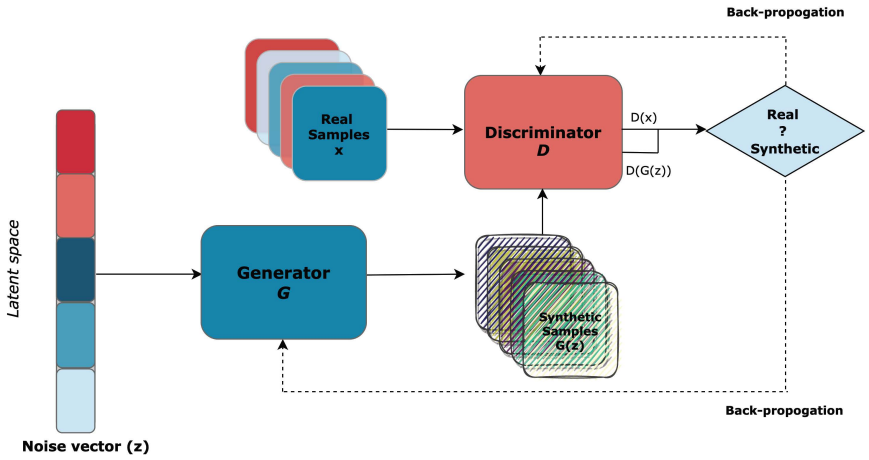


Figure 2.1: An overview of the architecture of GANs showing the function of both the *generator* and *discriminator* neural networks. The generator takes a noise vector $\mathbf{z}$ as input and outputs the synthetic data. The discriminator is trained to distinguish between the real and synthetic data. Both the *generator* and the *discriminator* are then fine-tuned by back-propagation.

GAN Variants

The original GAN framework has inspired numerous extensions that have aimed to improve performance and flexibility for specific applications. Conditional GANs (CGAN) [182] introduced the concept of conditioning the generator on additional information, such as class labels, to guide the generation process. Deep Convolutional GANs (DCGAN) [199] incorporated convolutional layers to improve the quality of images generated from GANs. Furthermore, InfoGAN [39] added an interpretable latent space to provide greater control over generated features.

Recurrent GAN (RCGAN) [68] extended the GAN architecture to handle sequential data, which found applications in healthcare, particularly to generate synthetic EHRs. CycleGAN [288] and StarGAN [46] further extended GANs to domain adaptation and translation tasks. Each of these variants brought new strengths to the GAN framework, broadening its application areas.

GAN Training Challenges

Despite their popularity, GANs face significant training challenges. The most notable is *mode collapse*, where the generator learns to produce a limited variety of outputs, even when different inputs are provided [92]. Another common problem is vanishing gradients [10], where the discriminator becomes too strong, offering little useful feedback to improve the generator. Several techniques have been developed to address these challenges, including Wasserstein GAN (WGAN) [11] and its improved version [94], which use the Wasserstein distance to stabilize the training of GAN. Other strategies like mini-batch discrimination [212], noise injection, and unrolled GANs [179] have also been proposed to improve stability and diversity in GAN output. However, stabilizing GAN training remains an ongoing research focus.

2.2.3 Variational AutoEncoders (VAEs)

Variational Autoencoders (VAEs) [137] are another class of generative models that rely on probabilistic latent variable modeling. VAEs aim to learn a compressed latent space representation of the input data, from which new data points can be generated. Unlike GANs, which utilize a competitive framework, VAEs model the data generation process as probabilistic inference by assuming that the observed data are generated through latent variables.

VAEs work by approximating the true posterior distribution over latent variables through the use of an encoder-decoder structure. The encoder maps the input data to a latent space, while the decoder reconstructs the data from this latent representation. To train the model, VAEs employ a loss function called the Evidence Lower Bound (ELBO), which balances two objectives: the accuracy of data reconstruction and the regularization of the latent space to conform to a prior distribution, typically a Gaussian.

Mathematical Formulation

The generative process in VAEs assumes that there is a latent variable $\mathbf{z}$, drawn from a prior distribution $p(\mathbf{z})$, typically a Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$. The observed data $\mathbf{x}$ are generated from the conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$, where θ are the parameters of the generative model.

The key challenge in VAEs is to infer the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ over the latent variables $\mathbf{z}$ given the observed data $\mathbf{x}$. However, direct computation of the posterior is intractable, and VAEs instead approximate it using a variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$, where ϕ are the parameters of the encoder (also known as the inference network).

The training objective of VAEs is to maximize the likelihood of the observed data $\mathbf{x}$, which is equivalent to maximizing the marginal likelihood:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

However, this integral is intractable. Instead, VAEs maximize a variational lower bound, the Evidence Lower Bound (ELBO), which is given by:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

Here, the first term is the reconstruction likelihood, which ensures that the generated data from the latent variable $\mathbf{z}$ is close to the input $\mathbf{x}$. The second term is the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$, which regularizes the latent space.

Reparameterization Trick

The reparameterization trick is used to enable backpropagation through stochastic sampling of the latent variable $\mathbf{z}$. Instead of directly sampling $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$, we sample an auxiliary variable $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and express the latent variable as a differentiable function of ϵ:

$$\mathbf{z} = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon$$

where $\mu_\phi(\mathbf{x})$ and $\sigma_\phi(\mathbf{x})$ are the mean and standard deviation of the approximate posterior, and $\odot$ denotes element-wise multiplication.

This formulation allows the gradients of the ELBO with respect to the parameters θ and ϕ to be computed using standard backpropagation, enabling efficient training of VAEs using stochastic gradient descent.

Training Objective

The overall training objective for VAEs is to maximize the ELBO with respect to the parameters θ and ϕ:

$$\max_{\theta,\phi} \mathcal{L}(\theta, \phi; \mathbf{x}) = \max_{\theta,\phi} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

This balances the trade-off between reconstructing the input data accurately and regularizing the latent space to encourage smooth and meaningful latent representations.

In summary, VAEs provide a powerful framework for learning latent representations of data while maintaining a probabilistic generative process. However, they often struggle with producing sharp outputs compared to GANs, particularly in tasks such as image generation, where fine detail is important.

2.2.4 Related Deep Generative Models

Although GANs and VAEs are two of the most prominent deep generative models we focused on in this dissertation, other approaches, like diffusion models and transformers, have gained traction in the field. The diffusion models [110] map the data to a noisy latent space and gradually remove the noise to generate new samples. These models have recently demonstrated state-of-the-art performance, often surpassing GANs in image generation tasks [62]. However, diffusion models can be inefficient to train and may generate less private synthetic data than other models [33].

Transformers, originally introduced for natural language processing [249], have expanded to tasks involving image and tabular data generation [130, 13]. Although powerful, transformer-based models require large datasets and considerable computational resources, making them more challenging to train and optimize compared to VAEs or GANs.

In summary, while each of these generative models offers unique advantages, they also come with trade-offs. VAEs excel in capturing probabilistic representations of data, though they struggle with output quality. GANs generate high-quality output, but suffer from instability during training. Diffusion models and transformers provide additional alternatives, though they introduce new complexities in training and scalability. Each model serves different purposes, depending on the specific application requirements.

2.3 Applications of Generative Models for structured EHRs

The applications of GANs in the medical domain are very diverse, specifically in medical imaging. For example, GANs have been used for various radiology tasks that ranged from data augmentation to data segmentation and denoising [169, 284, 272]. However, much less work has been done on the use of GANs to generate realistic structured healthcare data, such as EHRs. The delay in the use of GANs for EHR data can be attributed to many data challenges, such as complexity, heterogeneity, and missingness [263]. Compared with other data modalities, such as images and text, which can be intuitively and visually evaluated for realism, it is difficult to assess the quality of the generated EHR data. In Table 2.1, we summarize the main works that used GANs for EHR applications and group them based on their target application. The main groups are (1) generation of diverse types of EHRs, (2) imputation of missingness and (3) treatment effect estimation. The works are reviewed in terms of the used models, task, dataset size, open access code, and data as well as evaluation components used to assess the quality of the synthetic data.

2.3.1 Generation of Diverse Types of EHRs

Generative Adversarial Networks (GANs) have been increasingly applied to synthesize electronic health records (EHRs), addressing challenges related to privacy, data imbalance, and data availability. Existing works can be broadly categorized into the generation of tabular EHRs, time-series EHRs, and heterogeneous EHRs comprising mixed data types. Early GAN frameworks focused on structured, discrete EHR data such as diagnosis and billing codes. *medGAN* [44] was among the first to integrate an autoencoder within the GAN architecture to model binary and count-based tabular data. Its extensions, including *medWGAN* and *medBGAN* [14], employed Wasserstein GAN with gradient penalty [94] and boundary-seeking loss [109] to enhance training stability and generation quality. *MC-medGAN* [29, 91] further improved the handling of multi-categorical variables using Gumbel-softmax sampling [118]. Other approaches targeted feature-level correlation modeling. *CorGAN* [245] leveraged convolutional autoencoders, while *GcGAN* [269] encoded treatment-disease relationships. *SMOOTH-GAN* [202] introduced a conditional generator with smoothed labels to represent disease stages more flexibly. Additionally, *EMR-WGAN* [287] removed the autoencoder to better handle rare clinical concepts and improve fidelity.

Table 2.1: Summary of the various uses of GANs for EHRs and comparison of target application, evaluation measures, medical datasets and open access.

		Problem		Evaluation							Medical Dataset		Open Access	
References	Year	Model	Task	DWS	LDS	JDS	IDRS	PP	DU	Qual	Dataset	Dataset size (N/R)	Dataset	Code
Generation of Diverse types of EHRs														
[44]	2017	medGAN	Generating discrete tabular EHR data	✓	✗	✗	✓	✓	✗	✓	MIMIC-III Sutter PAMF Sutter Heart failure Cohort	46,520/ NA 258,559/ NA 30,738/ NA	✓* ✗ ✗	✓
[68]	2017	RGAN, RCGAN	Generating continuous time-series EHRs	✗	✗	✓	✗	✓	✓	✓	Philips eICU	17,693/ NA	✓*	✓
[266]	2017	GAN for DLEs	Generating continuous time-series Drug Laboratory Effects (DLEs)	✗	✗	✓	✗	✗	✗	✓	Private New York EHR dataset	4,830/ NA	✗	✗
[276]	2018	RadialGAN	Leveraging multiple tabular datasets by using multiple GAN	✗	✗	✗	✗	✗	✓	✗	14 RCTs from MAGGIC	528-13279/ NA	✓*	✓
[14]	2019	medWGAN, medBGAN	Itegrating medGAN with WGAN-GP & BGAN for generating discrete tabular EHRs	✓	✗	✗	✓	✗	✗	✗	MIMIC-III NHIRD Taiwan	46,517 / 46,517 498,909 / 498,909	✓* ✗	✓
[42]	2019	WGAN	Generating heterogeneous discrete tabular EHRs	✓	✗	✓	✗	✓	✓	✓	NMDS	NA / 2,873,466	✗	✗
[254]	2019	SC-GAN	Generating continuous sequentially coupled time-series EHRs data for patient sate & medication dosage	✓	✗	✗	✓	✗	✓	✓	MIMIC-III	29,278 / NA	✓*	✗
[287]	2020	EMR-WGAN, EMR-CWGAN	Improved EHR generation training stability and evaluation	✓	✓	✓	✓	✓	✗	✗	VUMC Synthetic Derivative	2,246,444 / NA	✗	✗
[91]	2020	MC-medGAN	Generating multi-categorical tabular EHRs	✓	✓	✗	✓	✓	✗	✓	SEER's research dataset	NA / 366,631	✓*	✓
[268]	2020	HGAN	Generating Heterogeneous tabular EHRs while preserving feature constraints and inter-dimensional dependencies	✓	✓	✓	✓	✓	✗	✓	VUMC Synthetic Derivative	928,089 / NA	✗	✗
[269]	2019	GcGAN	Generating tabular EHRs while preserving grouped coorlations	✓	✗	✗	✓	✗	✓	✗	Private Pediatric EHR	NA /17,000	✗	✗
[245]	2020	CorGAN	Correlation-Capturing generation of continuous and discrete tabular EHRs	✓	✗	✗	✓	✓	✓	✗	MIMIC-III UCI Epileptic Seizure Recognition	NA / 46,000 500 / 11,500	✓* ✓	✓
[202]	2020	SmoothGAN	Generating tabular EHRs with smooth conditions	✓	✗	✗	✓	✓	✓	✓	Cerner HealthFacts	NA /47,412	✗	✗
[286]	2021	SynTEG	Generating discrete time-series EHRs of diagnostic events	✓	✓	✗	✗	✓	✓	✗	VUMC Synthetic Derivative	NA / 2,187,629	✗	✓
[147]	2020	DAEE	Generating time-series EHRs of discrete diagnostic codes	✗	✗	✓	✗	✓	✗	✓	MIMIC-III UT-Physicians	7,537 / 19,993 13,025 / 85,845	✓* ✗	✓
[149]	2021	EHR-M-GAN	Generating mixed-type time-series EHRs	✓	✗	✓	✓	✓	✓	✓	MIMIC-III Philips eICU HiRID	28,344 / 28,344 99,015 / 99,015 14,129 / 14,129	✓* ✓* ✓*	✓
[259]	2021	WGAN	Federated learning for GAN for discrete, binary EHRs	✓	✗	✓	✗	✓	✗	✓	MIMIC-III	NA / 46,520	✓*	✗
Semi-Supervised Learning and Data Augmentation														
[35]	2017	ehrGAN,SSL-GANs	Augmenting data for imbalanced SSL tasks using EHRs of sequences of diagnosis codes	✗	✗	✗	✓	✗	✓	✓	Private insurance dataset	218,680 / 14,969,489	✗	✗
[152]	2018	GAN for SSL	SSL based GANs for detecting rare diseases in unlabelled tabular discrete & continuous EHRs	✗	✗	✗	✗	✗	✓	✗	IQVIA Rx & Dx	2,961,750 / NA	✗	✓
[279]	2019	GAN for SSL	SSL based GANs for detecting rare diseases in unlabelled time-series EHRs	✗	✗	✗	✗	✗	✓	✗	IQVIA Rx & Dx	1,792,760 / NA	✗	✗
[271]	2019	GAN	SSL based labeling of unlabled data, and GAN-based data augmentation in tabular EHRs	✗	✗	✗	✗	✗	✓	✓	20 datasets from UCI Cerebral stroke private dataset	NA / 80-2,000 11,039 / NA	✓ ✗	✗

[1] The included evaluation components are (DWS): Dimension-wise Similarity, (LDS): Latent Distribution Similarity, (JDS): Joint Distribution Similarity, (IDRS): Inter-dimensional Relationship Similarity, (PP): Privacy Preservation, (DU) Data Utility, and (Qual) Qualitative Evaluation, which are explained in details in section 2.3.4.

[2] The dataset size is reported in the format of (N/R) where N refers to the number of patients and R refers to a number of records, reported in each of the works.

[3] The symbol ✓* refers to data sources that can be accessed after going through an application process

Table 2.2: Summary of the various uses of GANs for EHRs (continued)

References	Year	Problem		Evaluation							Medical Dataset		Open Access	
		Model	Task	DWS	LDS	JDS	IDRS	PP	DU	Qual	Dataset	Dataset size (N/R)	Dataset	Code
Imputation of Missingness														
[275]	2018	GAIN	GAN-based discrete & categorical tabular data imputation	✗	✗	✓	✗	✗	✓	✗	UCI Breast dataset	NA / 569	✓	✓
[281]	2018	Stackelberg GAN	Stabilizing GAIN imputation for discrete, continuous, & categorical EHRs using Stackelberg principles	✗	✗	✗	✗	✗	✓	✗	MIMIC-III	38,645 / 58,000	✓*	✗
[163]	2018	GAN with GRUI	GAN-based multivariate time-series EHR imputation	✗	✗	✗	✗	✗	✓	✗	PhysioNet Challenge 2012	NA/ 4,000	✓	✓
[164]	2019	E^2GAN	Improved GAN-based multivariate time-series EHR imputation	✗	✗	✗	✗	✗	✓	✗	PhysioNet Challenge 2012 dataset	NA/ 4,000	✓	✓
[270]	2019	Categorical GAIN	Improving GAIN imputation of categorical tabular EHRs	✗	✗	✗	✗	✗	✓	✗	UCI Breast Cancer PRAEGNANT study	NA/ 286 1234 / NA	✓ ✗	✗
[30]	2019	GAIN adaptation	Improving GAIN imputation of mixed tabular EHRs, including multi-categorical features	✗	✗	✗	✗	✗	✓	✗	UCI breast dataset	NA / 569	✓	✓
[58]	2021	MI-GAN	GAN-based multiple imputation for categorical time-series EHRs	✗	✗	✗	✗	✗	✓	✗	ADNI dataset	NA / 649	✓*	✗
[95]	2021	Bi-GAN	Concurrent imputation and prediction in time-series EHRs	✗	✗	✗	✗	✗	✓	✗	Nemours Pediatric All of Us	66,878 / NA 34,226 / NA	✓* ✓	✓
Treatment Effect Estimation														
[277]	2018	GANITE	Generating missing counterfactual data and individualized treatment effects estimation in tabular EHRs	✗	✗	✗	✗	✗	✓	✗	Twins IHDP	11,400 / 11,400 747 / 747	✓ ✓	✓
[174]	2018	CWR-GAN	Generating time-series post-treatment outcomes for ITE estimation in biomedical translation tasks	✗	✗	✗	✗	✗	✓	✗	MIMIC-III	2,000 / NA	✓*	✓
[79]	2020	MGANITE	Estimating effects of continuous, binary and categorical treatments via conditional GANs on tabular EHRs	✗	✗	✗	✗	✗	✓	✗	AML dataset	NA/212	✓*	✓
[153]	2020	GAD	Continuous treatment effect estimation by deconfounding in tabular EHRs	✗	✗	✗	✗	✗	✓	✗	Twins	4,821 / NA	✓	✗
[83]	2021	PSSAM-GAN	Propensity score augmentation matching for tabular EHRs	✗	✗	✗	✗	✗	✓	✗	S. aureus dataset IHDP	NA / 2,006 747 / 747	✗ ✓	✓
Privacy Preservation														
[264]	2018	DPGAN	Generating differential private EHR data using moment-accounting techniques	✓	✗	✗	✓	✓	✗	✗	MIMIC-III	NA / 46,520	✓*	✓
[123]	2018	PATE-GAN	Generating differential private tabular data using PATE	✗	✗	✗	✗	✓	✓	✗	UCI Epileptic Seizure Recognition Kaggle Cervical Cancer UNOS Transplant MAGGIC	NA / 11,500 NA / 858 NA / 23,706 NA / 30,389	✓ ✓ ✓* ✓*	✓
[16]	2019	AC-GAN	Generating Differntially private GAN via discriminator clipping for tabular EHRs	✓	✗	✗	✓	✓	✓	✓	SPRINT MIMIC-III	6,502 / NA 8,260 / NA	✓* ✓*	✓
[256]	2020	PART-GAN	Improving private GAN training of time-series EHRs	✓	✗	✓	✗	✓	✗	✗	Philips eICU	200,000 / 224,026,866	✓*	✗
[274]	2020	ADS-GAN	Anonymizing generated tabular EHR data while minimizing patient identifiability	✓	✗	✓	✓	✓	✓	✓	MAGGIC (RCT data) 3 UNOS Transplant datasets	30,389 / NA 23,706-56,822 / NA	✓* ✓*	✓
[267]	2020	HealthGAN	Improved End-to-End privacy-preserving WGAN-GP with a focus on privacy & resemblance metrics	✓	✗	✓	✗	✓	✓	✓	MIMIC-III	NA / 27,000	✓*	✓
[116]	2021	HCGAN	Improving robustness to privacy attacks by training Cramér GANs for tabular EHRs	✗	✗	✗	✗	✓	✓	✗	UCI Breast dataset Texas Hospital Data	NA / 699 NA / 186,976	✓ ✗	✗

[1] The included evaluation components are (DWS): Dimension-wise Similarity, (LDS): Latent Distribution Similarity, (JDS): Joint Distribution Similarity, (IDRS): Inter-dimensional Relationship Similarity, (PP): Privacy Preservation, (DU) Data Utility, and (Qual) Qualitative Evaluation, which are explained in details in section 2.3.4.
[2] The dataset size is reported in the format of (N/R) where N refers to the number of patients and R refers to a number of records, reported in each of the works.
[3] The symbol ✓* refers to data sources that can be accessed after going through an application process

Recognizing that tabular data fails to capture temporal dynamics, various GAN-based models have been proposed for time-series EHRs. *SynTEG* [286] employed a two-stage pipeline combining self-attention [249] and Wasserstein GANs to generate time-stamped diagnostic sequences. Similarly, *DualAAE* [147] used recurrent adversarial autoencoders to model visit-based medical code sequences. Further, *Yahi et al.* [266] generated continuous lab value trajectories to monitor drug effects. *RCGAN* [68] incorporated LSTM-based recurrent GANs for continuous vital sign sequences. *SC-GAN* [254] introduced a sequentially coupled generation mechanism to reflect clinical practices, such as adjusting medication doses based on patient state.

To mimic real-world EHRs' multimodal nature, recent efforts have targeted mixed-type data generation. *WGAN* [42] synthesized tabular EHRs with administrative and diagnostic codes. *HGAN* [268] integrated constraint-aware penalization to preserve logical consistency across binary, categorical, and continuous variables. Finally, *EHR-M-GAN* [149] proposed a dual-VAE and LSTM-based architecture for longitudinal mixed-type EHR generation, enabling temporal coherence and inter-type correlation learning.Overall, the evolution of GANs for EHR generation demonstrates increasing sophistication in modeling structure, time, and heterogeneity, contributing significantly to the availability of high-quality synthetic data for downstream tasks.

2.3.2 Imputation of Missing EHR Data

Handling missing data remains a significant challenge in EHR-based machine learning. Missingness in clinical datasets can arise from a variety of causes including machine failure, irregular sampling, omission of tests due to clinical irrelevance, or high costs and risks associated with procedures [171, 139, 65, 27, 135, 7]. Such missingness is typically categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) [157].

Generative Adversarial Networks (GANs) offer a powerful framework for imputing missing values by learning to generate plausible data samples without explicitly modeling the data distribution [270]. The seminal work in this area, *GAIN* [275], introduced a modified GAN architecture where the generator imputes missing values and the discriminator distinguishes between observed and imputed entries, using hint vectors to improve performance. This framework was benchmarked against conventional methods such as MICE [247], missForest [234], and Expectation-Maximization [186].

Further enhancements of GAIN have been proposed to address specific challenges in EHR imputation. For instance, *Stackelberg-GAN* [281] adapted GAIN to include multiple generators acting as followers and a discriminator as a leader, applying

game-theoretic principles to improve performance on highly incomplete datasets like MIMIC-III [121]. For categorical data, *Cat-GAIN* [270] proposed fuzzy binary encoding to preserve category semantics. To handle mixed-type EHR data, Camino et al. [30] introduced architectural improvements such as multi-output generators and gumbel-softmax activation [118]. Dai et al. [58] extended GAIN with a theoretically supported model for MAR scenarios, resulting in the *MI-GAN*, which outperformed existing methods in both accuracy and computational efficiency.

While most of these approaches were limited to static tabular data, time-series EHRs present unique challenges due to temporal irregularity. Luo et al. [163] introduced a GRU-based GAN model that uses a modified GRU cell (GRUI) to handle irregular time lags. The model operates in two stages: first, training a GAN to generate realistic sequences; then, optimizing latent noise vectors to find the best imputation for each sample. Due to high computational cost, an end-to-end alternative E^2GAN [164] was proposed to speed up training by bypassing the latent space search through a compression-reconstruction pipeline.

More recently, Gupta et al. [95] developed *Bi-GAN*, which simultaneously imputes missing values and forecasts future events using Bi-directional RNNs. This architecture enables "any-time" prediction without requiring fixed prediction windows, improving flexibility and clinical applicability. Collectively, these advancements establish GANs as a promising class of models for imputing missing data in both tabular and time-series EHRs, addressing structural irregularities and diverse missingness mechanisms.

2.3.3 Treatment Effect Estimation

Estimating treatment effects is a complicated causal inference task with many data challenges, where the aim is to estimate the patient's response to a specific treatment. The major challenges in this field arise from missing counterfactual data, and the unobserved outcomes of untaken treatments [108]. In randomized controlled trial (RCT) settings, patients in the treatment group are matched to those in the control group to compensate for missing counterfactuals. However, despite being the gold standard for various clinical applications, RCT-based treatment effect estimation suffers from multiple issues concerning their high cost [214], relatively small size [101], ethical issues [90], and short follow-up duration, which could miss out the long-term effects of medications [23]. A low-cost alternative to RCT data is regularly collected EHR data. Specifically, longitudinal EHRs, which include diverse patient cohorts,

long-term outcomes without strict exclusion criteria, making EHRs more representative of the patient population [23, 188]. However, in EHR data, treatments are not assigned at random, and there is no clearly defined control group. Thus, estimating treatment effects from EHRs requires measures to control confounding effects and perform covariate adjustment [175, 207] to avoid selection bias.

The generative capabilities of GANs are a valuable option for various treatment effect estimation applications. In [277], Yoon et al. made use of the generative properties of GANs to generate counterfactual outcomes. In their novel design, GANs for inference of Individualized Treatment Effects *(GANITE)*, they considered counterfactual outcomes to be missing labels, similar to their earlier work in [275]. *GANITE* utilized a pair of GANs: one for counterfactual imputation and another for treatment effect estimation. In the first GAN, the generator's task is adjusted to generate missing counterfactual outcomes, while the discriminator's task is to tell the factual from the counterfactual outcomes. In the presence of counterfactual outcomes, a treatment effect estimation function can be predicted using traditional machine learning models. However, in *GANITE*, Yoon et al. utilised another GAN to model treatment effect estimation by taking the output of the counterfactual GAN as input and generating a potential outcome vector with confidence intervals [277]. While *GANITE* focused on binary treatment, [48] focused on generating time-series post-treatment outcomes. Chu et al.'s work was motivated by the scarcity of paired pre- and post-treatment patient time-series data in settings such as ICU ventilation and vasopressor assignment. Their proposed model, Cycle Wasserstein Regression GAN *(CWR-GAN)*, is a hybrid of several architectures; original GAN [92], Wasserstein GAN [11], and cycle-consistent GAN [288]. The Chu et al.'s of CWR-GAN tested their model in regression-based tasks and provided an alternative to the traditional unidirectional regression approaches, where unpaired data would be ignored during training [48].

To extend GAN-based treatment estimations from binary to various types of treatment variables, including categorical and continuous, [79] applied modifications to *GANITE*, which they named *MGANITE*. Estimation of continuous treatment is of high importance in applications involving dosage adjustment, especially in oncology [178]. One of the main modifications was a mathematical adjustment to the loss function that takes a treatment assignment vector in both counterfactual and ITE estimation blocks, to allow for simultaneous treatment effect estimation [79]. When using observational data where treatments are not randomly assigned, controlling confounding factors, such as using propensity scores, is essential [57]. In [153], Li et

al.' proposed a GAN-based model that generates a "calibration" distribution, one that eliminates associations between covariates and treatment assignment by a random perturbation process of the treatment variable. The generative capabilities of GANs are used to learn a weight vector that is used to adjust the distribution of observed data and construct the calibration data. Li et al. refer to their model as Generative Adversarial De-confounding (*GAD*) [153].

Statistical approaches such as propensity score matching (PSM) are commonly used by classical treatment effect estimation works to balance the characteristics of the population assigned to an intervention or a control group [28]. However, despite their popularity, PSM approaches can lead to large reductions in sample sizes due to unmatched control samples [28]. Recently, a GAN-based synthetic augmentation matching model, *PSSAM-GAN*, was proposed to mitigate the problem of sample size reduction using PSM approaches [83]. First, Ghosh et al. matched their samples based on calculated propensity scores. Then, to be able to use unmatched samples, the Ghosh et al. used a GAN-based model to generate treatment matches for unmatched control samples [83]. Finally, the original EHR data was augmented with the newly generated matched samples to be used for the downstream treatment estimation tasks [83].

2.3.4 Evaluation methods

The evaluation of GANs for EHR data generation remains complex due to the absence of universally accepted metrics and methodologies [242]. Effective evaluation is essential to ensure that generated data closely approximate the real data distribution, maintain privacy, and provide utility for downstream tasks. Existing evaluation methods are classified as either qualitative or quantitative, each addressing different aspects of synthetic data quality.

Quantitative evaluation focuses on metrics that assess similarity, utility, and privacy. Resemblance evaluation examines how well the synthetic data mirrors the distribution of real data across individual features (dimension-wise similarity) and multi-feature relationships (joint distribution similarity). Metrics like Kullback-Leibler divergence (KLD) [17], Jensen-Shannon Divergence (JSD) [167], and Maximum Mean Discrepancy (MMD) [93] are commonly used to assess distribution similarity, while methods such as association rule mining[265] and Pearson correlation [18] gauge inter-dimensional relationships.

For utility, methods like Train on Synthetic, Test on Real (TSTR)[68] verify if machine learning models trained on synthetic data perform well on real data, thus

indicating the synthetic data's effectiveness in predictive tasks. Privacy preservation is assessed using differential privacy[66] and defenses against membership inference attacks [228], attribute disclosure attacks [172], and model inversion attacks [76] to ensure synthetic data minimize re-identification risks.

Qualitative evaluation methods supplement quantitative metrics by visually inspecting data distributions, feature correlations, or patient trajectories [267]. Visualization techniques and clinical evaluations are often used to assess the realism of synthetic data and its adherence to clinical expectations. Clinician evaluations can involve rating the plausibility of synthetic data [16] or distinguishing synthetic from real data [259]. Ablation studies examine the role of individual model components in the quality of synthetic data by iteratively removing parts of the model to observe performance changes [268, 277].

The diversity of evaluation metrics reflects the complex nature of EHR data and highlights several gaps. Currently, there is no standardized metric for evaluating synthetic EHRs in all relevant dimensions, complicating fair benchmarking and model comparison [105]. Many metrics also require significant computational resources, especially for high-dimensional data [127]. Addressing these gaps will require developing evaluation metrics that balance interpretability, efficiency, and relevance to clinical applications, ultimately enabling more robust and trustworthy synthetic EHR data generation.

2.4 Time Series Imputation

The handling of missingness in longitudinal EHRs is a critical step before performing statistical or machine learning analysis, as missingness can reach as high as 80% or more in many EHR datasets [230, 121, 196]. Missing data processing usually involves deletion or imputation of data. The simple deletion of records or variables with missing values often leads to compromised results and is not recommended practice, especially when the data is highly missing [217, 136, 187]. Missing value imputation and replacing unobserved values with substitute value methods gained more attention with the advancements in statistical and machine-learning-based models. Current missing-value imputation methods can be classified into three main categories: The first category of methods, known as simple imputations, includes replacing missing values with population mean, median, or personalized Last (observed) Value Carried Forward (LOCF) [4, 224]. Although useful and commonly used in various clinical

applications such as clinical trials, imputing by the LOCF presents an issue in observational EHRs, particularly when the feature is never observed for the patient, hence no value to carry forward. The second type of imputation method is a simulation-based statistical class of methods, often called multiple imputation methods. For multiple imputation methods, N complete datasets are simulated from the incomplete dataset by imputing the missing values N times. These completed data sets N are analyzed and then combined into a single dataset [117]. A common multiple imputation method is multivariate imputation by chained equations (MICE), which is often used by many epidemiological and clinical research [235]. Despite the usefulness of many approaches, they often fail to handle missingness in longitudinal EHRs as they cannot account for time-series sequential data. Although MICE is useful for certain imputation tasks, they may not be as effective in modeling the sequential nature of time-series data as they read the data in flattened form. This approach often treats each time point independently, which can lead to suboptimal imputations when dealing with the continuous and interdependent nature of time-series data. The third and most advanced category of imputation methods is based on machine learning algorithms. Matrix Factorization (MF) [201], Expectation Maximization (EM) [88], K-Nearest Neighbor (KNN) -based imputation [168], along with Recurrent Neural Network (RNN) [31].

2.4.1 Deep Generative Imputation Models

Recently, deep generative neural networks such as Generative Adversarial Networks (GAN) [92] and Variational Autoencoder (VAE) [137] were proposed to generate synthetic data by learning the underlying distribution and dynamics of real datasets. The applications of deep generative models in healthcare range from generating synthetic patient records [44, 149], to estimating treatment effects [277], detecting undiagnosed patients on a large scale [152], and generating missing value imputations [275, 185, 75, 164]. The generative capabilities of deep generative models make them naturally suitable for generating not only new synthetic records but also missing data imputations, where they show superior performance to most of the commonly used statistical methods in tabular and longitudinal data [134, 85, 253]. Despite their high reported performance, most deep generative methods assume that *missingness is missing at random*, which limits their potential in real healthcare applications, where missingness can be informative and reflect the state of the individual underlying patient [217]. Furthermore, many of these models can only generate *population*

level rather than *personalized level* information conditioned on the already observed patient data, characteristics, and treatments.

Table 2.3: Comparison of different missing data imputation methods in terms of the architecture, compatibility with time-series data, data type and ability to account for individualized missingness.

Model	Architecture	Generative Model	Time-series Data Type	Individualized Missingness Indicators
MICE [247]	Statistical regression	✗	N.A.	✗
GAIN [275]	GAN	✓	N.A.	✗
BRITS [31]	RNN	✗	continuous	✗
HI-VAE [185]	VAE	✓	N.A.	✗
GP-VAE [75]	VAE	✓	continuous	✗
TimesNet [262]	Temporal 2D-Variation Modeling	✗	continuous	✗
SAITS [63]	Self-attention-based model	✗	continuous	✗

* N.A: Note Applicable, as the methods are proposed for static data.
[a] Acronyms in Full Form: GAN: Generative Adversarial Networks, RNN: Recurrent Neural Network, and VAE: Variational Autoencoder :

2.5 Treatment Effects Estimation

The study of treatment effects has gained much attention in recent years, where various tools and approaches have been proposed to help mitigate the financial cost and optimize the effectiveness and efficacy of prescribed treatments. One of the fast-growing applications in clinical ML is studying ITEs [21], where ML capabilities, in many cases, have superseded those of state-of-the-art clinical and pharmaceutical advancements. Using the wealth of observational EHRs data, the individualized patient response to various treatments can be estimated [21].

Although there has been increased attention to ITE works from static observational EHR data [277, 43], much less attention has been given to those from EHR-data measured over time, often referred to as time-series EHRs. Despite similarities in the concepts for treatment estimation between static and time-series data, challenges related to the time-varying nature of covariates make many existing works not directly applicable to time-series data. To this end, we aim to provide an overview of the main concepts and ideas for estimating ITE in general followed by ITE applications for time-series data.

2.5.1 Randomized Controlled Trial Data

RCT data are considered the gold standard for studying treatment effects. This mainly stems from the random assignment of participants to treatment groups, which eliminates confounding bias effectively. The randomness and control measures used in RCTs make them a lucrative option for studying the effect of treatments; an unbiased

estimate of the average treatment effects (ATE) can be directly calculated from the data [160]. Although RCTs offer methodological strengths, various challenges hinder their optimal and complete use for studying treatment effects [23]. Firstly, despite the power of "randomness" to eliminate confounding, certain biases may remain in RCTs. This bias does not come from the level of treatment assignment, but from representatives of samples presented in the study, denoted as sample selection bias hereafter. Most RCTs tend to employ stringent exclusion criteria for enrolled participants, which could introduce bias in the results for members of the unrepresented population [155, 131, 219, 23, 236]. This becomes a bigger issue when a significant percentage of treated patients in real-world data belong to the population excluded from the RCT. For example, the aging population is rarely enrolled in diabetes RCTs due to their age and multimorbid health conditions, although it represents a large proportion of diabetic patients [125]. Such factors might introduce bias that makes RCTs' generalizability and external validity questionable [101, 23].

The generalizability of the results of the RCT is further limited by the fact that current RCTs are typically designed to only measure ATE. In other words, they estimate how the "average patient" will respond to a given treatment. For a more personalized approach, the unique characteristics of the patient would be needed to predict an individual patient's response to a given treatment, and it may differ significantly from the average response of a population. Furthermore, RCTs tend to be financially costly to design and implement [112, 214] and their data tend to be hard to share due to privacy concerns [218]. Additional limitations exist in terms of their relatively small sample size [101], ethical issues [90], and short lengths of follow-ups which might miss out on the long-term effects of medications [23]. For example, the effects of oral contraceptives were not quantified until the presence of long-term data, which were not captured in RCTs [226].

2.5.2 Observational Data: From Efficacy to Effectiveness

Despite all the stated challenges, RCTs are still essential for determining the *Efficacy* for treatments [138] but are not necessarily optimal for studying treatment *Effectiveness*. Making a clear distinction between these two terms will help one understand when data-driven and statistical approaches can improve the use of RCT data. Efficacy refers to the effect of interventions under ideal "theoretical" circumstances, while effectiveness means that an effect is detected not under ideal conditions, but under real-world conditions [138]. Observational EHR data presents a good candidate for better testing of treatment effectiveness. Specifically, longitudinal observational

data collected in EHRs typically includes a diverse cohort of patients without strict exclusion criteria, making it more representative of the real or targeted population of patients. Observational data are also less expensive to collect compared to RCTs and capture long-term results [23, 188]. Furthermore, with the widespread use of EHR systems worldwide, observational data can allow one to estimate treatment effects from various clinical settings such as low-middle-income countries (LMICs), where performing RCTs would not be feasible due to high costs. To this end, observational data is a promising resource for studying treatment effects with more inclusive estimations for various patient groups while maintaining low cost and learning from real-world evidence.

2.5.3 Challenges of Treatment Effects Estimation

Estimation of treatment effect is a subfield of causal inference and as such suffers from the fundamental problem that counterfactual outcomes are never observed [208]. A counterfactual outcome refers to the hypothetical outcome that would have been observed if a different treatment had been given than the actual (factual) one [21]. Of course, for treatments that were not administered, it is not possible to extract the ground truths of individual patient outcomes directly, either from observational or clinical trial data. In RCTs, this problem is resolved by estimating the ATE across the entire study population but not the ITE of an individual. Randomization of treatment allocation ensures that the underlying distribution of patient characteristics is similar in both the treatment and control groups. This allows us to compare the average outcomes in both groups and thus determine the ATE.

Omitting the randomized control measures used in RCTs and relying on observational data precludes the direct computation of treatment effects. This is because treatment assignment is not random, but biased, as treatment selection in observational data is often driven by the patient's characteristics, such as treatment allocation flow charts in clinical practice guidelines [244]. Therefore, clinical practice recorded in real-world observational data results in systematic differences in the characteristics of treated and untreated patients. To be able to estimate treatment effects from observational data, it is essential to remove the confounding bias, introduced by the non-random treatment assignment. A confounder is a variable that influences both the intervention and the outcome, which can lead to a spurious association between them [9]. For example, in clinical practice, patients with more severe health conditions might be more likely to receive stronger medications while still being expected

to have poorer outcomes. However, it would be wrong to conclude that stronger medication leads to poorer patient outcomes. Failing to adjust for the severity of a health condition as a confounding factor in this case would lead to measuring a spurious association that does not reflect the actual effect of treatment.

2.5.4 From Average to Individualized Treatment Effects Estimation

The modern understanding of estimating treatment effects is highly attributed to the work of Neyman-Rubin's "potential outcomes framework" [210]. In the Neyman-Rubin model, the ITE between a treatment A and a treatment B is defined as the difference between the two potential outcomes (e.g. blood pressure) after administering a treatment A or B to a given patient. Subject to certain assumptions [21], the ATE can be calculated directly from the RCT data by calculating the difference between the average outcomes in both treatment groups. However, because of the absence of counterfactuals in real-world data, ITEs cannot be calculated directly, but must be estimated through the use of models. An illustrative figure of the difference between ATE and ITE is shown in Figure 2.2. Based on statistical methods, but also driven by recent developments in machine and deep learning approaches, various models have been proposed to estimate treatment effects on a personalized level [277, 83, 20, 176]. Overall, most of these models estimate the potential outcome for an individual patient by learning the underlying effects and interactions between patients' characteristics, treatment, and outcome. The patient's ITE can then be calculated as the difference between the predicted potential outcomes with and without treatment. Furthermore, various methods were proposed to address the problem of confounding bias introduced in observational studies as a result of unobserved confounders [205, 19, 142], paving the way for ITE estimations for personalized medicine.

2.5.5 Estimating Individualized Treatment Effects from Time-series Data

Various works for estimating ITE using observational EHR data have recently been proposed [277, 222, 83]. Despite the plethora of works, most of them estimate the treatment effect using static data, where each patient is represented as a snapshot of covariates at the exposure of the treatment. Although useful, using only static data and disregarding the time component have many limitations in estimating the

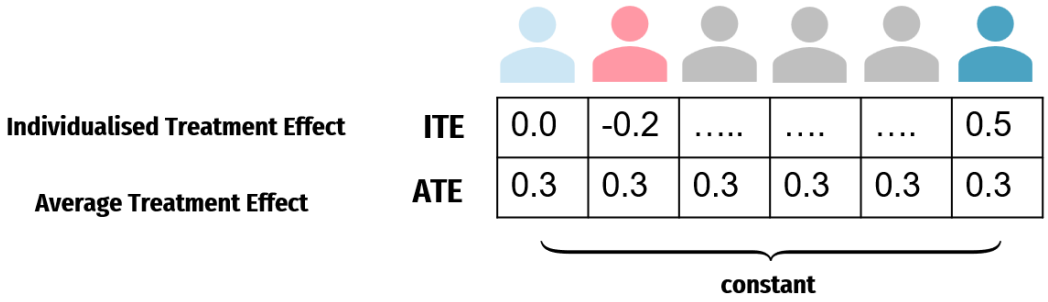


Figure 2.2: Figure Showing the difference between the Average Treatment Effect (ATE) and Individualised Treatment Effects (ITE). As shown, the ATE is constant across the population whereas the ITE varies across the population, reflecting heterogeneity.

impact of treatment over time. Furthermore, ITE estimation from static data limits the opportunities for learning when to stop or change treatments when the outcome is dynamic and varies over time, a critical clinical application for ITE. Additionally, in the static treatment effect estimation setup, the treatments are assigned at a single time point and often remain static over time. However, using time-series data for the estimation of treatment effects would allow monitoring and dynamical change of the treatment plan while simultaneously observing the effect of time-varying treatment on patient covariates and outcomes of interest. A typical example of a time-series ITE problem is in the cancer application, where a patient's treatment option (e.g., radiation or chemotherapy) is adjusted according to his or her clinical response (tumor size) over time. In the static setting, such a problem cannot be addressed due to the absence of the time factor.

Despite the promise and potential of treatment effect estimation from time-series data, the major problem lies in the time-varying or temporal confounders. Similarly to static confounders, a temporal confounder is typically a variable that varies over time and affects both the assignment of treatment and the outcome. For example, consider that angiotensin receptor blockers (ARBs) (treatment 1) are given when the blood pressure of a hypertension patient (covariate) is outside the normal range value. Suppose also that this patient's covariate was affected by the previous administration of an ACE inhibitor (treatment 0), another type of treatment for uncontrolled blood pressure. Estimating the effect of a different sequence of treatments on patient outcome would require adjusting for bias in the current step (treatment 1) and bias introduced by the previous application of ACE inhibitors (treatment 0). Adjusting for

time-varying confounders remains a major challenge hindering the direct application of methodologies developed for static treatment effects tasks to dynamic problem settings. Here we have reviewed works in the literature for estimating treatment effects from time-series data, including the estimation frameworks, model architectures, and assumptions used. An overview of the existing work on ITE from time-series data is presented in Table 2.4. They are categorized into two main groups: (i) outcome estimation methods which focus on inferring the ITE by estimating the potential outcomes of different treatments, and (ii) deconfounder methods, which estimate the ITE in the presence of hidden confounders. More details are presented in Sections 2.5.6 and 2.5.7, which cover the estimation of the ITE outcome and the deconfounder methods for time-series data, respectively. In Figure 2.3, we show an example causal model that underlies a dynamic ITE estimation setting with time-varying treatments and covariates.

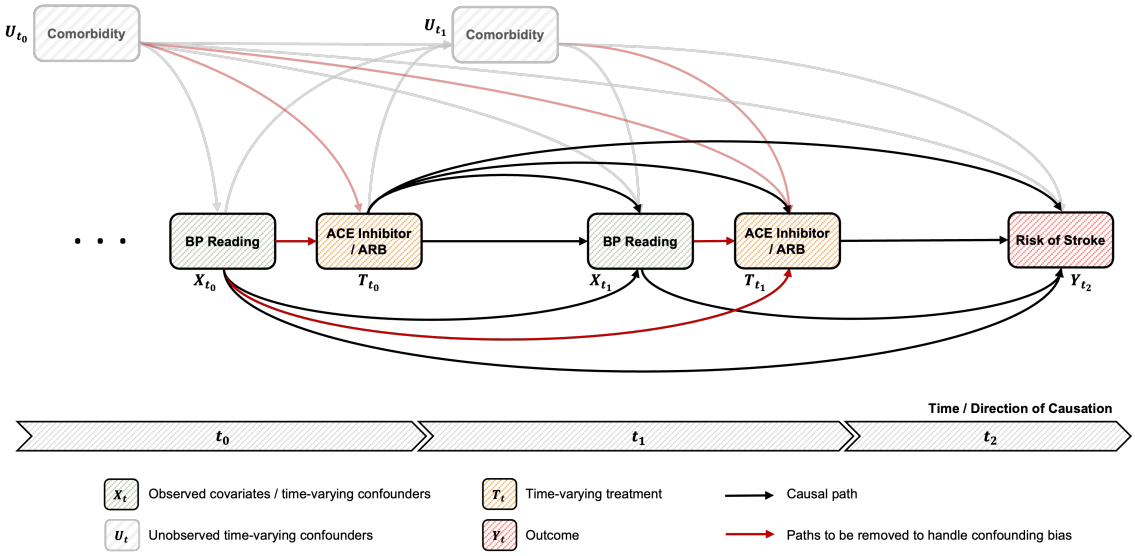


Figure 2.3: Illustration of a causal model underlying a dynamic ITE estimation setting with time-varying treatments and covariates. The arrows indicate causal dependencies between variables. Here, the observed covariates are blood pressure (BP) readings that act as time-varying confounders affecting all subsequent treatment decisions (ACE inhibitor vs. ARB) and the outcome of interest (risk of stroke). We also showcase an example of an unobserved "hidden" confounder, namely potential comorbidity that is not directly represented in the data. All connections from covariates and unobserved variables to the treatment variables are depicted in red, and would need to be accounted for to remove confounding bias effectively.

Table 2.4: A summary of ITE works for time-series data.

	Proposed Methods	Estimation Framework	Assumptions	Model Architecture	Validation Data
Outcome Estimation Methods	MSM [205]	MSM	C/P/SSI	LR	NA
	RMSN [154]	MSM	C/P/SSI	LSTM	simulated tumor dynamics data
	CRN [20]	Balanced Representation	C/P/SSI	LSTM	simulated tumor dynamics data, ICU data
	G-Net [151]	G-formula	C/P/SSI	LSTM	simulated tumor dynamics data, simulated cardiovascular data
	Causal Transformer [176]	Balanced Representation	C/P/SSI	Transformer	simulated tumor dynamics data, (semi-synthetic) ICU data
Deconfounder Methods	Time Series Deconfounder [19]	Latent Factor Model	C/P/SSSI	RNN Factor Model	simulated data, ICU data
	Sequential Deconfounder [99]	Latent Factor Model	C/P/TIUC	GPLVM	simulated data, ICU data
	Deconfounding Temporal AutoEncoder [142]	Noisy Proxies	C/P	AutoEncoder	simulated data, ICU data

*The abbreviations in full form. (C): Consistency, (SO): Sequential Overlap, (SSI): Sequential Strong Ignorability, (SSSI): Sequential Single Strong Ignorability, (TIUC): Time-Invariant Unobserved Confounding, (LR): Logistic Regression, (LSTM): Long Short-Term Memory, (RNN): Recurrent Neural Network, (GPLVM): Gaussian Process Latent Variable Model.

Assumptions

Estimating treatment effects from time-series data relies on the potential outcome framework [210] and its extensions to the time-varying setting [204]. In the potential outcome framework, an ITE is the difference between the potential outcomes for a specific individual given different treatments. Three main assumptions are typically required for treatment effects to be identifiable from time-series data. We explain each in lay terms and provide examples from real-world clinical problems. For interested readers, we provide references to works that explain and include mathematical notations.

1. **Consistency.** This assumption states that the potential outcome of a patient should be consistent with his/her "factual" outcome if the same treatment plan is applied. For example, consider a patient who has been on a diabetes treatment plan A for several years, where blood glucose is considered a monitored outcome over time. If a model is built to estimate the potential outcome of this patient given the same treatment plan A, this outcome would be equivalent to the outcome observed for that patient (that is, blood glucose). The mathematical notation and theoretical basis for this consistency are explained in [20, 176].

2. **Sequential Overlap or Positivity**. This assumption means that each treatment option has a non-zero probability of being given to the patient at each timestep. For example, let us consider a cancer treatment in which options are radiotherapy or chemotherapy. If a patient was given radiotherapy last month, a doctor might give this patient chemotherapy or radiotherapy this month, and both have a non-zero probability of being given to the patient.

3. **Sequential Strong Ignorability** This assumption means that conditioned on the observed patient history, the current treatment assignment is independent of the potential outcome. In some works, this assumption is referred to as sequential exchangeability or "no unobserved confounders". In other words, no unobserved confounders affect both treatment and outcome. While useful, this assumption cannot be tested in practice since various factors may impact the treatment and the outcome, yet they might not be recorded or observed. For this purpose, various works have been proposed to relax this assumption (see the reference works for estimation with hidden confounders described in Section 2.5.7). Example variants include *Sequential Single Strong Ignorability*,

where the assumption is limited to no hidden single cause confounders [19]. Another example is *Time-Invariant Unobserved Confounding*, where confounders exist under the condition that they are the same random variable at each time step [99].

2.5.6 ITE Outcome Estimation Methods in Time-series data

Marginal Structural Models (MSM)

Various epidemiological approaches have been proposed to account for confounders varying over time, one of which is *inverse probability of treatment weighting* (IPTW) [41]. The main idea behind IPTW involves assigning weights that will redistribute or balance the population so that the effect of time-varying confounding is removed. The weights are derived based on the inverse probability of receiving the patient's treatment at each respective time point conditional on their covariate's history. The IPTW setup creates a pseudo-population set in which receiving treatment assignment is independent of the underlying patient characteristics and previous treatment assignments [12]. One way to implement IPTW in the time-series setting is via *marginal structural models* (MSMs) [205]. MSMs focus on controlling the effects of time-varying confounders affected by previous treatment exposure. The name "Marginal" refers to the approach of estimating the marginal distribution of treatment over time for the outcome [261]. Similarly, the name "structural" refers to the approach of exploring causal relationships inspired by econometrics [261]. An MSM first calculates the weights, most commonly via a regression-based IPTW model, and assigns such weights to each observation. The estimated weights indicate whether each of the observations in confounders is under-represented or over-represented in the sample for a target population [261]. The use of sample reweighing aims to remove the imbalance and bias caused by the uneven distribution of time-varying confounders across treatment groups. Finally, the treatment effect can be estimated using the calculated weights.

While powerful and useful, MSMs have limitations when dealing with high-dimensional and complicated data dynamics. This is because the treatment effect predictions are calculated using linear or logistic regression models. To address this, [154] proposed *recurrent marginal structural networks* (RMSMs) where recurrent neural networks [50] were used to estimate the inverse probability of treatment weights and counterfactual treatment outcomes. Similarly to the standard two-stage MSM approach, RMSM has two main networks. The first network calculates the treatment probability weights

used for the IPTW. The second network, on the other hand, is the prediction used to determine the response to treatment given a sequence of treatments and the calculated weights [154]. Despite the promise shown by the statistical and deep learning approaches, MSMs can be unstable if the IPTW results in extreme weights, leading to model misspecification [20].

G-formula

Another method for estimating treatment effects in time-varying confounders is *G-formula*. The G formula was first described by Robins et al. [203], where the author proposed a method for generalizing standardization to time-varying treatments and confounders and referred to it as the formula of the G computation algorithm. Most of the epidemiological works use the term G-formula or G-computation to refer to the same method proposed by [203]. The key assumption for the G computation formula is that the treatment received at each time was assigned conditional on the observed past treatment and the history of the covariates [203]. G-formula works by estimating the conditional distribution of relevant covariates given the covariate and treatment history at each time point, then producing Monte Carlo estimates of counterfactual outcomes by simulating forward patient trajectories under treatment strategies of interest [184]. In most statistical works, the estimation of patient trajectories and outcomes is done using simplistic regression estimators. While useful, it is important to remember that simple regression models fail to capture complex dependencies over time when dealing with high-dimensional time-varying data. In terms of implementation, there are no well-established G-formula implementations in statistical packages, which limits its applicability when compared to MSMs. Recently, [151] proposed G-NEt, the first deep-learning work that estimates ITEs using an LSTM-based G-formula model. The G-Net results showed improved performance compared to those estimated using a logistic regression estimator and other deep-learning-based models [154, 20].

Balanced Representations

Unlike MSM and G-formula-based estimations, a new class of deep learning estimation evolved based on learning representations that balance the distribution of treatment and control groups. Original works for learning balanced representations were first proposed for static settings [120], then several studies used a deep learning architecture to learn treatment-invariant representation for each time step to remove the association between patient history and treatment assignment. For example, the *counter-*

factual recurrent network (CRN) [20] is the first work to use a sequence-to-sequence model to learn balanced representations through adversarial training. In their proposed work, CRN aims to learn representations that are not predictive of treatment assignments, but achieve the highest performance in predicting the outcome. Another related work is that of [176], where Melnychuk proposed *Causal Transformer*, which is a transformer-based model that aims to learn treatment-invariant balanced representations to estimate ITE over time. To do so, the Causal Transformer comprises three transformer sub-networks for processing the time-varying covariates, treatments, and outcomes, all of which are combined via a joint network with cross-attentions.

2.5.7 ITE Deconfounding methods for Time-series data

Latent Factor Models

All the aforementioned studies to estimate ITE in time-series data focus on settings where all confounders are observed, or in other words, they require a sequential strong ignorability assumption to hold. Despite the potential of such works, sequential ignorability is not testable in practice. To this end, several studies have proposed approaches in which sequential ignorability is relaxed to account for settings where forms of hidden confounders exist in the data. For example, the first work to propose a deep learning model for deconfounding time-series data was the *Time Series Deconfounder* [19]. In their proposed work, the Bica et al. focus on addressing a specific type of hidden confounders which they refer to as multi-cause hidden confounders. The Time Series Deconfounder builds on a factor model to learn the distribution of treatments over time. Using the dependence between multiple treatment options at each given time step, the factor model infers substitutes for unobserved confounders at each time step. The assumption of sequential strong ignorability is relaxed to sequential single strong ignorability where the assumption is limited to the absence of hidden single-cause confounders [19]. The Times Series Deconfounder can be applied to the datasets before passing the deconfounded data to other outcome estimation models such as RMSM [154] or CRN [20].

While the results show great promise, the Time Series Deconfounder only works when there are multiple treatment options and fails when there is a single treatment option at each time step. This limitation of the Time Series Deconfounder is related to its use on the dependence between multiple treatment options to infer substitutes of hidden confounders. *Sequential Deconfounder*, on the other hand, is a method that deconfounds time-series data for ITE by fitting a Gaussian Process

(GP) latent variable model to capture any sequential dependence between the assigned treatments [99], without the limitation of depending on multiple treatments. The GP-based latent variable model aims to control for substitutes and uses them in conjunction with outcome estimation models such as RMSM [154] or CRN [20] to estimate treatment effects over time. Although the sequential deconfounder does not require multiple treatments at each time step, it requires a special case of the ignorability assumption, which they refer to as *Time-Invariant Unobserved confounding* [99]. This assumption requires that the hidden confounder is the same random variable at each time step.

Noisy Proxies

Most aforementioned studies utilize latent models to infer the hidden confounder in time-series ITE by capturing sequential dependencies. Recently, the Deconfounding Temporal AutoEncoder (DTA) is an autorencoder-based model that utilizes noise proxies as an alternative to latent factor models to learn hidden embeddings that resemble the true hidden confounders [142]. The main assumption in DTA builds on the fact that the observed covariates are not necessarily true confounders and assumes that the observed covariates are noisy proxies of the true confounders. DTA aims to learn a hidden embedding for which the ITE is the same when hidden confounders are present and when Sequential Strong Ignorability applies. To do so, DTA optimizes on a special loss that is referred to as a cause regularization penalty to produce results and treatment assignments that are conditionally independent for each hidden embedding [142].

Most studies found in the literature make use of simulated datasets to evaluate their methods for ITE estimation from time-series data. Unlike real-world data, where only the factual outcome is observed, simulations provide ground truths for all potential outcomes. Some simulated datasets used in the literature [19, 99, 142] do not aim to mimic specific medical scenarios; instead, they are based on purely mathematical modeling of time-varying covariates, hidden confounders, treatments, and outcomes. In contrast, models such as the pharmacokinetic-pharmacodynamic (PK-PD) model by Geng et al. [80], which simulates cancer dynamics, strive to provide a realistic perspective on actual medical processes. Simulated "observational" cancer growth data, derived from the PK-PD model, is widely used in the literature [154, 20, 151, 176] for evaluation purposes. The model has been adapted to simulate the change in tumor volume over time under the influence of different treatment options, such as chemotherapy and radiation. Time-dependent confounding

can be incorporated into the model by expressing the probabilities of administering chemotherapy and radiation as a function of tumor size [154].

In addition to the PK-PD model, Li et al. [151] evaluate the performance of G-Net on longitudinal data, simulated using Heldt et al. CVSim [102]. CVSim provides a mechanistic model of the human cardiovascular system and enables simulation of the trajectories of outcome parameters such as mean arterial pressure (MAP) or central venous pressure (CVP) in interventions such as different fluid or vasopressor administration strategies. Moreover, Melnychuk et al. [176] evaluate their Causal Transformer on semi-synthetic and real-world datasets of patient trajectories in the ICU that are based on the MIMIC-III dataset [122]. For their semi-synthetic data, they combine real-world covariates from the MIMIC-extract by Wang et al. (2020) [255] with simulated trajectories of control outcomes. The treated results are obtained by simulating synthetic binary treatments, incorporating confounding, and applying those treatments to the control outcomes [176]. For their experiments on real-world data, they used the same patient's covariates from MIMIC-III again and considered the effect of vasopressors and mechanical ventilation on blood pressure. However, since counterfactual outcomes are not available for real-world data, they can only report the performance in predicting the factual outcomes.

The fundamental problem of causal inference and the resulting reliance on (semi-) simulated data sets for comprehensive validation poses a major challenge to developing models for ITE estimation. Although models such as CVSim or the PK-PD model can offer valuable insights from a medical standpoint, their data generation process is less complex with simple assumptions, which could result in lower performance when compared to real-world applications.

2.5.8 Irregular Sampling and Missingness

All studies included in this work have shown great promise in estimating ITE in discrete time-series data. Transforming multi-variate time-series data to one with discrete time steps requires covariate alignment. Since most time-series covariates are measured at irregular time-steps where treatment is also administered at various time-steps, missing values are inevitably produced in the data. Despite the increasing number of proposed approaches for estimating ITEs and deconfounding the data, none of the previous work has investigated the impact of missing values on counterfactual predictions and model performance. With a wide variety of methods proposed to handle time-series missingness with various underlying assumptions on the nature

of missingness [75, 31, 63], we believe that new research directions should investigate the impact of such assumptions on the assumptions and models of ITE. This is particularly important when data missingness can be related to the underlying patient's health state, which can lead to a lack of follow-up, making it relevant patient information for the ITE estimation.

2.6 Summary

This chapter explored the role of deep generative models, missing-value imputation, and ITE estimation in advancing clinical machine learning applications. Deep generative models, such as GANs and VAEs, offer promising solutions for creating synthetic EHR data and augmenting training data in privacy-sensitive environments. Although these models excel in generating realistic structured healthcare data, their application to EHRs remains challenging due to the high dimensionality, heterogeneity, and sparsity of clinical data. Future efforts must focus on refining these models to improve stability and interpretability, which are critical for clinical applications. The imputation of missing values in EHR data is essential for maintaining data quality in predictive modeling. Deep generative methods offer robust solutions for handling missingness across diverse data types, allowing clinical models to leverage complete datasets for improved prediction accuracy. Resolving missing data issues is particularly valuable for time-series data, as it preserves the temporal structure essential for understanding patient progression over time. ITE estimation has become central to personalizing treatments by predicting individual patient responses, thus optimizing treatment efficacy and reducing costs. Although substantial progress has been made with static EHR data, adapting ITE methods to time-series data presents unique challenges due to time-varying covariates and the complexity of longitudinal data. Successfully incorporating time-series data and imputation methods promises more accurate and individualized predictions, further supporting effective clinical decision-making.

In summary, deep generative models, robust imputation strategies, and advanced ITE estimation techniques offer transformative potential in clinical machine learning. However, continued work is necessary to address specific clinical challenges, such as dynamic time-series modeling and the need for interpretable, reliable predictions. Closing these gaps will be essential to achieve truly individualized, data-driven treatment strategies in healthcare, which will be the focus of the subsequent chapters of this dissertation.

2.7 Relevant Publications

- **Ghosheh, G.**, Gogl. M., Zhu, T. (2025). A Perspective on Individualized Treatment Effects Estimation from Time-Series Data. Journal of the American Medical Informatics Association, 2025; https://doi.org/10.1093/jamia/ocae323
- **Ghosheh, G. O**., Li, J., & Zhu, T. (2025). Understanding Missingness in Time-series Electronic Health Records for Individualized Representation. arXiv preprint arXiv:2402.15730. NPJ Artificial Intelligence [Under Review]
- **Ghosheh, G.**, Li, J., Zhu, T. (2023). A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Computing Surveys*, 56(4), Article 63.

Chapter 3

Data Description

3.1 Overview

This thesis aims to explore the application of treatment effect estimation in various clinical domains, focusing on primary care and Intensive Care Unit (ICU) settings. Using multiple electronic health record (EHR) datasets, we seek to model patient outcomes and the effects of different treatments, with an emphasis on handling missing data and utilizing advanced machine learning techniques for robust estimations.

This study incorporates key datasets that were accessed after meeting all ethical and regulatory requirements. These include four widely used open access ICU datasets, a specialized single-center ICU dataset, and a comprehensive primary care dataset from the Clinical Practice Research Datalink (CPRD). These datasets cover a variety of clinical scenarios, from high-acuity ICU patients to population-level data in primary care, offering a diverse set of features for estimating treatment effects.

The specific datasets are outlined as follows:

- **ICU Datasets:** Four ICU datasets are used to study critically ill patients, focusing on time-series data, treatment administration and patient outcomes. These datasets are highly valuable for understanding acute interventions and their effects on mortality and morbidity.

- **Primary Care Data:** The CPRD dataset allows a population-based analysis of treatment effects, particularly in the management of chronic diseases such as hypertension and their associated risks. This dataset includes longitudinal data, allowing for a detailed study of treatment outcomes over time.

Through careful preprocessing and imputation techniques, we address the high level of missingness in the time-series data, ensuring that the models can reliably

estimate treatment effects. A combination of feature selection, treatment aggregation, and outcome definitions was applied across these datasets, tailored to their specific characteristics and the clinical questions being addressed. In Table 3.1, we present an overview of the datasets used in various chapters of the thesis.

Table 3.1: Comparison of Datasets Used in the Thesis

Dataset	Type of Data	Outcome Studied	Care Setting	Tasks Used For	Sample Size	Country
PhysioNet 2012 [89]	Time-series	Mortality	ICU	Time-series Imputation	12,000	United States
HiRID [69]	Time-series	Mortality	ICU	Time-series Imputation	33,000	Switzerland
eICU [197]	Time-series	Mortality	ICU	Time-series Imputation	54,423	United States
MIMIC III [121]	Time-series	Diastolic Blood Pressure	ICU	Treatment Effects Estimation	60,000	United States
Vietnam ICU Dataset [243]	Tabular	Hospital Acquired Infection	ICU	Generative Modelling for Data Augmentation	364	Vietnam
STRATIFY Primary Care [61]	Tabular	Falls, AKI, Syncope	Primary Care	Treatment Effects Estimation	4,000,000	United Kingdom

Since missingness of EHRs is an important factor in this thesis, we describe each of the time-series datasets in terms of missingness. The datasets are pre-processed so that the measurements taken in each hour are aggregated, resulting in 48 time steps for each feature. If the feature was measured more than once, the arithmetic mean of the measurement was recorded each hour. The value is considered missing if the feature is not measured in that respective hour. We formally define two types of missingness in EHRs as follows:

- **Feature-wise missingness**: This refers to the percentage of patients who have no observations for a specific feature. It captures the extent to which a particular feature is completely missing for a subset of patients. In this case, the missingness is evaluated at the level of the feature for each patient, highlighting if some features are completely absent in certain patient records.

- **Sample-wise missingness**: This refers to the overall percentage of missing values in all patient records, features, and time steps. It is calculated as the ratio of the total number of missing entries to the total number of possible entries. This metric captures how much data is missing at the level of individual data points, giving a sense of the overall missingness across the dataset.

The following sections describe each dataset in detail, followed by an analysis of the statistical properties of patient populations and key information on missingness patterns within these large-scale EHR datasets.

3.2 PhysioNet Challenge 2012 Dataset

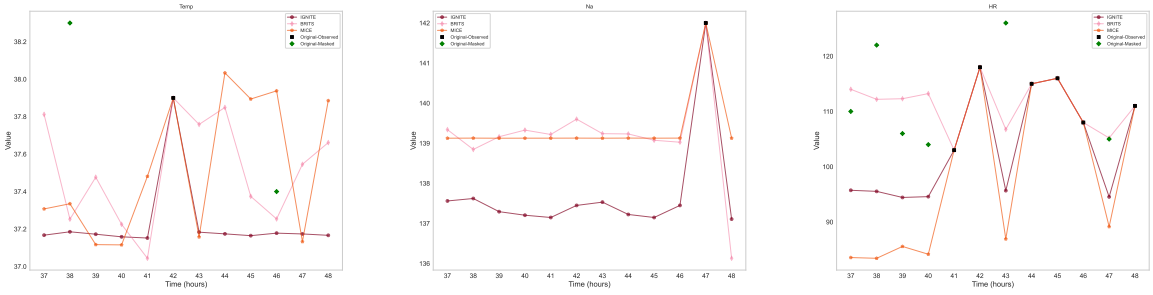
The PhysioNet Challenge 2012 dataset is a well-known resource comprising physiological time-series data from 12,000 ICU patients [230]. It is widely used in research for

clinical machine learning and time-series imputation, providing a realistic representation of ICU data characterized by high sparsity and variability in feature recording frequencies.

Data Composition

The dataset includes up to 42 features for each patient, categorized into static and time-series variables. The six static variables are collected at the time of admission to the ICU and include demographic and admission-specific information, such as age, sex, and type of ICU. The remaining variables are time-series features that represent physiological measurements and administered treatments recorded during the first 48 hours after admission to the ICU.

Each time-series observation is associated with a timestamp indicating the elapsed time from admission to the ICU. Key features include physiological measurements such as heart rate (HR), temperature (Temp), and blood pressure, as well as laboratory values such as albumin and alkaline phosphatase (ALP). Treatment features, including the use of mechanical ventilation, are also recorded. The dataset includes a diverse ICU population admitted for various medical, surgical, cardiac, and trauma conditions. To ensure sufficient data coverage, stays in the ICU for less than 48 hours were excluded. In addition, patients with Do Not Resuscitate (DNR) or Comfort Measures Only (CMO) directives were omitted. These selection criteria result in a dataset focused on patients requiring substantial care and monitoring from the ICU. We show an example of vital signs and laboratory tests for a patient over time, plotted in the Figure 3.1.



(a) Temperature (b) Na (c) Heart Rate

Figure 3.1: Visualized Patient Time-series Variables in Physionet 2012 Dataset. The original observed values are shown as black.

Missingness Characteristics

The dataset demonstrates significant missingness, with an overall missingness rate exceeding 80%. However, the level of missingness varies considerably across features. We analyze the missingness of features in the time-series data and categorize them into two types: feature-wise and sample-wise missingness (Table 3.2 and Table 3.4). The PhysioNet dataset, for instance, has 98.36% missingness for ALP at the sample level, indicating high levels of missing data for this feature, while heart rate (HR) had minimal missingness (9.89%).

3.3 eICU Collaborative Research Database

The eICU Collaborative Research Database is a large, multi-center repository of deidentified critical care data collected from 208 hospitals in the United States between 2014 and 2015 [197]. This database comes from Philips Healthcare's eICU telehealth system, which provides continuous remote monitoring and decision support to ICU patients, complementing the bedside care teams. By aggregating data from diverse patient populations and clinical settings, the eICU database offers a rich resource to investigate a wide range of research questions, including predictive modeling, outcomes analysis, and quality improvement initiatives.

Database Composition and Structure

The eICU Collaborative Research Database includes health information for over 200,000 admissions to the ICU, which yields approximately 827 million observations. Data are stored in a relational format with multiple tables linked by primary and foreign keys. This structure encompasses:

- **Patient Demographics and Administrative Data:** Deidentified identifiers, admission and discharge details, ICU unit type, and hospital metadata.
- **Vital Sign Measurements:** Time-stamped records of parameters such as blood pressure, heart rate, respiratory rate, and temperature, typically captured at regular intervals or as clinically indicated.
- **Laboratory Results:** Serum chemistry values (e.g., electrolytes, kidney and liver function tests), arterial blood gases, hematologic parameters, and other diagnostic results.

Table 3.2: Description of Included Features of PhysioNet Challenge 2012 dataset

	Time-series Missingness	
Feature	**% Sample-level**	**% Feature-level**
ALP	98.36	57.53
HR	9.89	1.53
DiasABP	45.87	29.78
Na	92.90	1.77
Lactate	95.89	45.20
NIDiasABP	57.97	12.62
PaO2	88.49	24.55
WBC	93.27	1.78
pH	87.94	24.09
Albumin	98.74	59.42
ALT	98.32	56.56
Glucose	93.18	2.45
SaO2	95.99	55.28
Temp	62.91	1.54
AST	98.32	56.55
Bilirubin	98.30	56.55
BUN	92.77	1.53
RespRate	75.96	72.27
Mg	92.91	2.43
HCT	90.51	1.58
SysABP	45.86	29.78
FiO2	84.32	32.37
K	92.42	2.08
GCS	67.98	1.54
Cholesterol	99.83	92.10
NISysABP	57.94	12.40
TroponinT	98.92	78.03
MAP	46.18	29.93
TroponinI	99.80	95.29
PaCO2	88.47	24.54
Platelets	92.64	1.65
Urine	30.80	2.58
NIMAP	58.55	12.80
Creatinine	92.73	1.53
HCO3	92.92	1.73

- **Medications and Treatments:** Records of drug administrations, including vasopressors, antibiotics, sedatives, and other critical medications, as well as detailed care plans (e.g., oxygen therapy, mechanical ventilation).
- **Diagnoses** ICD-based 9 codes for diagnosis, comorbidities, complications, and care interventions.

The hospital and unit identifiers have been removed, and the dates have been shifted or generalized to prevent reidentification. This comprehensive deidentification strategy enables data sharing across institutions and supports large-scale multicenter research while ensuring patient privacy.

Preprocessing and Study Inclusion Criteria

For this study, we followed a preprocessing approach similar to that of Li et al. [149], applying the following inclusion criteria to define our final cohort:

1. **ICU Stay Duration:** Patients must have been admitted for at least 48 hours to ensure adequate temporal data for time-series analyzes.

2. **Mortality Window:** Encounters require at least 24 hours of continuous data without a mortality event, allowing sufficient time for predictive modeling.

Applying these criteria yielded 54,423 patient encounters. This filtered set of admissions focuses on clinically relevant stays in the ICU with sufficient data density to support robust analysis.

Feature Selection and Intervention Aggregation

From the extensive range of variables available in the eICU dataset, we identified 55 time-series physiological features that capture trends in core vital signs and laboratory measurements that are consistently recorded at multiple sites. In addition, we extracted 19 discrete-valued medical interventions, grouped into six main treatment categories based on clinical function. To consolidate these interventions into treatment concepts suitable for our predictive modeling, we applied a prevalence threshold of at least 10% in the study cohort. As a result, three high-level treatment categories remained in the final analysis: Oxygen Therapy, Vasopressors and antibiotics.

Missingness in eICU

The features included for the eICU dataset, along with the sample-wise and feature-wise missingness statistics, are shown in Table 3.3. The eICU dataset, which includes more than 200,000 admissions to the ICU from multiple hospitals, presents a slightly different pattern of missingness (Table 3.3). The high number of included variables results in notable variability in both feature-wise and sample-wise missingness. For example, glucose measurements have a sample-wise missingness of 83.53%, indicating substantial gaps in the time-series data. Similarly, key vital signs such as heart

rate exhibit considerable sample-wise missingness (34.00%). However, feature-wise missingness is lower for essential variables such as systolic blood pressure (Noninvasive BP Systolic), which has a feature-wise missingness of only 1.12%, indicating that this variable is consistently recorded for most patients, although with temporal gaps. However, advanced laboratory measurements, such as troponin-I (99.66% feature-wise missingness), are rarely recorded, reflecting the specific clinical focus on certain types of patients. This type of missingness suggests that some features may not be essential for every admission to the ICU, but are crucial for specific diagnoses or conditions.

Data Utility

The eICU Collaborative Research Database of ICU practices at more than 200 sites, combined with robust patient-level granularity, makes it an invaluable resource for evaluating and validating predictive algorithms. The final subset of 54,423 encounters with 55 carefully chosen features provides a balanced foundation for analyzing patient trajectories, understanding disease progression, and identifying the impact of key interventions. By leveraging such a large, multi-center dataset, researchers can increase the generalizability of their findings and better account for variations in clinical practice and patient populations.

3.4 HiRID Dataset

The HiRID dataset is a single center critical care database collected in the Intensive Care Medicine Department of the Bern University Hospital, Switzerland. It contains data for over 33,000 patients admitted for various medical, surgical, and critical care reasons. Similarly to the eICU dataset, HiRID offers a wide range of physiological measurements and clinical documentation, but from a single institution, providing a complementary view of ICU data compared to multi-center databases.

Preprocessing and Study Inclusion Criteria

In line with our inclusion criteria, we restricted the dataset to encounters with:

1. A minimum stay in the ICU of 48 hours.
2. At least 24 consecutive hours without an in-hospital mortality event.

Feature Selection and Intervention Aggregation

After applying the selection criteria above, we obtained a subset of encounters with 50 time-series physiological features. In addition, we identified medical intervention data, applying a *treatment concept* grouping strategy. Specifically, we included those treatments that had a prevalence of at least 10% in the final cohort, resulting in seven primary treatment categories. Namely, oxygen therapy, crystalloids, vasopressors, vasodilators, insulin, painkillers, and anticoagulants.

Missingness Patterns

Despite its comprehensive coverage of admissions to the ICU, HiRID exhibits variability in both feature-wise and sample-wise missingness. As shown in Table 3.4, certain core physiological measurements, such as heart rate (HR), are nearly fully recorded (feature-wise missingness of 0.00%), indicating consistent documentation for most patients. However, more specialized or advanced measurements (e.g., pulmonary blood pressure systolic, mixed venous oxygen saturation) demonstrate high feature-wise missingness (over 90%), reflecting their restricted application to specific subsets of patients.

Sample-wise missingness also differs substantially across features. Common vitals such as respiratory rate show moderate sample-wise gaps (32.66%), whereas body temperature measurements reach 78.08% sample-wise missingness, indicating sporadic recording frequency over time. This irregularity in data collection is typical in real-world ICU settings, where measurements are taken according to clinical protocols rather than rigid schedules.

Data Utility

The HiRID dataset provides a single-institution perspective on ICU care, complementing the multi-center scope of eICU. Although certain physiological measurements and interventions are consistently recorded, others show substantial missingness, reflecting the reality of selective clinical measurements. The final subset of 50 time-series physiological features and seven aggregated treatment categories (39 discrete intervention features) ensures both diversity in captured variables and sufficient data prevalence for meaningful analysis. This balance between breadth and depth makes HiRID a valuable resource for evaluating imputation strategies, modeling clinical trajectories, and studying the impacts of various treatments under more controlled institutional conditions.

3.5 MIMIC-III Dataset

The Medical Information Mart for Intensive Care (MIMIC-III) dataset is a comprehensive repository of deidentified health data from intensive care units (ICUs) in the United States. It offers a rich resource for analyzing critical care practices and evaluating clinical interventions [120]. In the following, we describe its composition, preprocessing steps, outcome focus, and utility in this study.

Preprocessing and Study Inclusion Criteria

To ensure consistency and relevance for the analysis, we implemented the following preprocessing and inclusion criteria:

1. Selected stays in the ICU for at least 30 hours, truncating longer stays at 60 hours to standardize data length while capturing meaningful trends.

2. Focused on patients with complete data for key covariates and outcomes, excluding records with significant missingness or irregular sampling.

3. Extracted 25 time-varying covariates, including vital signs (e.g., heart rate, oxygen saturation) and laboratory results (e.g., blood gas levels, metabolic panels).

For temporal prediction tasks, we structured the data to accommodate varying projection horizons τ. Sub-trajectories were extracted using a rolling origin strategy, and masking techniques were applied to prevent data leakage during multi-step predictions.

Data Composition and Preparation The MIMIC-III dataset provides detailed temporal data that includes demographics of the patient, vital signs, laboratory measurements, and treatment records. For this study, 25 time-varying covariates were selected, focusing on key physiological and laboratory variables. Temporal features were aligned at regular intervals to enhance comparability and simplify downstream modeling.

Outcome of Interest The primary outcome of interest was diastolic blood pressure, a critical marker for the management of tissue perfusion and the avoidance of complications related to hypoperfusion. Accurate prediction of blood pressure informs clinical decision-making, particularly concerning the administration of vasopressors and mechanical ventilation. These treatments were identified as key interventions with potential impacts on patient outcomes.

Data Utility The MIMIC-III dataset's comprehensive coverage of ICU patient journeys makes it a vital resource for evaluating time-series individualized treatment effects (ITE). Its detailed temporal structure and rich feature set enable the development and validation of predictive models for treatment optimization in critical care.

Table 3.3: Description of Included Features of eICU dataset

	Time-series Missingness	
Feature	**% Sample-wise**	**% Feature-wise**
glucose	83.53	3.87
SpO2	45.19	15.19
Noninvasive BP Diastolic	19.40	1.12
Noninvasive BP Systolic	19.40	1.12
Noninvasive BP Mean	19.80	1.42
RR	36.36	8.26
-basos	98.20	50.07
-eos	98.10	47.34
-lymphs	97.96	44.91
-monos	97.97	45.14
-polys	98.18	50.36
ALT (SGPT)	98.73	62.18
AST (SGOT)	98.71	61.69
BUN	95.62	6.05
FiO2	99.06	78.71
HCO3	98.95	76.76
Hct	95.64	7.88
Hgb	95.59	7.94
MCH	96.39	15.38
MCHC	96.21	10.94
MCV	96.20	10.93
PT	98.93	71.43
PT - INR	98.89	70.55
PTT	99.15	82.92
RBC	96.13	8.74
RDW	96.38	14.63
WBC x 1000	96.12	8.47
albumin	98.49	56.90
alkaline phos.	98.72	62.11
anion gap	96.49	25.14
bicarbonate	95.84	10.83
calcium	95.76	8.73
chloride	95.58	5.83
creatinine	95.60	5.94
lactate	99.53	88.61
magnesium	97.46	39.11
pH	98.90	75.86
paCO2	98.90	75.79
paO2	98.88	75.50
phosphate	98.35	57.94
platelets x 1000	96.10	8.89
potassium	94.81	5.34
sodium	95.32	5.70
total bilirubin	98.75	62.82
total protein	98.72	61.92
troponin - I	99.66	91.34
Base Excess	99.13	80.24
urinary specific gravity	99.84	93.12
Heart Rate	34.00	8.29
Temperature	75.99	6.12
Tidal Volume (set)	97.24	84.81
MPV	97.42	39.02
Exhaled MV	98.31	89.33
Exhaled TV (patient)	98.70	89.80
SaO2	97.74	81.36

Table 3.4: Description of Included Features of HiRID Dataset

	Time-series Missingness	
Feature	**% Sample-wise**	**% Feature-wise**
Heart rate	1.44	0.00
Body Temperature Core	78.08	38.45
Arterial Blood Pressure systolic	11.23	4.12
Arterial Blood Pressure diastolic	11.25	4.12
Arterial Blood Pressure mean	11.23	4.09
Non-invasive Blood Pressure systolic	89.82	37.85
Non-invasive Blood Pressure diastolic	89.82	37.86
Non-invasive Blood Pressure mean	89.82	37.89
Pulmonary Blood Pressure mean	92.18	84.95
Pulmonary Blood Pressure systolic	92.17	84.94
Pulmonary Blood Pressure diastolic	92.32	84.97
Cardiac Output	92.23	85.18
mixed venous oxygen saturation	92.34	85.37
central venous pressure	37.26	29.32
ECG ST1	38.97	32.14
ECG ST2	56.94	36.75
ECG ST3	64.99	43.55
Saturation Oxygen peripheral/capillary	3.95	0.03
End tidal CO2 concentration	68.95	43.68
Respiratory rate	32.66	0.07
Hourly urinary output	73.52	1.48
ICP	94.15	90.97
Measurement of output from drain	96.65	92.36
Base excess in Arterial blood by calculation	90.60	14.77
Carboxyhemoglobin in Arterial blood	90.65	15.29
Hemoglobin in Arterial blood	91.52	24.18
Bicarbonate in Arterial blood	90.60	14.77
Lactate in Arterial blood	90.67	14.04
Methemoglobin in Arterial blood	90.65	15.30
a_pH	91.43	24.73
a_pCO2	91.63	24.93
a_PO2	91.52	24.82
a_SO2	90.65	15.27
Potassium	89.48	2.19
Sodium	89.40	2.10
Cl-	90.84	15.89
Ca2+i	90.61	15.26
Phosphate	97.12	22.66
Magnesium in Blood	97.27	22.95
Urea	96.98	21.28
Crea	96.36	8.49
INR	97.89	45.10
Glucose	79.55	0.94
C-reactive protein	96.53	8.96
Hemoglobin in Blood	95.49	3.83
Total white blood cell count	95.55	3.73
Platelet count	95.71	4.57
MCH	95.57	3.81
MCHC	95.57	3.87
MCV	95.56	3.80

3.6 Vietnam Tropical Hospital ICU Dataset

This study uses a dataset obtained from the Ho Chi Minh City Hospital for Tropical Diseases in Vietnam, which has been made openly available for research [243]. A total of 364 patients are included, each of whom was admitted to the hospital's Intensive Care Unit (ICU) and remained there for at least 48 hours. Given the regional prevalence of tropical infections, the dataset provides a focused lens on critical care practices in an environment where diseases such as dengue and other infectious conditions are common.

Data Composition and Categories

All variables in this dataset were collected at admission to the ICU and can be categorized into three principal groups: *comorbidities*, *demographics*, and *admitting diagnoses*.

- **Comorbidities:** This category includes diabetes (reported in 9.62% of patients), steroid use (4.12%), chronic liver disease (15.11%), and chronic kidney disease (0.82%). In particular, the relatively low incidence of chronic kidney disease indicates that very few patients required specialized renal support prior to or upon admission.

- **Demographics:** Of the 364 patients, 66.48% were female, highlighting a noticeable predominance of female patients in this cohort. In terms of age distribution, 36.54% of the sample fell into the 16–44 range, while 39.01% were between 45–59 years old, and 24.45% were aged 60 years or older. This age breakdown underlines a slightly higher proportion of middle-aged adults, although a significant minority still consists of older adults.

- **Admitting Diagnosis:** Primary reasons for ICU admission include tetanus (4.67%), sepsis (12.36%), local infections (20.60%), dengue (56.04%), and internal medicine diseases (6.32%). In particular, dengue alone accounts for more than half of all admissions. Meanwhile, local infections comprise a cluster of conditions such as pneumonia, cellulitis, urinary tract infection, and spontaneous bacterial peritonitis. In contrast, the internal medicine category includes various diagnoses, including kidney failure, myocarditis, myocardial infarction, malignant hypertension, diabetic ketoacidosis, and epilepsy [243].

Table 3.5 summarizes these categories along with the prevalence of each feature. By highlighting both the most common and the least frequent conditions, the table underscores the diversity within this ICU population, encompassing both infectious and non-infectious etiologies.

Outcome of Interest

The primary outcome of interest is the acquisition of hospital acquired infections (HAI) during the ICU stay. These infections include pneumonia, bloodstream infections, and urinary tract infections. Of the 364 patients, 86 (23.60%) developed at least one HAI. This figure suggests that nearly one in four ICU patients in this cohort encountered an HAI, highlighting the need for stringent infection control measures within such settings.

Clinical Relevance and Data Utility

The Vietnam Tropical Hospital ICU dataset offers critical information on patient management and disease patterns within a single-center context, especially in a region where tropical and emerging infections are prevalent. Although it involves a relatively moderate cohort size (364 patients), it encompasses a wide array of clinical presentations, from local infections (20. 60%) to conditions such as dengue (56. 04%). Understanding the interplay of preexisting comorbidities, such as chronic liver disease (15.11%), with the risk of HAIs can help refine prevention and management strategies. Furthermore, the substantial presence of women (66.48%) in this population raises additional questions about sex-based differences in disease progression or response to treatment.

Overall, these data provide a valuable platform for studying both infectious and non-infectious outcomes in a tropical clinical environment. The results generated from this dataset may guide the development of targeted clinical interventions, improve local ICU practices, and inform broader global health strategies, particularly those focusing on infection prevention, patient risk stratification, and early intervention protocols.

Table 3.5: List of Included Patient Features and Their Prevalence in the Vietnam Tropical Hospital ICU Dataset (n = 364).

Comorbidities (n, %)	
Diabetes	35 (9.62%)
Steroids	15 (4.12%)
Chronic Liver	55 (15.11%)
Chronic Kidney	3 (0.82%)
Demographics (n, %)	
Female	242 (66.48%)
Age	
16–44	133 (36.54%)
45–59	142 (39.01%)
60+	89 (24.45%)
Admission Diagnosis (n, %)	
Tetanus	17 (4.67%)
Sepsis	45 (12.36%)
Local Infections	75 (20.60%)
Dengue	204 (56.04%)
Internal Medicine	139 (6.32%)
Outcomes (n, %)	
Hospital-Acquired Infections	86 (23.60%)

3.7 CPRD STRATIFY Dataset

A dataset for the STRAtifying Treatments In the multi-morbid Frail elderlY (STRATIFY) project, which aims to investigate the treatment effects and potential harms of antihypertensives based on data extracted from the Clinical Practice Research Datalink (CPRD). Our collaborator, Prof. James Sheppard leads the STRATIFY project from the Nuffield Department of Primary Care Health Sciences at the University of Oxford. We are granted access to longitudinal data extracted from CPRD's primary care data for more than 4,000,000 patients from the UK. The included treatments were Antihypertensive therapy such as ACE inhibitors, calcium channel blockers, thiazides and diuretics similar to thiazides, beta-blockers, alpha-blockers, vascular vasodilators, and renin inhibitors.

Descriptive Statistics

In this subsection, we present the descriptive statistics for the STRATIFY data used in this study. These statistics include demographics, clinical indicators, medical history, and outcomes for patients in multiple datasets.

Demographics

The patient demographics provide a general overview of the population involved in the study. As shown in Table 3.6, the total number of patients is 3,834,056, divided into two groups: 478,737 in the treatment group and 3,355,319 in the control group.

- **Age:** The mean age of the population is 58.87 years (SD: 13.24). Patients in the intervention group had a mean age of 65.32 years compared to 56.44 years in the control group.
- **Gender:** Females represented 48.34% of the overall population. The proportion was almost equal between the intervention groups (48.42%) and the control groups (48.31%).
- **BMI:** The mean BMI was 27.51 (SD: 5.42), and the patients in the intervention group had a slightly higher BMI (28.62) than those of the control group (27.08).

Clinical Indicators

Clinical indicators, particularly related to blood pressure, are crucial to understanding the baseline health conditions of patients (Table 3.6).

- **Systolic Blood Pressure (SBP):** The mean SBP at inclusion was 142.59 mmHg (SD: 11.67), and the patients in the intervention group had a significantly higher mean SBP (150.95 mmHg) compared to the control group (141.39 mmHg).
- **Diastolic Blood Pressure (DBP):** The mean DBP was 83.85 mmHg (SD: 9.51), and the intervention group showed a higher mean DBP of 88.28 mmHg versus 83.22 mmHg for the control group.

Medical History

Patients' medical history was evaluated based on various chronic conditions (Table 3.6). The prevalence of these conditions helps to understand the overall health burden of the population.

- **Cardiovascular conditions:** Conditions such as myocardial infarction (1.16%), stroke (1.60%), and heart failure (0.70%) were present in the population. The intervention group had a significantly higher prevalence of these conditions.
- **Diabetes:** Overall, 5.78% of patients were diagnosed with diabetes, with a higher proportion in the intervention group (14.77%) compared to the control group (4.5%).

Outcomes

Key health outcomes, such as falls, fractures, and acute kidney injury (AKI), are shown in Table 3.6. These outcomes are essential to understand the risks associated with antihypertensive treatment.

3.8 Summary

In this chapter, we have provided a detailed description of the datasets used in this thesis. These datasets, which span both the ICU and primary care settings, offer a comprehensive basis for estimating treatment effects in a range of clinical contexts. Datasets cover diverse patient populations, varying in both geographic location and

Table 3.6: Data Exploration Table with Patient Characteristics for the Antihypertensives Case Study

Variable	Total	Intervention	Control
Patients, n	3,834,056	478,737	3,355,319
Demographics			
Age, mean (SD)	58.87 (13.24)	65.32 (13.21)	56.44 (12.4)
Female, n (%)	2,696,935 (48.34)	737,605 (48.42)	1,959,330 (48.31)
BMI, mean (SD)	27.51 (5.42)	28.62 (5.77)	27.08 (5.22)
Clinical Indicators, mean (SD)			
Systolic Blood Pressure at Inclusion	142.59 (11.67)	150.95 (13.69)	141.39 (10.84)
Diastolic Blood Pressure at Inclusion	83.85 (9.51)	88.28 (11.78)	83.22 (8.96)
Cardiovascular Disease Risk Score	0.12 (0.12)	0.22 (0.15)	0.11 (0.11)
Electronic Frailty Index	0.05 (0.05)	0.08 (0.06)	0.04 (0.05)
Medical History, n (%)			
Myocardial Infarction	44,453 (1.16)	25,880 (5.41)	18,573 (0.55)
Stroke	61,202 (1.60)	19,907 (4.16)	41,295 (1.23)
Heart Failure	25,900 (0.70)	13,779 (2.88)	12,121 (0.36)
Transient Ischemic Attack (TIA)	28,713 (0.75)	9,368 (1.96)	19,345 (0.58)
Peripheral Vascular Disease (PVD)	22,401 (0.58)	7,720 (1.61)	14,681 (0.44)
Agina	250,656 (4.49)	31,331 (6.54)	26,841 (0.80)
CABG	9,574 (0.25)	6,169 (1.29)	3,405 (0.10)
Diabetes	221,655 (5.78)	70,714 (14.77)	150,941 (4.5)
Atrial Fibrillation	58,769 (1.53)	24,197 (5.05)	34,572 (1.03)
Chronic Kidney Disease (CKD)	49,528 (1.29)	21,261 (4.44)	28,267 (0.84)
Cancer	140,892 (3.67)	23,897 (4.99)	116,995 (3.49)
Previous Falls	153,640 (4.01)	21,054 (4.4)	132,586 (3.95)
Previous Acute Kidney Injury (AKI)	4,451 (0.12)	1,434 (0.3)	3,017 (0.09)
Previous Fracture	732,806 (19.11)	89,250 (18.64)	643,556 (19.18)
Previous Hypotension	14,078 (0.37)	3,007 (0.63)	11,071 (0.33)
Previous Syncope	60,928 (1.59)	7,665 (1.6)	53,263 (1.59)
Previous Electrolyte Abnormality	25,751 (0.67)	6,911 (1.44)	18,840 (0.56)
Previous Gout	90,973 (2.37)	18,945 (3.96)	72,028 (2.15)
Outcomes, n (%)			
Falls	368,833 (9.62)	67,382 (14.07)	301,451 (8.98)
Acute Kidney Injury (AKI)	36,833 (9.62)	7,813 (1.63)	24,029 (0.72)
Fracture	354,171 (9.24)	48,387 (10.11)	305,784 (9.11)
Hypotension	46,305 (1.21)	11,654 (2.43)	34,651 (1.03)
Syncope	65,619 (1.71)	11,144 (2.33)	54,475 (1.62)
Gout	139,615 (3.64)	29,392 (6.14)	110,223 (3.29)
Electrolyte Abnormality	88,844 (2.32)	22,959 (4.8)	65,885 (1.96)

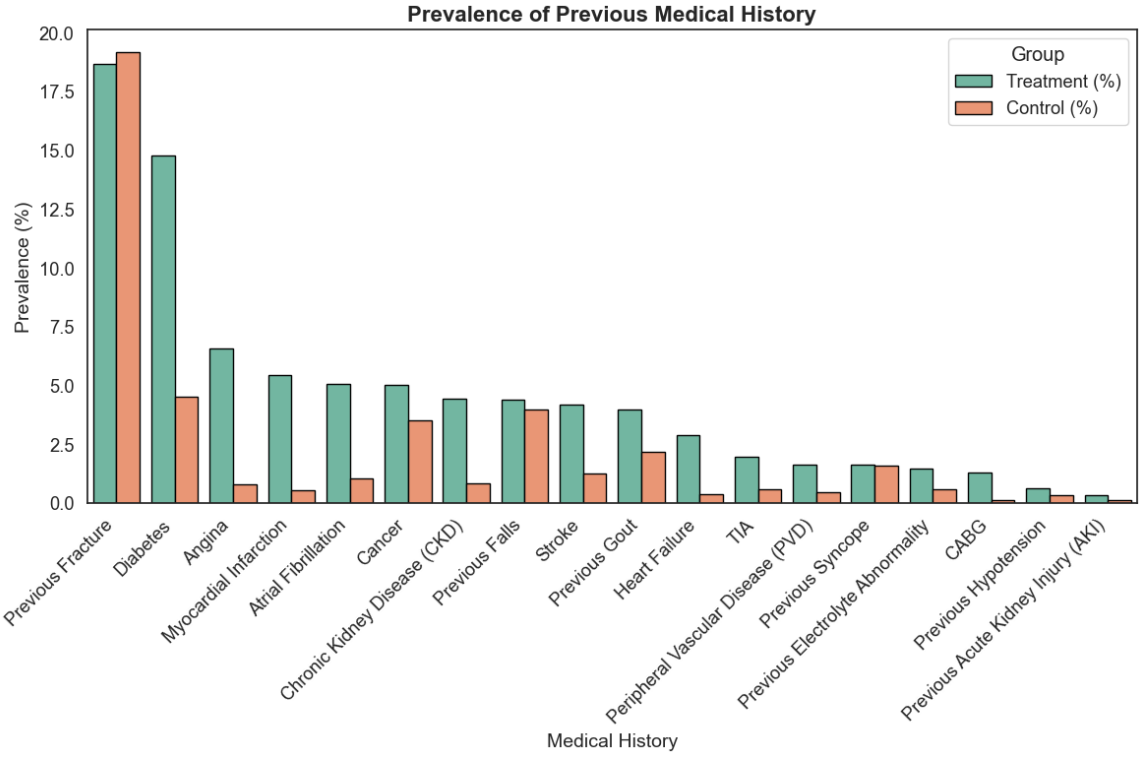


Figure 3.2: Bar plot showing the prevalence of previous medical history indicators (e.g., diabetes, hypertension, heart disease) among patients in the STRATIFY dataset. These indicators serve as baseline covariates in subsequent analyses.

clinical conditions, and include a wide range of features such as demographics, comorbidities, physiological measurements, and treatment records.

The intensive care datasets, including PhysioNet Challenge 2012, eICU, and HiRID, are crucial to studying critically ill patients and understanding the impacts of acute interventions. However, the CPRD primary care dataset enables the exploration of long-term treatment effects in a population-based setting, focusing on the management of chronic diseases. Together, these datasets provide a diverse and rich foundation for applying treatment effects estimation techniques.

One of the key challenges addressed in this chapter is the issue of missingness, particularly in time-series data from ICU settings. By analyzing the extent and patterns of missingness across different datasets, we have outlined the steps taken to preprocess the data and ensure that they are suitable for machine learning models.

The next chapters will build upon this dataset foundation by describing the methodological approach used for synthetic data generation, treatment effects estimation, and other methods including the handling of missing data. The combination of robust datasets and advanced modeling techniques will allow us to derive meaningful

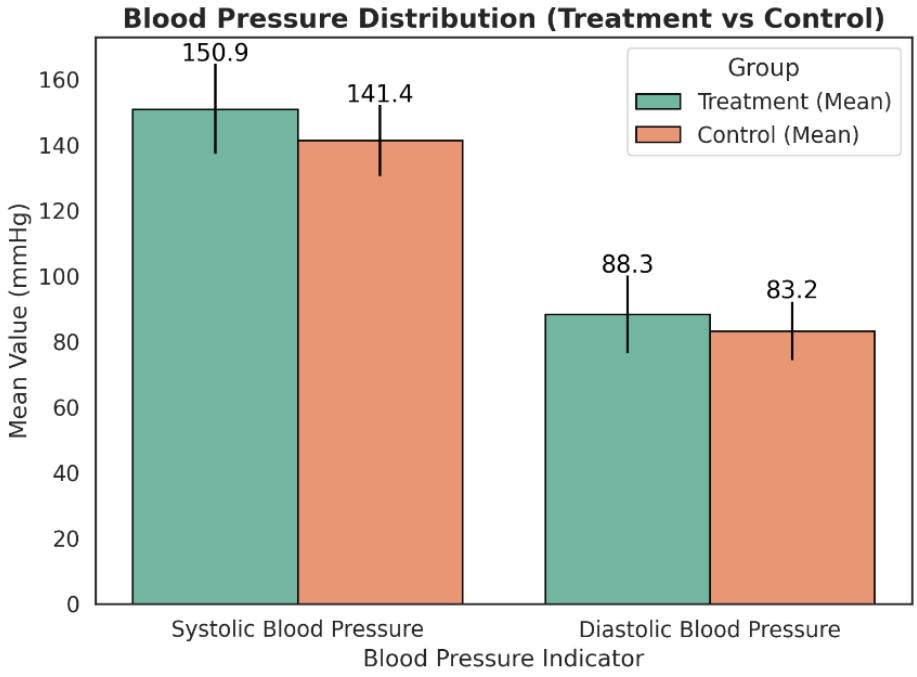


Figure 3.3: Distribution of blood pressure measurements for patients in treatment versus control groups within the STRATIFY dataset. The plot highlights variations in systolic and diastolic readings prior to treatment initiation.

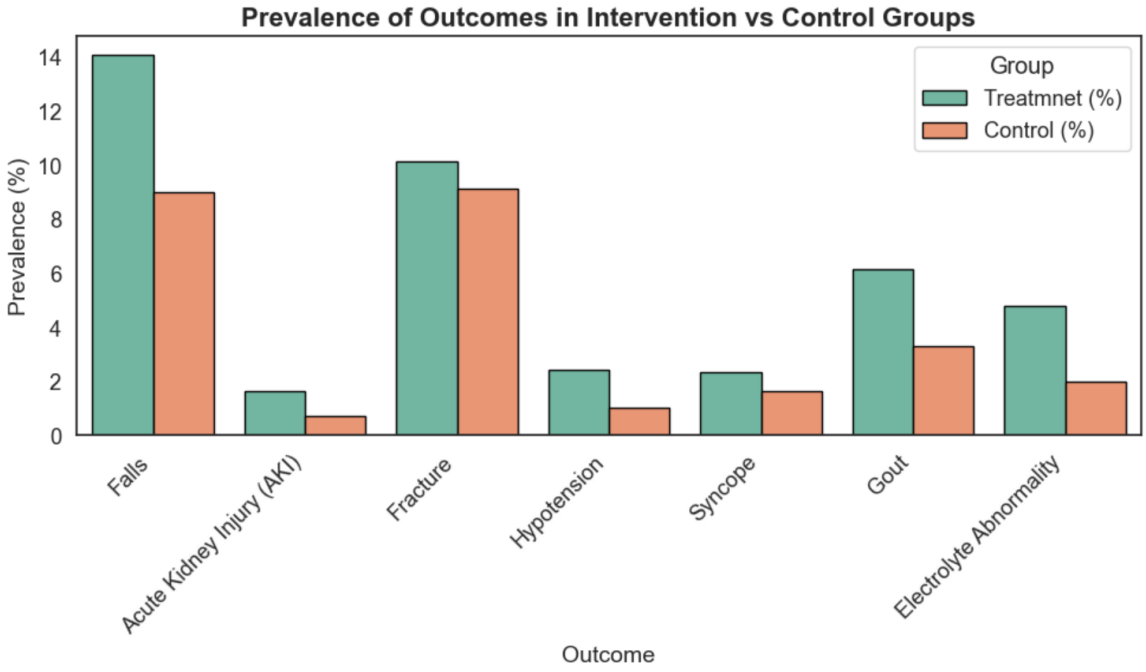


Figure 3.4: Bar chart showing the proportion of patients experiencing each studied outcome in the STRATIFY dataset. This informs outcome modeling in subsequent chapters.

insights into the effects of treatments on patient outcomes.

Chapter 4

Synthetic Patient Generation for Low-Resource Settings

Building on the foundational understanding of the datasets described in the previous chapter, we now explore how generative models can enhance predictive performance in tabular EHR data. Rather than revisiting the challenges of missingness and data sparsity, this chapter focuses on applying and evaluating generative approaches to improve downstream prediction tasks. Specifically, we investigate how synthetic data generation can mitigate class imbalance and support model training in real-world clinical datasets.

4.1 Introduction

Clinical decision support systems (CDSS) are important tools to promote optimal patient care, safety, and resource use. In low and middle-income countries (LMICs), where an estimated 8,000,000 deaths occur every year as a result of low-quality (but accessible) care, such systems have the potential to have a huge impact [140]. Developing CDSS applications using electronic health records (EHRs) and machine learning (ML) techniques has gained increased interest from the research community [263]. Despite the promising results of many of these applications, the performance of ML models is highly dependent on the availability of training data [119, 248]. ML models tend to be data hungry, and can easily overfit and underperform when trained on a small dataset [119, 248].

Most CDSS have been developed in high-income countries using massive datasets from EHRs [2, 56]. Consequently, they do not support decision making in many diseases prevalent in low-resource settings, resulting in an unmet need for ML research applications that are developed and validated for low-resource settings. Even if

CDSS addresses problems common to all resource settings, those developed from high-income datasets are usually unsuitable for direct deployment in LMIC settings due to differences in the prevalence of diseases and demographic distribution [5, 15, 71], and require adaptation using data from these populations [211]. For example, hospital-acquired infections (HAI) are well-established markers of healthcare quality, as well as being significant causes of mortality and morbidity in patients around the world. They are a particular concern in LMICs and the prediction of CDSS of those at risk of HAI would be of great value in improving patient outcomes. However, since HAI is closely related to the local context, the development of these is particularly dependent on high-quality local data. EHRs are rarely available in LMICs [177] and their healthcare systems often suffer from constraints in infrastructure and diagnostic capacity [77], frequent changes in strategic healthcare policies, and political instability [181], all of which could impact the quantity and quality of routine healthcare data collected from such clinical settings. Manual collection of high-quality large-scale data is unfeasible in terms of cost and personnel. Data dependency inhibits the optimal development and utilization of CDSS, specifically in resource-constrained clinical settings.

Many of the current medical statistics and data-driven models rely on methods such as Synthetic Minority Oversampling TEchnique (SMOTE) which oversample the training data, especially in imbalanced settings. Oversampling methods could introduce flawed correlations and dependencies between samples and result in limited data variability [73], all of which could severely underperform in testing environments. Recent works in deep learning research proposed generative models that learn the underlying data distribution and generate realistic looking data while preserving the privacy of the original samples. Those deep generative models, including Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs) [137, 92] have been originally proposed and validated for the imaging domain where quantitative and qualitative evaluation by experts could not differentiate the real images from those generated by the models. Although very relevant and highly needed, the use of deep generative models for the synthesization of EHRs for low-resource clinical applications is often not discussed or motivated in most of the proposed works [84].

To this end, this chapter proposes synthetic data as a solution for developing models based on small datasets collected from LMIC countries. To do so, we trained a GAN-based model to learn the underlying data distribution and generated synthetic samples that could be used for training purposes. We refer to our proposed synthetic data generation and predictive modeling framework as SynthEHR-LMIC, designed

specifically for small, tabular ICU datasets from LMICs. SynthEHR-LMIC is built on a modified medGAN architecture and includes interpretability and size-sensitive performance evaluation. Specifically, we utilize a small already published dataset (364 patients) collected from an Intensive Care Unit in Vietnam [243], with variables collected at admission and a binary outcome indicating if the patient had a hospital-acquired infection. With the increasing burden of antimicrobial resistance, especially in LMICs, it is vital to develop risk scores to predict the probability of developing such infections. This could allow clinical personnel to take antiseptic measures, reduce unnecessary antibiotic prescriptions, and introduce timely interventions to prevent a prolonged stay. The proposed method provides a plausible solution that could be used to develop diagnostic models despite data scarcity in LMICs. Our contributions could be summarized as follows.

1. **Deep generative models for LMICS.** For the first time, we demonstrate the feasibility of using generative models to synthesize data that are used to develop ML models from small datasets from LMIC healthcare settings.

2. **Comprehensive data utility evaluation.** We evaluate the utility of synthetic data compared to other commonly used approaches and demonstrate superior performance using models trained on synthetic data. We also show the impact of the synthetic tabular data size on the performance of the predictive model in a series of experiments where the synthetic data training size is varied.

3. **Interpretability analysis**: We conducted a post-hoc SHapley Additive exPlanations (SHAP) interpretability analysis to investigate the impact of using various training sets on the feature importance in the test set predictions, which is the new approach for evaluating deep generative models for EHRs.

The structure of the chapter is as follows. In the methods, we first describe the dataset used in the study, followed by an explanation of the model used to generate the synthetic data. The other subsections in the materials and methods discuss the predictive modeling task and the baseline methods used to compare the performance of the proposed model, followed by an overview of the interpretability analysis. In the results section, we present the performance of the models and the findings of the feature importance analysis. In the discussion, we interpret the findings, discuss the limitations and strengths of the work, and outline the future outlook for related research directions.

4.2 Related Works

Generative models have gained increasing traction in healthcare for their potential to augment limited data and improve model generalizability. Early work on synthetic data generation for clinical applications largely focused on oversampling methods like SMOTE [73], which artificially balance class distributions by interpolating between minority samples. While simple and computationally efficient, these approaches often fail to capture the true data distribution and may introduce spurious correlations [73].

To overcome these limitations, deep generative models such as VAEs [137] and GANs [92] have been explored for healthcare data synthesis. medGAN [44] was one of the first models to generate realistic EHR data using an adversarial framework with an autoencoder-based generator. Since then, models such as ehrGAN [14] and CorGAN [68] have introduced refinements to better preserve temporal dynamics, covariate dependencies, and clinical plausibility.

Despite progress, most generative modeling efforts have focused on structured EHRs from high-income settings. Studies using MIMIC-III [121] datasets dominate the literature, while real-world applications in LMICs remain limited. Recent reviews [84] emphasize the need for LMIC-specific investigations, especially for small and imbalanced datasets where privacy constraints and resource limitations restrict data availability. In parallel, there is growing recognition of the value of synthetic data in clinical decision support, particularly for early prediction of high-impact events such as HAIs. However, existing studies rarely evaluate generative model utility for such diagnostic applications in data-constrained environments. Moreover, few works explicitly analyze the interpretability of predictive models trained on synthetic data, an essential factor for clinical adoption.

Our work addresses these gaps by proposing SynthEHR-LMIC, a generative framework tailored to small ICU datasets from LMICs. Unlike prior approaches, we focus not only on data realism but also on downstream predictive performance and interpretability. This positions our work as one of the first to bridge deep generative modeling and practical CDSS development in resource-constrained clinical settings.

4.3 Methodology

4.3.1 Dataset Description

The data used in this work is collected from Ho Chi Minh City Hospital for Tropical Diseases, Vietnam, and are released for open access [243]. The patients included in

this study are 364 patients, a total of which were all admitted to the ICU and stayed at least two days. The variables included are those readily available on admission to the ICU, which we categorize into comorbidities, demographics, and admission diagnosis. The admission diagnosis included one of five categories: (1) Tetanus, (2) Sepsis, (3) Local infections, (4) Dengue, and (5) Internal Medicine disease. According to the original study documentation, local infections included cases of pneumonia, cellulitis, urinary tract infection, and spontaneous bacterial peritonitis, while internal medicine diseases included kidney failure, myocarditis, myocardial infarction, malignant hypertension, diabetic ketoacidosis, and epilepsy [243]. The outcome of interest is a binary label that indicates if the patient acquired an infection during their stay in the ICU. The infections acquired included in the dataset were pneumonia, bloodstream infection, and urinary tract infection, all of which were defined according to the 2014 Centers for Disease Control and Prevention criteria [1]. More details on this dataset can be found in Chapter 3, Section 3.6.

4.3.2 Synthetic Data Generation

To explore the feasibility of synthetic data generation for training purposes, we applied a random stratified train-test split, dividing the data into a 70-30 split. This allows us to dedicate a test set for model evaluation while using the training data to develop the generative model, ensuring a robust evaluation of the downstream predictive model [82]. Although K-Fold Cross Validation is a viable option, the 70-30 split better maintains an independent held-out test set, essential for consistent downstream performance comparisons.

Among the range of generative models, such as VAEs and GANs, we select GANs due to their superior capability to produce high-fidelity synthetic data. This often leads to enhanced performance in predictive modeling tasks [180]. Although VAEs are particularly useful for image [143] and time-series generation [147], they are less suitable for discrete, tabular data, such as those in our EHR dataset. For tabular synthetic data generation, we adopt and modify medGAN [45], a model designed to adapt GAN to handle the discrete nature of EHR data through an innovative autoencoder-based structure. In the original GAN framework, the generator G learns through error signals provided by the discriminator D through backpropagation. This setup is optimal for generating continuous values, but introduces limitations when dealing with discrete data, such as patient records $x \in \mathbb{Z}_+^{|C|}$ in our dataset. medGAN addresses this challenge by incorporating an autoencoder, allowing the model to map

input data into a lower-dimensional latent space and reconstruct it accurately, thereby capturing key features of the discrete data.

Our work introduces enhancements to the original medGAN framework to better capture the intricacies of discrete variables in patient records. Specifically, we employ an autoencoder to learn the essential features of these discrete variables, which are then used to map the continuous output of G into meaningful discrete values. This enables gradients to flow efficiently from D to the decoder Dec, supporting end-to-end fine-tuning and higher-fidelity data generation. The modified architecture involves an encoder $Enc(x; \theta_{\text{enc}})$ that compresses the discrete input $x \in \mathbb{Z}_+^{|C|}$ into a continuous representation $Enc(x) \in \mathbb{R}^h$, and a decoder $Dec(Enc(x); \theta_{\text{dec}})$ that reconstructs x from this compressed form. The autoencoder optimizes reconstruction accuracy by minimizing both mean squared error for count variables and cross-entropy loss for binary variables:

$$\frac{1}{m}\sum_{i=0}^{m} \|x_i - x_i'\|_2^2 \tag{4.1}$$

$$\frac{1}{m}\sum_{i=0}^{m} [x_i \log x_i' + (1 - x_i)\log(1 - x_i')] \tag{4.2}$$

where $x_i' = Dec(Enc(x_i))$ and m is the mini-batch size. ReLU activations are used in both Enc and Dec for count variables, while tanh and sigmoid are used for binary encoding and decoding, respectively.

After pre-training, the GAN generates representations in the latent space (the output of Enc) rather than directly producing patient records. The pre-trained decoder Dec then translates these generated latent representations into synthetic discrete records $Dec(G(z))$, while D differentiates between real patient records x and synthetic records $Dec(G(z))$. The medGAN training process is defined by the following gradient updates:

$$\theta_d \leftarrow \theta_d + \alpha \nabla_{\theta_d} \frac{1}{m}\sum_{i=1}^{m} [\log D(x_i) + \log(1 - D(x_{z_i}))] \tag{4.3}$$

$$\theta_{g,\text{dec}} \leftarrow \theta_{g,\text{dec}} + \alpha \nabla_{\theta_{g,\text{dec}}} \frac{1}{m}\sum_{i=1}^{m} \log D(x_{z_i}) \tag{4.4}$$

where $x_{z_i} = Dec(G(z_i))$. To ensure D is trained with discrete values, we can round $Dec(G(z))$ to the nearest integers. However, our experiments showed that training D without rounding resulted in better performance, as described in Section 4.2. Thus, for the remainder of this paper, we assume that D is trained without

rounding. To clarify, the autoencoder was first pre-trained for 200 epochs using the training set to ensure accurate reconstruction of the original records before adversarial training. This pre-training enables the decoder to act as a stable inverse mapping from latent to data space. Although medGAN permits rounding the decoder outputs before passing them to the discriminator to enforce discrete values, we observed empirically that using the continuous decoder outputs led to better convergence and downstream performance. As such, all discriminators in our experiments were trained using unrounded, continuous outputs from the decoder.

During optimization, the decoder parameters θ_{dec} are fine-tuned alongside G, effectively transforming G into a neural network with an additional pre-trained hidden layer, which maps continuous outputs to discrete records. For G, we use ReLU activation functions for all layers except the output layer.

Algorithm 1 Generative Augmentation and Predictive Modeling for Tabular EHRs

Require: Real EHR dataset $\mathcal{D}_{\text{real}} = \{X_{\text{real}}, Y_{\text{real}}\}$, synthetic ratio p_{synth}, generative model type G_{type}
Ensure: Trained prediction model M_{pred}, evaluation metrics
Preprocess $\mathcal{D}_{\text{real}}$ (normalize, encode, split into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$)
Initialize generative model G_θ (e.g., GAN, VAE) **while** *not converged* **do**
    Sample minibatch x_{batch} from $\mathcal{D}_{\text{train}}$
Update G_θ using adversarial or reconstruction loss

Generate noise samples $Z \sim \mathcal{N}(0, I)$
Compute $X_{\text{synth}} = G_\theta(Z)$
Assign labels Y_{synth} (e.g., from classifier or sampling)
Form augmented dataset: $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{train}} \cup \{X_{\text{synth}}, Y_{\text{synth}}\}$
Initialize predictive model M_{pred}
Train M_{pred} on $\mathcal{D}_{\text{aug}}$
Evaluate M_{pred} on $\mathcal{D}_{\text{test}}$ using AUROC, AUPRC, Accuracy
    return M_{pred}, evaluation metrics

4.3.3 Predictive Modeling Task and Baselines

The generated synthetic data is used to train a simple machine learning model to predict hospital-acquired infections during the patient's stay in the ICU. Three different types of machine learning models were evaluated, which were Random Forest [198], Support Vector Machines (SVM) [191] and K-Nearest Neighbor (KNN) [146], respectively. The choice of the three models is motivated by their relative simplicity, with

often comparable performance to many advanced models, making them a good candidate for deployment in hospitals in LMICs. We compare the performance of the models trained on the synthetic data with those trained on the (1) original small training set and (2) oversampled training data using SMOTE. To better understand the impact of the size of the synthetic data on the predictive model performance, we train the GAN model to synthesize data of various sizes in inference. The synthesized data are then used to train the predictive model, where the performance is compared to that of models trained with original and oversampled data. Each of the machine learning models was trained using 3-Fold Stratified K-Fold validation, to choose the best hyperparameters using GridSearch to make predictions on the held-out test set. The hyperparameter ranges used are included in the supplementary material Table A1. We selected 3-fold cross-validation due to the small size of the training set (approximately 255 patients). Using a larger number of folds (e.g., K=5 or 10) would further reduce the number of training samples per fold, increasing variance and risk of overfitting during model selection. Stratified sampling was used to preserve class distribution across folds. Nevertheless, exploring larger K-fold validation remains a future direction as synthetic data sizes increase. We opted not to use a separate validation set to avoid further splitting the already small training set. Instead, cross-validation was used within the training fold to tune hyperparameters. The final test set, reserved at the beginning using a 70–30 train-test split, remained untouched during training and validation, ensuring fair model comparison.

The final performance is reported on the held-out test set in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) [96], Area Under the Precision-Recall Curve (AUPRC) [193], and balanced accuracy with confidence intervals computed using bootstrapping with 1,000 iterations. Although there are a variety of metrics that can be reported for predictive models (e.g., precision, recall, specificity) [173, 280], our choice of AUROC and AUPRC was driven by their ability to summarize the trade-off between commonly reported metrics at various thresholds. For example, the AUROC metric quantifies the trade-off between specificity and sensitivity at various thresholds [290], while the AUPRC summarizes the trade-off between precision and recall at various thresholds [193]. We also choose to report balanced accuracy along with AUROC and AUPRC as they are more robust and indicative of the performance in the presence of imbalanced labels such as our dataset and outcome of interest when compared to normal metrics such as accuracy. Reporting metrics such as AUROC and AUPRC is a common practice in machine learning

models [156, 98], which can make it easier to interpret the findings and reduce the overoptimistic results of a single metric alone.

Predictive modeling and data preparation was performed using Python (version 3.7) and predictive models were trained using the `scikit-learn` package implementation. An overview of the predictive modeling and evaluation of our approach is presented in Figure 4.1.

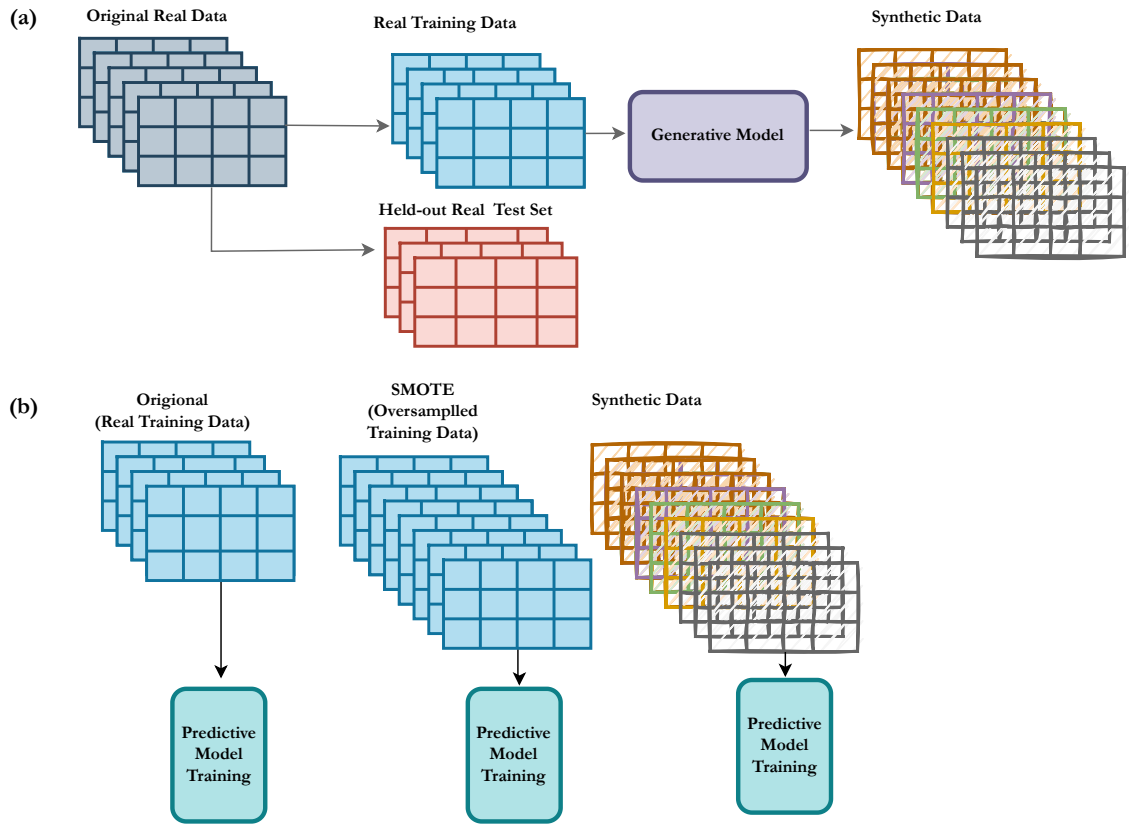


Figure 4.1: Overview of the proposed model trained on the synthetic data. (a) The dataset is split into training and a held-out test set. The training set is used to train the deep generative model, which generates synthetic data. (b) a predictive model is trained in three different setups, (1) original, (2) SMOTE, and (3) synthetic data, which are evaluated on the held-out test set and compared in terms of the performance metrics

4.3.4 Interpretability Analysis

In addition to reporting model performance, we evaluated the impact of using the various training sets on the model by conducting a feature importance and interpretability analysis using post hoc SHapley additive explanations (SHAP) [162]. We

use SHAP as the method to perform the interpretability analysis because of its relative simplicity to interpret the values, computational efficiency, and compatibility with a wide range of models. SHAP values are derived from a game-theoretic basis, where the goal is to explain the ML model's predictions for each instance by calculating the contribution of each of the features to the prediction. Although SHAP values are computed for each of the instances separately, they are often reported in their aggregated format for all the features across all samples. In this analysis, we report the mean absolute SHAP value for all features, which indicates the relative importance of the feature in terms of the impact on the prediction in the test set. Specifically, we perform SHAP analysis for the models trained using (1) original, (2) SMOTE, and (3) synthetic data, using a random forest classifier, and compare the relative importance of features. We perform the analysis using the SHAP open-source package, particularly the SHAP tree explainer [161], which works for tree ensemble models such as the random forest used for the predictive analysis in this work.

4.4 Results

4.4.1 Predictive Modeling Task

The original data used to develop our predictive modeling is made up of 364 unique patients with a positive outcome prevalence of 23.6%. The population was 66. 48% females with 39. 01% patients between 45-60 years of age. We describe the statistical distribution of our dataset in terms of outcomes and include features in section 3.6. In Table 4.1, we present the results of the models trained on the original training data (70% of the original data), the oversampled data, and synthetic data of various sizes, respectively.

In general, the models trained on synthetic samples of a size greater than 500 consistently outperformed the model trained using the original data, as well as the model trained on the oversampled data by SMOTE across the three classifier types. For the random forest model, the original model achieved a performance of 0.528 in AUROC, compared to 0.577 for the SMOTE baseline. However, the models trained on synthetic data outperformed the other baselines, with a performance of 0.610 0.344, and 0.596 for AUROC, AUPRC, and balanced accuracy, respectively. The models trained on the original data and SMOTE were first outperformed by the model trained with 1000 synthetic samples in terms of AUROC and AUPRC, where it also achieved the highest balanced accuracy of 0.592. We notice that performance gains after increasing

the synthetic data size from 1,000 to 10,000 are minimal, where the balanced accuracy did not change, with minor changes observed in the AUROC and AUPRC scores. Although there was a slight drop in the model trained on 10,000 in terms of AUROC and AUPRC, it maintained the same balanced accuracy and higher performance than the SMOTE processed and original models. Performance gains using synthetic data for the random forest model were 0.082 in AUROC, 0.088 in AURPC, and 0.107 in Balanced Accuracy. The results are also visualized in Figure 4.1.

We also report the performance using SVMs and KNN models, where the models trained on the synthetic data outperformed SMOTE and the original models. For SVM models, we note that SMOTE achieved a similar performance to the model trained on 1000 synthetic samples, in terms of AUROC but it outperformed in terms of AUPRC and balanced accuracy, respectively. The models trained on synthetic data first outperformed the original model using 200 synthetic samples, where the performance increased from 0.560 to 0.565 for the original model compared to the model trained using synthetic data. However, when using a KNN classifier, the performance of the models trained on SMOTE and the original data did not change across the three reported metrics, with an AUROC of 0.526, AUPRC of 0.255 and balanced accuracy of 0.500, respectively. We observe consistent performance gains for the model trained on 10000 synthetic samples with a performance of 0.564 for AUROC, 0.272 for AURPC, and 0.569 for balanced accuracy, respectively.

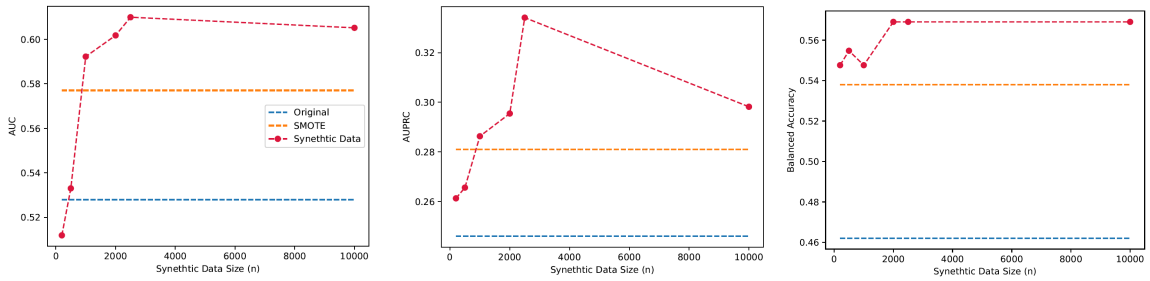
4.4.2 Interpretability Analysis

The post-hoc SHAP interpretability analysis for the random forest models revealed the relative importance of the features in making predictions for each of the baseline models trained with various training sets, as shown in Figure 4.3. We chose random forest models for this analysis, as they achieved the highest score among the three evaluated classifiers, and we provide the results of the SHAP analysis of the two other classifiers in the supplementary material Figures A2 and A3, respectively. The most predictive features in the random forest model trained on the original data were a patient age > 60 years, female sex, and a diagnosis of admission of Tetanus, which was in the top five for the models trained on synthetic datasets, with the highest predictive feature being patient age > 60 years. The model trained using oversampled data using SMOTE had a different order where patient age > 60 years ranked as the fifth most predictive feature after four features indicating admission at diagnosis. The original model's top five predictive features were patient age > 60 years, female sex, admission diagnosis of tetanus, and admission diagnosis of sepsis

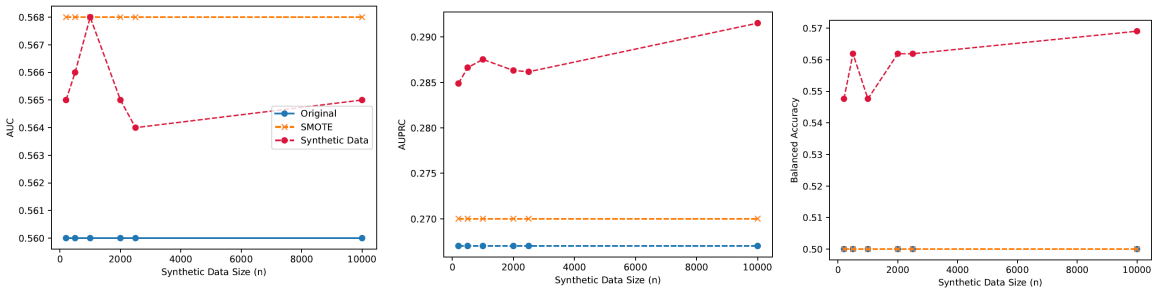
Table 4.1: Results of the predictive model using the various baselines for training data. The results are reported in terms of AUROC, AUPRC and balanced accuracy at a threshold of 0.5.

Estimator	Model	AUROC	AURPC	Balanced Accuracy
Random Forest	Original	0.528 (0.386, 0.649)	0.246 (0.157, 0.377)	0.462 (0.389, 0.542)
	SMOTE	0.577 (0.428, 0.713)	0.281 (0.169, 0.451)	0.538 (0.419,0.651)
	Synthetic 200	0.511 (0.370,0.658)	0.261 (0.153, 0.431)	0.548 (0.448,0.648)
	Synthetic 500	0.533 (0.397,0.677)	0.266 (0.162, 0.440)	0.555 (0.459,0.657)
	Synthetic 1000	0.592 (0.455,0.723)	0.286 (0.185, 0.462)	0.548 (0.450, 0.661)
	Synthetic 2000	0.602 (0.459, 0.743)	0.295 (0.182, 0.469)	0.569 (0.471, 0.675)
	Synthetic 2500	**0.610 (0.460, 0.751)**	**0.334 (0.185, 0.542)**	0.569 (0.470, 0.669)
	Synthetic 10000	0.605 (0.479, 0.742)	0.298 (0.191, 0.481)	**0.569 (0.477, 0.674)**
Support Vector Machines	Original	0.560 (0.418, 0.699)	0.267 (0.165, 0.434)	0.500 (0.500, 0.500)
	SMOTE	**0.568 (0.428, 0.707)**	0.270 (0.170 , 0.419)	0.500 (0.500, 0.500)
	Synthetic 200	0.565 (0.427, 0.703)	0.285 (0.181, 0.454)	0.548 (0.452, 0.662)
	Synthetic 500	0.566 (0.427, 0.707)	0.287 (0.176, 0.459)	0.562 (0.470, 0.672)
	Synthetic 1000	**0.568 (0.436, 0.712)**	0.288 (0.185, 0.470)	0.548 (0.450, 0.659)
	Synthetic 2000	0.565 (0.431, 0.707)	0.286 (0.177, 0.457)	0.562 (0.470, 0.660)
	Synthetic 2500	0.564 (0.427, 0.690)	0.286 (0.178, 0.449)	0.562 (0.465, 0.671)
	Synthetic 10000	0.565 (0.409, 0.708)	**0.292 (0.178, 0.460)**	**0.569 (0.476, 0.674)**
K-Nearest Neighbor	Original	0.526 (0.390, 0.666)	0.255 (0.154, 0.401)	0.500 (0.500, 0.500)
	SMOTE	0.526 (0.391, 0.657)	0.255 (0.157, 0.405)	0.500 (0.500, 0.500)
	Synthetic 200	0.528 (0.391, 0.675)	0.280 (0.167, 0.448)	0.548 (0.451, 0.650)
	Synthetic 500	0.520 (0.368, 0.669)	0.281 (0.168, 0.444)	0.555 (0.455, 0.662)
	Synthetic 1000	0.525 (0.386, 0.669)	0.281 (0.164, 0.445)	0.555 (0.465, 0.660)
	Synthetic 2000	0.542 (0.405, 0.687)	**0.290 (0.178, 0.457)**	0.555 (0.464, 0.669)
	Synthetic 2500	0.536 (0.394, 0.676)	0.281 (0.173, 0.437)	0.569 (0.469, 0.666)
	Synthetic 10000	**0.546 (0.404, 0.689)**	0.272 (0.171, 0.441)	**0.569 (0.476, 0.675)**

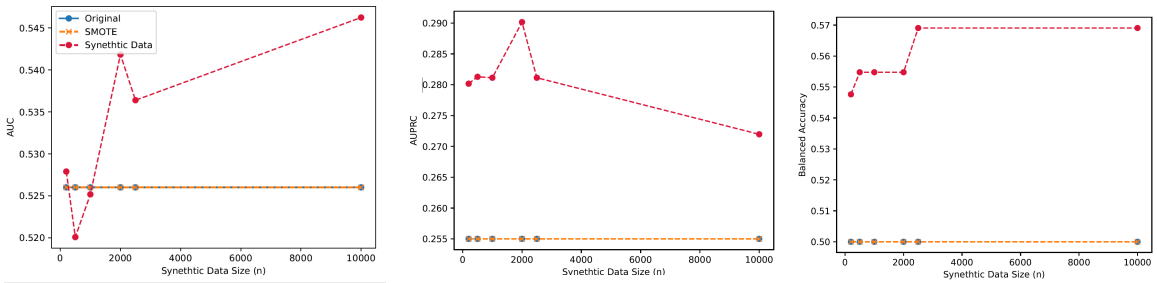
and chronic liver disease. However, the most predictive features for the model trained using oversampled training data via SMOTE were: admission diagnosis of sepsis, admission diagnosis of local infections, admission diagnosis of tetanus, admission diagnosis of dengue and patient age $>$ 60 years. We noticed that the synthetic model of 1000 patients had a different order of predictive features, where four out of five features were related to sex or age and one indicated an admission diagnosis of tetanus. Similarly, the SHAP analysis of the highest performing model, trained on 10000 synthetic samples, shows the patient age $>$ 60 years as the most predictive feature followed by the admission diagnosis of tetanus, the age of the patient 45-60 years, the age of 16-45 years and the sex of the female. The top predictive features for the models trained on synthetic samples were very similar with minor differences in the mean absolute SHAP value, which is also reflected in the similar performance in the predictive modeling tasks.



(a) Random Forest



(b) SVM

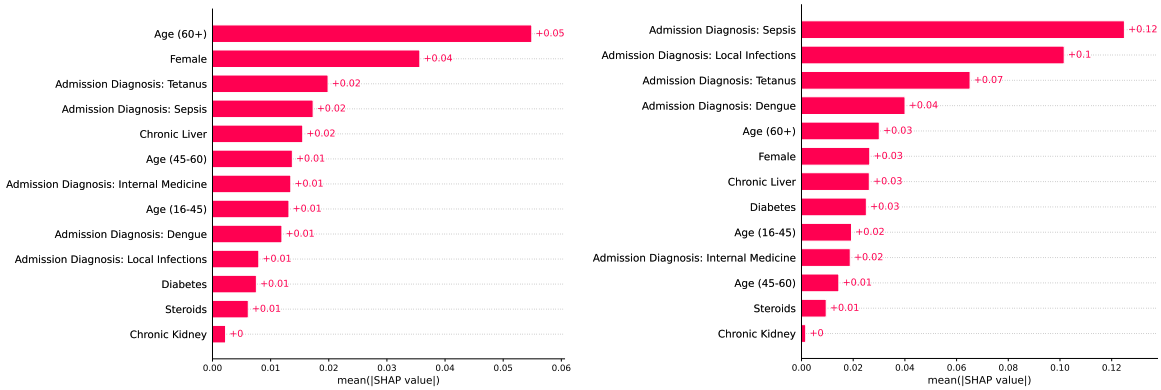


(c) KNN

Figure 4.2: Predictive performance across three classifier types (Random Forest, SVM, KNN) trained on different data variants: original data, SMOTE-augmented data, and synthetic datasets of increasing sizes (from 200 to 10,000 samples). Evaluation metrics include AUROC, AUPRC, and balanced accuracy.

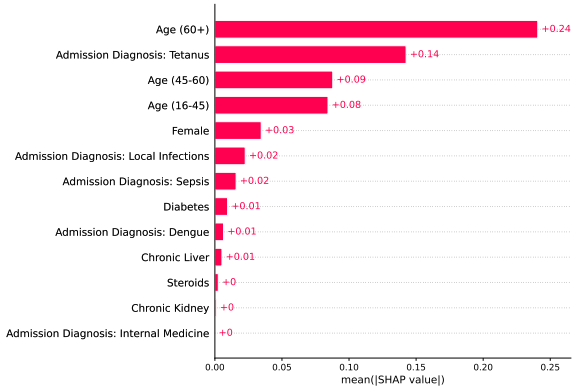
4.5 Discussion

Despite the increased research interest in using deep generative models, there is a gap in identifying the opportunities and limitations such models have in ML applications for low-source settings. To the best of our knowledge, this work is the first to investigate the use of deep generative models for generating EHRs from LMICs, where the datasets often come with small sizes and feature sets. Furthermore, our work validates the use of these synthetic data for real-world CDSS applications of high importance in LMICS, namely predicting HAI. Predicting HAI presents a challenge for

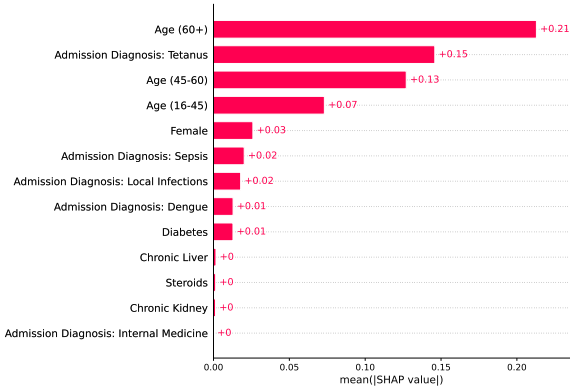


(a) Original Model

(b) SMOTE



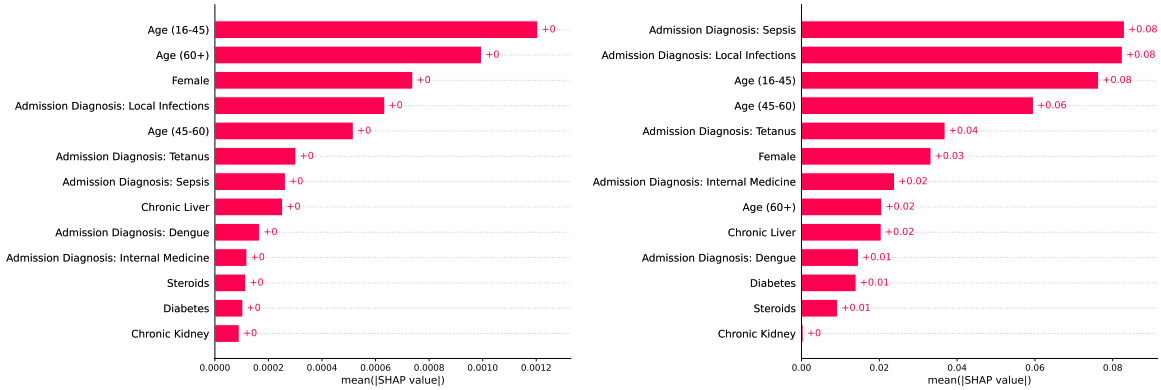
(c) Synthetic 1000



(d) Synthetic 10000

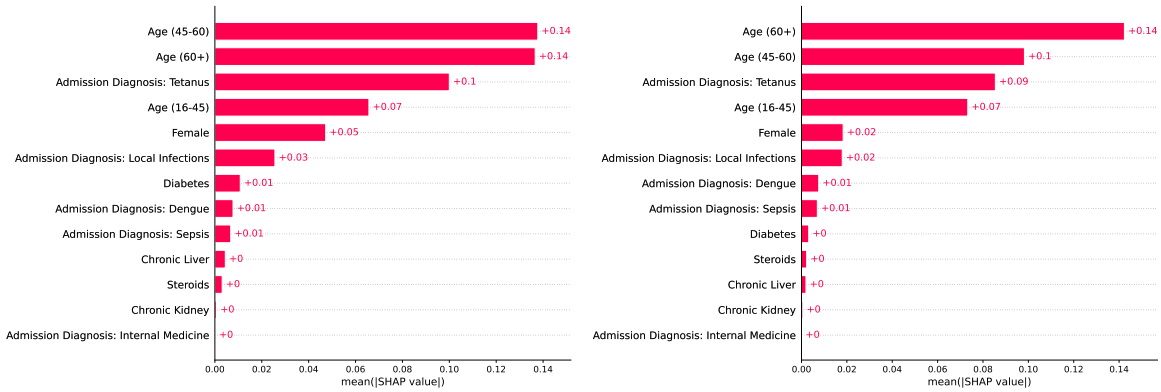
Figure 4.3: Mean absolute SHAP values computed for models trained on different datasets: original, SMOTE, and various sizes of synthetic data—using a Random Forest classifier. The figure highlights how feature importance rankings shift with the choice of training data.

clinicians, as very limited data is collected from such settings. However, improving the prevention and treatment of HAI using CDSS is a priority. Antibiotic resistance presents a global health challenge with an estimated death toll in 2019 alone, larger in magnitude than that of major diseases such as HIV and malaria [183]. LMICs tend to be one of the highest prescribers of antibiotics [190, 189], yet they remain with limited antibiotic stewardship programs [77]. With the increased burden of HAI and its link to antimicrobial resistance, especially in LMICs, our work aims to fill a gap by developing simple CDSS to predict the probability of developing such infections despite data scarcity. In this work, synthetic data was evaluated as an independent source to test its feasibility and performance potential. However, combining real and synthetic data could provide additional benefits by augmenting training diversity while retaining signal from the original distribution. We plan to explore such hybrid



(a) Original Model

(b) SMOTE



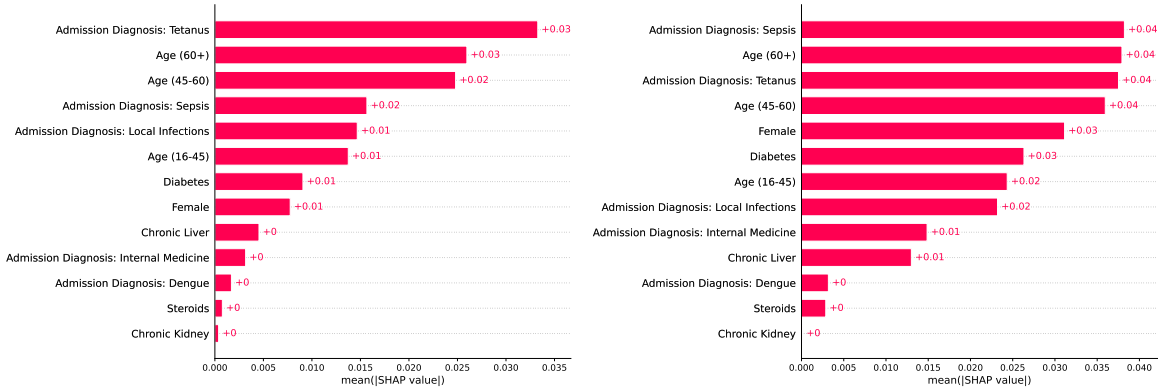
(c) Synthetic 1000

(d) Synthetic 10000

Figure 4.4: Mean absolute SHAP values from Support Vector Machine (SVM) models trained on original, SMOTE-augmented, and synthetic datasets. This comparison reveals how interpretability and feature contributions vary with data augmentation method.

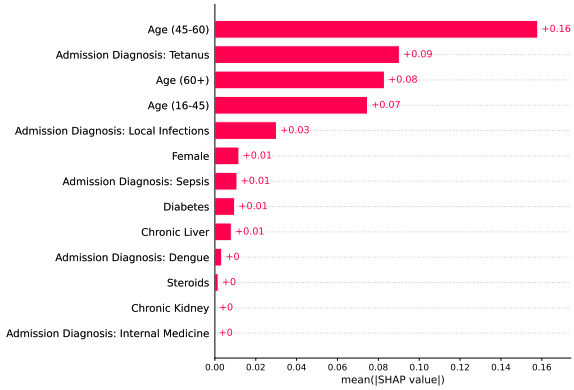
approaches in future work, which may further boost robustness in low-data regimes. The proposed approach allows improving diagnostic accuracy and performance without adding additional burden to the clinical staff involved in collecting more data, which is often not feasible. Furthermore, the interpretability component would allow for a better and more informed understanding of the risk scores predicted for each patient, toward machine learning transparency. The impact of a CDSS in providing early prediction to clinical staff would allow clinical staff to prioritize preventive strategies, optimize antimicrobial stewardship, and track care quality improvement, paving the way for better patient outcomes, with reduced operational costs related to hospital-acquired infections and patient deterioration.

Although many papers have shown the feasibility of developing CDSS, very few discuss the challenges associated with deployment and real-world validation, espe-

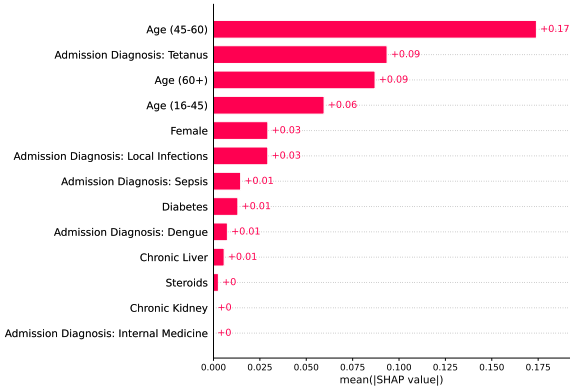


(a) Original Model

(b) SMOTE



(c) Synthetic 1000



(d) Synthetic 10000

Figure 4.5: Mean absolute SHAP values from K-Nearest Neighbor (KNN) models trained on the original dataset, SMOTE, and synthetic samples. The figure illustrates differences in feature influence across training configurations.

cially in LMICs. Examples of challenges commonly discussed include integration in clinical workflows and medical staff adaption and decision-making process [115], security [144] and interoperability [64], we believe that this work addresses an untapped area where data are scarce in terms of features and number of patients, which presents challenges for both the development and robust validation of CDSS. Another contribution of this work is demonstrating the impact of the size of the generated data on the performance of the predictive model, which we believe is an understudied area of research. We note that while many works have investigated using deep generative models and synthetic datasets [44, 68], to the best of our knowledge, this is the first to investigate the impact of the size of the synthetic dataset for EHR applications.

In addition, several related works investigated the impact of using synthetic data in downstream tasks [68, 150], this work is the first to investigate the interpretability of models trained on synthetic data compared to other baselines such as SMOTE

and the original training data. Older age and underlying medical conditions, such as diagnosis at admission, were identified as the most predictive features in both the original and models trained on synthetic datasets, which is also consistent with medical knowledge [132, 34]. The interpretability analysis showed consistency in the ranking of the five most predictive features in the models trained on the synthetic samples, which is reflected in the similar predictive performance of the models trained on the synthetic samples. Despite the comparable performance of the models trained on the original training samples and the oversampled training data using SMOTE, we noticed significant changes in the order and predictive value of the features. For example, a local infection admission diagnosis ranked as the ninth most important feature for the original model, while ranked as the second most predictive feature in the SMOTE model, compared to being the sixth and seventh most important feature for both models trained on synthetic baselines. In general, the order and SHAP values of the most predictive features of the models trained on synthetic samples did not change compared to the original model, where a diagnosis of admission of Tetanus, female sex, and an age > 60 was the most contributing to the predictions, but the model was able to achieve higher performance, indicating its ability to preserve predictive importance of features and data distribution.

In addition to evaluating predictive performance and interpretability, direct validation of the generated synthetic data is another important aspect. This can include comparing feature distributions (e.g., using JS divergence or Earth Mover's Distance), assessing pairwise correlations, and even expert clinical review. While not included in this study, incorporating such techniques would help ensure the realism and usefulness of synthetic data and will be a focus in future research.

Limitations

This work also has several limitations. The results of the predictive models were not very high, which is related to the choice of using a simpler model to simulate a close setup to the target application setting in resource-constrained settings. Future works can investigate the use of more advanced models, such as neural networks, and study the trade-off between computational complexity and impact on model performance. Another limitation of this work is related to the choice of the GAN model, where medGAN was used as one of the simpler and earlier works on GANs for EHRs. We believe that results could be improved by using conditional variants of GANs [150] where the generation can be conditioned on a specific class or outcome, or other variants with more stable training such as Wasserstein GANs and boundary-seeking

GANs [14, 67] instead of the vanilla architecture where Jensen-Shannon Divergence (JSD) can learn the distribution of the data.

It is important to note that this work and the validation performed are retrospective. Future works can investigate a prospective validation with a comparative analysis comparing the performance of models trained on the synthetic datasets, concerning the model trained on the small original data. Such an analysis would provide better insight to regulatory bodies on the approving models trained on syntactic data considering their impact on perspective deployment. Similarly, this work investigated the use of interpretability analysis as a way to study the underlying data distribution, however, future works could investigate the development of parsimonious models where a smaller set of features is used, which might result in significant reductions in the time associated with collecting the data.

Conclusions

The promising results of using synthetic data for training purposes will open the door for new research directions in building ML models for LMIC despite data scarcity, which can pave the way for new research and clinical decision support systems that best fit LMIC settings. Specifically, building tools that facilitate the development of quick models with minimal data has great potential to increase our understanding of rare and emerging diseases despite data scarcity, which in return will help improve evidence-based practice [195] without increasing the burden on clinical staff. Such efforts of synthetic data sharing will allow for bridging the gap in the CDSS for LMICs and evaluating the performance and feasibility, as well as fine-tuning the models built in developed countries in simulated settings without incurring deployment costs. Furthermore, in the absence of protection guidelines and regulations such as HIPAA [74] and GDPR [251] that are specific to low-resource settings, we believe that using deep generative models could encourage data owners in low-resource settings to share synthetic data for international research without compromising the privacy of patients in low-resource settings.

4.6 Relevant Publications

- **Ghosheh, G. O.**, Thwaites C.L., Zhu, T. (2023). Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs).

MDPI Biomedicines, 11(6), 1749.

- **Ghosheh, G.**, Li, J., Zhu, T. (2023). A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Computing Surveys*, 56(4), Article 63.

- **Ghosheh, G.**, Zhu, T. (2023). Synthesizing Electronic Health Records for Predictive Models in LMICs. *Practical ML for Developing Countries Workshop at the International Conference on Learning Representations (ICLR).*

Chapter 5

Individualized Generation of Imputations in Time-series EHRs

While Chapter 4 addressed static tabular data, EHRs often contain time-series information with irregular sampling and missingness. This chapter transitions to modeling such longitudinal data by introducing IGNITE, a model for generating individualized imputations in time-series EHRs. Here, we avoid reestablishing the motivation for handling missing data, as these issues have been covered extensively. Instead, the focus is on a tailored approach to missingness that adapts to patient-specific temporal patterns and preserves the dynamics of medical events.

5.1 Introduction

Personalized medicine is emerging as a new direction of research, in which patients are holistically examined as "individuals" rather than using symptoms or diagnoses based on the general population to determine the optimal treatment [25, 159]. Patients differ in physiology and response to treatment, where each individual has unique characteristics and patterns that reflect their genetic, molecular, and cellular makeup, as well as other environmental and behavioral factors [159]. Electronic health records (EHRs) have opened the door to the use of longitudinal data from real-world patients for a wide range of research, such as creating clinical decision support systems for different medical applications [52, 133]. The adoption of EHR has increased worldwide, including in lower-middle-income countries (LMICs) [240]. Furthermore, the wealth of data collected in EHRs presents a great source for developing approaches in medicine that tailor interventions to best suit each individual patient [3]. In addition, recent work investigated the use of EHRs to generate digital twins of patients, where

real-time monitoring, diagnosis, prognosis, and treatment optimization are applied to personalize medicine [126, 250].

Data-driven approaches, specifically machine learning models [263, 223], have shown promising results in various clinical applications, especially when using longitudinal EHR data. However, EHRs tend to be multivariate, highly missing, and irregularly sampled [166]. Therefore, the success of machine-learning models for personalized medicine is highly dependent on how the EHR data are represented conditioning on the time-varying physiology, treatments, and the missing values in the data. The existing body of work on personalized medicine mostly emphasizes the utilization of observable data [37, 257]. However, this approach tends to overlook the crucial aspect of missingness, which is essential to achieve true individualization and to obtain meaningful insights. Missingness in the healthcare domain can occur due to various reasons, such as recording errors and device or system failure, irregular sampling, and inconsistent medical visits [139]. Even high-cost and dangerous data acquisition, such as invasive or radiological procedures, may be missing due to the risk or lack of feasibility of collection for all patients [27, 135]. Furthermore, the measurement frequency of physiological data could also be related to factors such as patient severity and deterioration [7, 84], lack of medical need [65], bias and low quality of care [258, 260]. The aforementioned characteristics exhibit variability between patients, underscoring the importance of representing the individualized missingness in personalized models.

5.2 Related Works

The handling of missingness in longitudinal EHRs is a critical step before performing statistical or machine learning analysis, as missingness can reach as high as 80% or more in many EHR datasets [230, 121, 196]. Processing missingness usually involves data deletion or imputation. Simply deleting recordings or variables with missing values often leads to compromised results and is not recommended practice, especially when the data is highly missing [217, 136, 187]. Missing value imputation and replacing unobserved values with substitute values methods gained more attention with the advancements in statistical and machine-learning-based models. As discussed in Section 2.4, missing value imputation methods can be broadly grouped into three main categories: (1) simple statistical methods, (2) multiple imputation techniques, and (3) advanced machine learning-based approaches. In this work, we build upon the

third category, particularly leveraging deep learning methods tailored to longitudinal EHR data.

Recently, deep generative neural network-based models such as Generative Adversarial Networks (GAN) [92] and Variational Autoencoder (VAEs) [137] were proposed to generate synthetic data by learning the underlying distribution and dynamics of real datasets. The applications of deep generative models in healthcare range from generating synthetic patient records [44, 149], to estimating treatment effects [277], detecting un-diagnosed patients in large-scale [152], and generating missing value imputations [275, 185, 75, 164]. The generative capabilities of deep generative models make them naturally suitable to generate not only new synthetic records but also missing data imputations, where they show superior performance to most of the commonly used statistical methods in tabular and longitudinal data [134, 85, 253]. Despite their high reported performance, most deep generative methods assume *missingness is missing at random*, which limits their potential in real healthcare applications, where missingness can be informative and reflect individual underlying patient state [217]. Furthermore, many of these models can only generate *population-level* rather than *personalized-level* information conditioned on the already observed patient data, characteristics, and treatments.

In our evaluation of IGNITE, we have focused on multivariate time-series imputation within the context of electronic health records (EHRs), where continuous monitoring and recording of patient information over time is prevalent. This necessitates the inclusion of methods that are directly comparable in this domain. Although influential methods such as MD-MTS (Xu 2020) and SMILES (Zhang 2020) have been cited in the literature, these methods are primarily focused on static data imputation or specific to non-multivariate time-series data. Our work specifically targets the unique challenges of multivariate time-series imputation in EHRs. A comparison of previous works is shown in Table 2.3.

The utilization of binary missingness masks that indicate the presence or absence of observations as an additional input feature to prediction models has shown improved performance compared to models built on observed data alone [225, 36]. Despite the utility and predictiveness of binary missingness masks, they are simplistic and lack indications of the patterns and frequencies of individual-level missingness. We believe that to generate missing values at a personalized level, robust models that leverage heterogeneous and multiple data types (such as continuous physiological data, discrete treatment data, and individualized missingness masks) are required. Utilizing these mixed data types would capture the non-linear dependencies across

high-dimensional longitudinal patient records and generate EHR information that is truly personalized.

To this end, we propose a novel end-to-end generative model for (**I**ndividualized **G**e**N**eration of **I**mputations in **T**ime-series **E**lectronic health records (denoted as IGNITE). IGNITE generates a complete patient record taking into account differences in physiological data, treatment strategies, as well as individualized missingness patterns. Adopting a personalized approach to missingness generation, IGNITE could pave the way for creating digital twins for precision medicine, allowing real-time health monitoring, risk phenotyping and prediction, and treatment strategies at an individual level. We view that deep generative models present a great opportunity not only for imputing values missing at random but also for generating new features that were never measured for the patient. For this purpose, we bring new insights into missingness as a generative task for generating samples and even complete features for patients without any observations of the respective features, which we refer to as feature-wise missingness. By combining these elements, IGNITE not only addresses the persistent challenges of incompleteness and variability of data in EHRs, but also provides more accurate and personalized risk assessment and treatment strategies, which significantly contribute to improving patient outcomes and optimizing clinical workflows. An overview of our model is shown in Figure 5.1. Our main contributions are as follows:

1. **Individualized Time-series Imputation Model**: We propose a generative model comprising dual-variational autoencoders that can impute any personalized missing values in time-series EHR conditioned on discrete treatment (intervention) data and patient characteristics. Our model also includes dual-stage attention networks to improve representation learning across both the feature dimension and the temporal dimension.

2. **Personalized Missingness Mask**: We introduce a novel individualized missingness mask (IMM) that accounts for individual-level differences in missingness frequencies and patterns across time and feature dimensions. This mask is used to augment the input data to one of the dual-VAEs to better generate personalized imputations for missing EHR values.

3. **New Missingness Evaluation Framework.** We propose a framework for investigating the performance of the imputation models across sub-populations with various frequencies of sample and feature missingness. Our evaluation

framework aims to demonstrate the robustness of these models to various missingness patterns, especially when complete features were never observed.

4. **Benchmarking on real-world data.** We perform rigorous evaluations of models on three widely used publicly available Intensive Care Unit (ICU) datasets.

5.3 Methodology

We denote a multivariate time-series EHR dataset as $\mathbf{D} = \{(\mathbf{x}_{i,1:T_i})\}_{i=1}^{N}$, which includes a set of individual patient records indexed by $i \in \{1, 2, \dots N\}$. Each patient record comprises two main components: time-series and static components. The time-series component includes a continuous time-series of physiological data and a discrete time-series of treatment data, recorded over time steps $\mathbf{T} = (\mathbf{t}_1, \dots, t_N)$. Each record has a corresponding static outcome label $\mathbf{y} = (\mathbf{y}_1, \dots, y_N)$ and static demographic information, indicating the age and sex of the patient. For each individual patient time-series matrix, we extract a binary mask indicating the missingness of each value, where 0 and 1 denote the missing value and the observed value, respectively.

5.3.1 Proposed Model

Dual Variational-AutoEncoders with Dual-Stage Attention

We utilize a pair of VAEs, each with an encoder and a decoder, to map the multivariate time-series into reversible low-dimensional dense representations with no missingness. We use long- and short-term memory (LSTM) [111] neural networks for the architecture of VAE. The first VAE is trained to reconstruct the multivariate time-series by calculating the evidence lower bound (ELBO) on the observed values only and imputing the missing values with zero, an approach adopted by [185, 75]. The second VAE, on the other hand, generates the data based on the ELBO calculated on the full data that is first imputed with the last value carried forward and masked with the *Individualized Missingness Mask* (IMM) that is discussed in section 5.3.1. We refer to the first and second VAE as $VAE^{\mathcal{OO}}$and $VAE^{\mathcal{IMM}}$, respectively. Each LSTM encoder is complemented with input attention highlighting feature importance. Specifically, the attention component computes the weights for each feature conditioned on the encoder's hidden state of the previous time step. The computed weights are obtained using a deterministic fully connected network, by referring to the previous hidden state $\mathbf{h}_{t-1}$ and the cell state $\mathbf{s}_{t-1}$ in the encoder LSTM unit with:

$$e_t^k = \mathbf{v}_e^\top \tanh\left(\mathbf{W}_e\left[\mathbf{h}_{t-1}; \mathbf{s}_{t-1}\right] + \mathbf{U}_e \mathbf{x}^k\right)$$

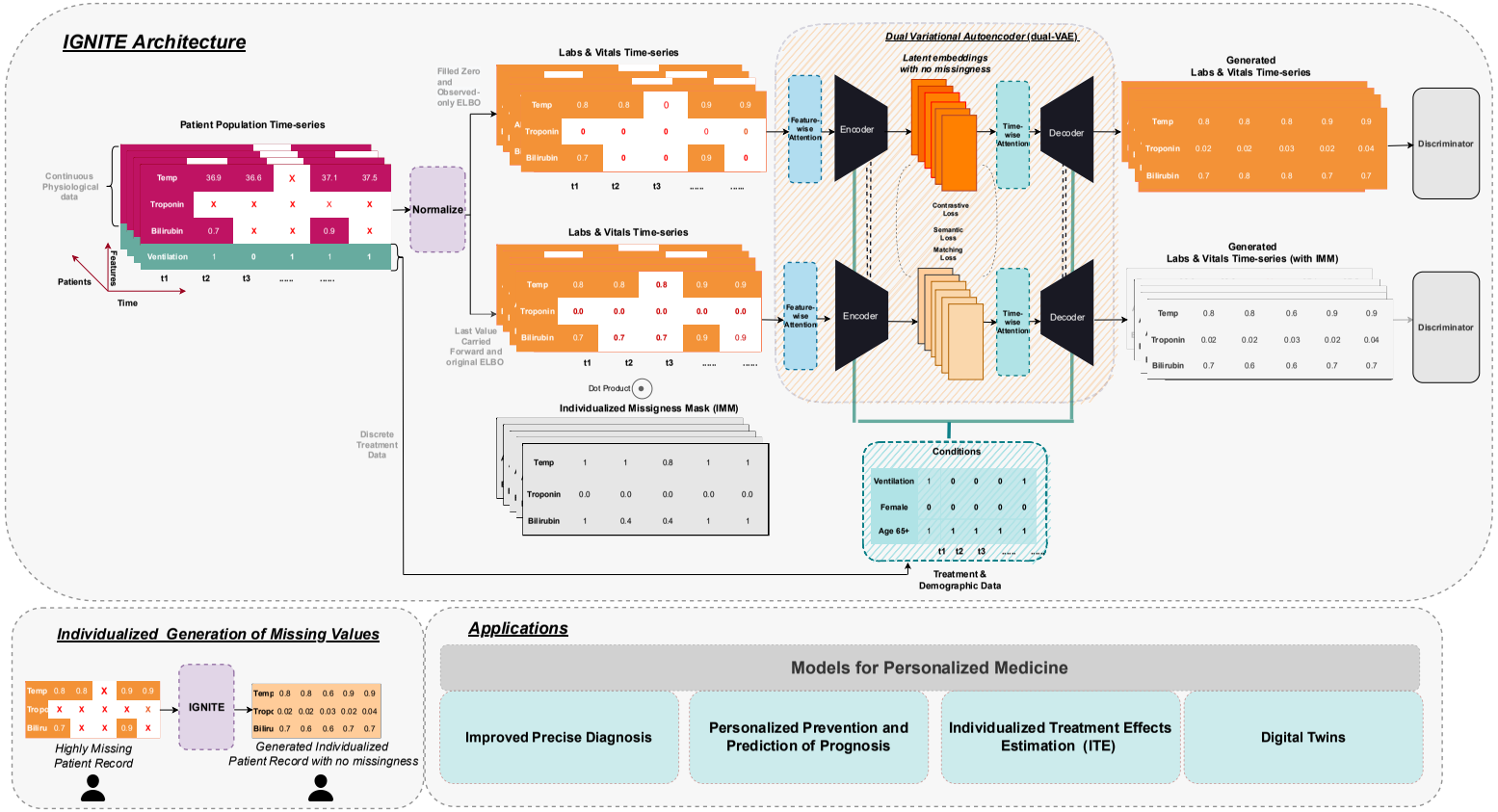


Figure 5.1: An overview of the architecture and applications of our proposed model, IGNITE, for generating individualized time-series EHRs. The evidence lower bound (ELBO) is calculated for the observed value only for the upper VAE and for the augmented data from the full individualized missingness mask (IMM) in the lower VAE. By utilizing treatment data and individualized missingness patterns, IGNITE is capable of generating EHRs that facilitate various applications of personalized medicine.

and

$$\alpha_t^k = \frac{\exp\left(e_t^k\right)}{\sum_{i=1}^{n} \exp\left(e_t^i\right)},$$

where α_t^k is the attention weight of feature k at time t. The learnable parameters are denoted in the form of $\mathbf{v}_e \in \mathbb{R}^T, \mathbf{W}_e \in \mathbb{R}^{T \times 2m}$ and $\mathbf{U}_e \in \mathbb{R}^{T \times T}$. The learned attention weights are passed to a softmax function, followed by a dense network trained with the encoder parameters. To this end, the learned weights are multiplied by the input features as shown in

$$\tilde{\mathbf{x}}_t = \left(\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \cdots, \alpha_t^n x_t^n\right)^\top.$$

which can be used to update the hidden state at time t:

$$\mathbf{h}_t = f_1\left(\mathbf{h}_{t-1}, \tilde{\mathbf{x}}_t\right),$$

The LSTM encoder unit is fed with $\tilde{\mathbf{x}}_t$ instead of $\mathbf{x}_t$, giving more attention to important features to encode a better representation of the time-series input. After encoding the data with feature-wise attention, a temporal attention mechanism is used in the decoder to select relevant encoder hidden states across all time steps. To do so, an attention weight is calculated for each encoder hidden state at time t. The calculation of the attention weight is based on the previous hidden state at time $t-1$. Specifically, the attention weight of each encoder hidden state at time t is calculated based on the previous hidden state of the decoder $\mathbf{d}_{t-1} \in \mathbb{R}^p$ and the cell state of the LSTM unit $\mathbf{s}'_{t-1} \in \mathbb{R}^p$ with

$$l_t^i = \mathbf{v}_d^\top \tanh\left(\mathbf{W}_d\left[\mathbf{d}_{t-1}; \mathbf{s}'_{t-1}\right] + \mathbf{U}_d \mathbf{h}_i\right), \quad 1 \leq i \leq T$$

and

$$\beta_t^i = \frac{\exp\left(l_t^i\right)}{\sum_{j=1}^{T} \exp\left(l_t^j\right)},$$

where $\left[\mathbf{d}_{t-1}; \mathbf{s}'_{t-1}\right] \in \mathbb{R}^{2p}$ is a concatenation of the previous hidden state and the cell state of the LSTM unit. $\mathbf{v}_d \in \mathbb{R}^m$, $\mathbf{W}_d \in \mathbb{R}^{m \times 2p}$ and $\mathbf{U}_d \in \mathbb{R}^{m \times m}$ are parameters to learn. The attention weight β_t^i represents the importance of the i-th encoder hidden state at time t. Since each encoder hidden state $\mathbf{h}_i$ is mapped to a temporal component of the input, the attention mechanism computes the context vector $\mathbf{c}_t$ as a weighted sum of all the encoder hidden states $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T\}$,

$$\mathbf{c}_t = \sum_{i=1}^{T} \beta_t^i \mathbf{h}_i.$$

We note that the context vector $\mathbf{c}_t$ is distinct at each time step. Once we get the weighted summed context vectors. The decoded data are based on the stacked weighted content vectors, allowing for better representation over time.

Individualized Missingness Mask (IMM)

Although multiple works investigated the importance of missingness indicators and their potential impact on prediction tasks [225, 36, 246], these works did not account for individual patient-record observation frequencies. Specifically, a time-varying physiological feature that is never measured for a patient can inform the model about the patient's state and medical needs. Moreover, the frequency of measurements can be related to the patient's physiological state and possible deterioration. For example, multiple measurements for cardiac troponins, sensitive biomarkers of myocardial injury, can indicate that doctors suspect a cardiac-related diagnosis or complication [59]. In the same intuition, missing measurements for the same cardiac troponins can indicate that the patient is not suspected of developing such complications, as the clinical guidelines only recommend this measurement for diagnosing myocardial injury [49]. However, this kind of feature-level missingness can not be represented in traditional binary missingness masks. For this purpose, we design a mask inspired by the frequency metrics of natural language document processing, such as the term frequency-inverse document frequency (TFIDF) [200], where a metric is calculated based on the relative frequency of appearance of words in a record. We refer to our mask as the Individualized Missingness Mask (IMM). Specifically, for each time-series feature in an individual patient record, we assign a value of 1 if the sample was observed. We formally define the mask for each time-series feature in a patient record:

$$\mathbf{IMM} = \begin{cases} 1 \text{ if } m_t^f = 1 \\ \dfrac{\sum_t^T m_i^f = 1}{T} \text{if } m_t^f = 0 \end{cases} \tag{5.1}$$

where f is a single time-series feature of length T, and m_t^f, corresponds to the binary mask indicating the presence/missingness of a single sample in feature f at time t for an individual patient record.

Conditional Generation

To incorporate discrete treatment data and demographic information from the patient, we use a conditional VAE architecture that allows for targeted data generation

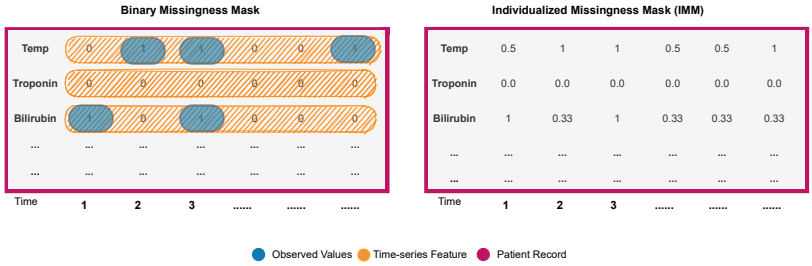


Figure 5.2: An example showcasing the difference between a binary missingness mask and an individualized missingness mask (IMM)

[232], making the imputation more specific to treatments and the age group and sex of the patient. Specifically, we utilize the 3D matrix of time-series treatments and concatenate it with the one-hot encoding of two demographic features (age and sex), which are used as conditions for our generative model. We note that the treatment and demographic features are fully observed, which makes them ideal for conditioning the VAEs.

Latent-Space Losses

While the vanilla VAE losses are based on ELBO losses, the sum of the Kullback-Leibler divergence and reconstruction loss in the observed space, we define multiple loss constraints for the optimization of the generative imputation model in latent space, namely: matching, semantic and contrastive losses.

Matching Loss. By mapping the time-series into a shared low-dimensional space, a shared weight constraint, and assuming that both data instances are representations of the same patient, we expect that the data generated by $VAE^{\mathcal{OO}}$ and $VAE^{\mathcal{IMM}}$ should be close to the latent space. Therefore, we optimize the Euclidean distance between the representations as follows:

$$\mathcal{L}^{\text{Match}} = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}\left[\left\|\mathbf{z}^{\mathcal{OO}} - \mathbf{z}^{\mathcal{IMM}}\right\|^2\right] \tag{5.2}$$

where $\mathbf{z}^{\mathcal{OO}}$ represent and $\mathbf{z}^{\mathcal{IMM}}$ represent the latent space mappings produced by each of the VAEs.

Semantic loss. Semantic loss is used to ensure that the learned representations are useful for performing downstream tasks. For this purpose, we concatenate the learned representations from both VAEs and implement a classifier to predict an outcome of interest. The loss is calculated based on the cross-entropy:

$$\mathcal{L}^{\text{Semnatic}} = \mathbb{E}_{\mathbf{z}=\text{CE}(f_{\text{classifier}}(\mathbf{z}),\mathbf{y})} \tag{5.3}$$

where z represents the representations, and y is the predicted outcome. CE stands for the standard cross-entropy loss used for classification models [170].

Contrastive loss. We also measure how close each pair of representations learned by both VAEs to the same patient, compared to those of other patients using contrastive loss [47]. Given a dataset of N patient records, the objective of contrastive loss is to have latent vectors derived from $\mathbf{VAE}^{\mathcal{OO}}$ and $\mathbf{VAE}^{\mathcal{IMM}}$ for the same patient to be similar to each other, yet different from those derived from different patients. Therefore, the equation is based on minimizing the distance between the latent space vectors $\mathbf{z}^{\mathcal{OO}}$ and $\mathbf{z}^{\mathcal{IMM}}$ and when they correspond to the same patient and maximize the distance when the latent vectors are for other patients. For this, we use the same formulation derived by [149]:

$$\mathcal{L}_{i,j}^{\text{Contra}} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i^{\mathcal{OO}}, \mathbf{z}_j^{\mathcal{IMM}}\right)/\tau\right)}{\sum_{i^{k=1}}^{2N}\left[k \neq i\right]\exp\left(\text{sim}\left(\mathbf{z}_i^{\mathcal{OO}}, \mathbf{z}_k^{\mathcal{IMM}}\right)/\tau\right)} \tag{5.4}$$

where the similarity function in the numerator is the cosine distance between the latent-space vectors, while the denominator is the cosine similarity between a vector belonging to one patient, and those belonging to all others. Here, τ is a temperature parameter that controls the concentration level of the distribution; smaller values of τ increase the separation between similar and dissimilar pairs in the latent space. The final contrastive loss is calculated for all patients included in the dataset and summed for both.

Masked Imputation Task (MIT) Loss

In addition to the reconstruction loss used in VAEs, we introduce a new loss that forces the model to minimize the loss of artificially introduced missing values, which we refer to as the Masked Imputation Task (MIT) loss as suggested in the work of [63]. The reconstruction loss used in calculating the ELBO in the VAE component is calculated only for the observed values, since the ground truth of those values cannot be accessed. The MIT loss masks a percentage of the observed values and introduces them to the model as missing values, simulating a missingness setup to force the model to generate accurate imputations. The MIT loss is calculated based on the mean squared error of the ground truth and the predicted values for the masked values.

Discriminator

The dual-VAE architecture is complemented by an LSTM-based classifier that outputs the probability that each sample belongs to the real or imputed (generated) classes. Inspired by the vanilla GAN's discriminator [92], we refer to this classifier as the discriminator. As IGNITE generates high-quality data samples, the aim is to have the discriminator not be able to tell which ones are real and which ones are the generated imputations, implying that the imputations are as real as possible. The error produced by the discriminator is optimized along with the overall model's losses.

In summary, IGNITE is optimized jointly using the total loss function

$$\mathcal{L}_{OO} = \gamma\mathcal{L}^{\text{Reconst.}} + \delta\mathcal{L}^{\text{KL}} + \epsilon\mathcal{L}^{\text{Match.}} + \zeta\mathcal{L}^{\text{Semantic}} + \eta\mathcal{L}^{\text{Contra.}} + \theta\mathcal{L}^{\text{MIT}} + \mathcal{L}^{\text{Discriminator}} \quad (5.5)$$

where $\gamma, \delta, \epsilon, \zeta, \theta,$ and ι are scalars used to determine the ratio of each loss in the overall model's loss, which are all finetuned on each dataset.

The optimization of IGNITE relies critically on the precise selection of hyperparameters, which was achieved by using a Bayesian hyperparameter search while minimizing the reconstruction loss. For each dataset, a customized Bayesian search was performed to accommodate its unique characteristics, such as variability in missing data patterns and feature scale. The final selection of hyperparameters was determined based on their ability to maximize imputation effectiveness and prediction accuracy, as evidenced by extensive validation experiments. This targeted approach ensured that the IGNITE framework was optimally tuned to each dataset's specific challenges, enhancing both imputation quality and clinical utility.

The optimization of IGNITE relies critically on the precise selection of hyperparameters. To this end, we employed a Bayesian hyperparameter search strategy, minimizing the reconstruction loss as the objective function. For each dataset, a tailored search space was defined to account for its unique characteristics, including the scale of features, temporal resolution, and the structure and extent of missingness. The hyperparameter configurations were selected based on their ability to improve both imputation quality and downstream predictive performance, as validated across AUROC and AUPRC.

Although both VAEs generate reconstructions, we are interested in the one produced by $VAE^{\mathcal{OO}}$, as it maps the data to the observed space and incorporates information from the IMM through joint training with $VAE^{\mathcal{IMM}}$. The full model is described in the pseudocode below.

Algorithm 2 Training Procedure of IGNITE

Input: Time-series input $\mathbf{X}$, binary mask $\mathbf{M}$, individualized missingness mask $\mathbf{IMM}$, treatments $\mathbf{T}$, demographics $\mathbf{d}$

Output: Trained IGNITE model

Initialize: $VAE^{\mathcal{OO}}$, $VAE^{\mathcal{IMM}}$, Discriminator D **Set:** Hyperparameters $\gamma, \delta, \epsilon, \zeta, \eta, \theta, \iota$

while *not converged* **do**

`// Step 1: Encode input using dual VAEs with attention`

$\mathbf{z}^{\mathcal{OO}} \leftarrow \text{Encoder}_{\mathcal{OO}}(\mathbf{X})$ $\mathbf{z}^{\mathcal{IMM}} \leftarrow \text{Encoder}_{\mathcal{IMM}}(\mathbf{X} \odot \mathbf{IMM})$

`// Step 2: Decode with conditional inputs`

$\hat{\mathbf{X}}^{\mathcal{OO}} \leftarrow \text{Decoder}_{\mathcal{OO}}(\mathbf{z}^{\mathcal{OO}}, \mathbf{T}, \mathbf{d})$ $\hat{\mathbf{X}}^{\mathcal{IMM}} \leftarrow \text{Decoder}_{\mathcal{IMM}}(\mathbf{z}^{\mathcal{IMM}}, \mathbf{T}, \mathbf{d})$

`// Step 3: Masked Imputation Task (MIT)`

Randomly mask observed values in $\mathbf{X}$ to form $\mathbf{X}^{\text{MIT}}$ $L_{\text{MIT}} \leftarrow \text{MSE}(\mathbf{X}^{\text{masked}}, \hat{\mathbf{X}}^{\text{masked}})$

`// Step 4: Update discriminator`

$L_{\text{Disc}} \leftarrow -\log D(\mathbf{X}) - \log(1 - D(\hat{\mathbf{X}}^{\mathcal{IMM}}))$ Update D to minimize L_{Disc}

`// Step 5: Compute loss terms`

$L_{\text{Reconst}} \leftarrow \text{MSE}(\mathbf{X}_{\text{obs}}, \hat{\mathbf{X}}^{\mathcal{IMM}}_{\text{obs}})$

$L_{\text{KL}} \leftarrow \text{KL}[\mathbf{z}^{\mathcal{IMM}} \| \mathcal{N}(0, I)]$

$L_{\text{Match}} \leftarrow \|\mathbf{z}^{\mathcal{IMM}} - \mathbf{z}^{\mathcal{OO}}\|^2$

$L_{\text{Semantic}} \leftarrow \text{CE}(f_{\text{cls}}([\mathbf{z}^{\mathcal{OO}}, \mathbf{z}^{\mathcal{IMM}}]), y)$

$L_{\text{Contra}} \leftarrow \text{ContrastiveLoss}(\mathbf{z}^{\mathcal{OO}}, \mathbf{z}^{\mathcal{IMM}})$

`// Step 6: Total loss`

$\mathcal{L}_{\text{IGNITE}} \leftarrow \gamma L_{\text{Reconst}} + \delta L_{\text{KL}} + \epsilon L_{\text{Match}} + \zeta L_{\text{Semantic}} + \eta L_{\text{Contra}} + \theta L_{\text{MIT}} + \iota L_{\text{Disc}}$

`// Step 7: Model update`

Update $VAE^{\mathcal{OO}}, VAE^{\mathcal{IMM}}$ and attention parameters to minimize $\mathcal{L}_{\text{IGNITE}}$

return *Trained IGNITE model*

5.3.2 Datasets Description

To evaluate the proposed model, we use three large-scale EHR datasets for intensive care unit patients from the United States and the Netherlands. The datasets include patients admitted to the ICU for medical, surgical, and trauma care, all of whom have not experienced a mortality outcome during the first 48 hours. The included datasets are as follows:

1. **PhysioNet Challenge 2012**: An open-access ICU dataset frequently used for time-series imputation, with over 80% overall missingness. It comprises 12,000 patient encounters, 35 time-series physiological features, and one treatment feature (mechanical ventilation). More details can be found in section 3.2.

2. **eICU**: A multi-center dataset from 208 US hospitals (2014–2015), containing more than 200,000 admissions and 827 million observations. Following preprocessing based on Li et al., 54,423 encounters with 55 time-series physiological features and at least 48 hours of stay in the ICU were retained. Further details can be found in section 3.3.

3. **HiRID**: A single-center critical care dataset from Bern University Hospital, Switzerland, with more than 33,000 ICU admissions with 50 time-series physiological features. Further details can be found in section 3.4.

For each of the datasets, we studied two types of missingness, namely feature-wise and sample-wise missingness. We formally define the two types of missingness below:

- **Feature-wise missingness**: This refers to the percentage of patients who have no observations for a specific feature. It captures the extent to which a particular feature is completely missing for a subset of patients. In this case, the missingness is evaluated at the level of the feature for each patient, highlighting if some features are completely absent in certain patient records.

- **Sample-wise missingness**: This refers to the overall percentage of missing values in all patient records, features, and time steps. It is calculated as the ratio of the total number of missing entries to the total number of possible entries. This metric captures how much data is missing at the level of individual data points, giving a sense of the overall missingness across the dataset.

Full details on the distribution of missingness across the three datasets are provided in3.2, 3.3 and 3.4, respectively.

5.3.3 Experimental Setting

Downstream Task

Due to the absence of ground-truth labels for highly missing EHR data, we evaluated IGNITE's imputations and compared them with those produced by the baseline methods using an extensive downstream task framework based on the original data missingness. To do so, we train an LSTM [111] classifier to predict mortality in any of the components of the framework explained below. During the development of the IGNITE model, we strictly adhered to protocols that ensure that there is no overlap or information exchange between training and test sets. The imputation model was

trained exclusively on the designated training data. This data set includes comprehensive features, but is entirely separate from the test dataset, which was set aside at the outset of the study. All imputation models were trained in the training set with a ratio of 80% of the full datasets and evaluated inference by producing imputations for an unseen test set. The final performance is reported in terms of both the area under the receiving operating curve (AUROC) [70] and the area under the precision recall curve (AUPRC) [231]. AUPRC measures the area under the precision and recall curve, making it more informative for class imbalance cases in all three datasets. In addition to the downstream task usually evaluated in imputation works, we propose an evaluation framework that tests IGNITE's ability to impute if trained and tested on data with various types and percentages of naturally occurring missingness. The proposed downstream task framework includes the following.

1. **Performance across the full population** In this analysis, we evaluate the performance of the machine learning model in a mortality classification task using the entire dataset, regardless of the missing rate. This is the common evaluation used by related works in the literature [75, 31, 164, 163].

2. **Performance across patients with overall sample-wise missingness** In this analysis, we investigate the performance of the mortality prediction classifier when trained in population subgroups where the sample-wise missingness across all features and time-steps is $\leq 25\%$, 25-75% and $\geq 75\%$, respectively. We did not report results for a subset that has fewer than 800 samples, as they are unreliable.

3. **Performance across patients with high feature-wise missingness** While modeling overall missingness is important, for this analysis, we focus on patients with a high percentage of features that were completely missing, i.e., features that were not observed for a particular patient. This analysis aims to evaluate the impact of various imputation methods when some features are never measured for population subgroups. Like the sample-wise missingness experiments, the investigated subgroups were also $\leq 25\%$, 25-75% and $\geq 75\%$, respectively. Similarly to the sample-wise missingness experiments, the results of subsets with low sample sizes were excluded.

Reconstruction Task

Despite not having the ground-truth labels for missing values in the EHR dataset, there are observed values which we can use to evaluate the accuracy of reconstruction

using the proposed model. Therefore, we randomly mask a percentage of the observed values for each patient in the test set as hidden and then observe the model's ability to recover the original values. A similar approach was previously used by [163, 75, 185]. The main purpose of this validation method is to measure the models' ability to reconstruct the original data and quantitatively compare it to other baseline methods. In our evaluation, the quality of the imputed data is assessed by comparing the reconstructions to the observed values using Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE). To account for variability in the amount of data missing from each patient's record, we introduce missingness randomly across the test set population. The RMSE for each patient is then calculated and weighted according to the proportion of missingness introduced in their records, ensuring that our metrics accurately reflect the severity and distribution of missing data. For this analysis, missingness was introduced at varying levels of 10%, 20%, and 50% to test the robustness of our imputation methods under different scenarios. The final results are presented as the mean and standard deviation of RMSE and MAE across all patients for each baseline method, providing a comprehensive view of performance across different degrees of incompleteness of the data.

Baselines

We benchmark our model with several commonly used and state-of-the-art imputation methods listed below:

- **LOCF**: Last Observation Carried Forward, an imputation method often used in clinical studies, where the last observed value for the patient is used to impute all subsequent missing observation points [224].

- **MICE**: Multivariate Imputation by Chained Equations is a simulation-based statistical model where N complete datasets are produced from the incomplete dataset by imputing the missing N times. These completed N datasets are analyzed and then combined into a single dataset [117]. This method is often used in epidemiological studies and is designed for tabular datasets.

- **CATSI**: Context-Aware Time Series Imputation is a bidirectional LSTM-based model that enhances clinical time series imputation by capturing global health contexts of patients and using cross-functional correlations [273].

- **GP-VAE**: a deep probabilistic model that utilizes a Gaussian process with a Cauchy kernel and a VAE to impute the missing multi-variate time-series data [75].
- **Transformer**: An implementation of attention is all you need paper [249], adapted for time-series imputations.
- **BRITS**: Bidirectional Recurrent Imputation for time-series, an RNN-based model where bidirectional long short-term memory network (BiLSTM) is used to predict the intermediate states of partially observed data, given its past and future states [31].
- **TimesNet**: a neural network for long-term time series forecasting that transforms 1D time series into a 2D space to unify intraperiod and interperiod variations. By leveraging multi-resolution representations, temporal convolutions, and attention mechanisms, it effectively captures complex temporal patterns, achieving state-of-the-art performance across diverse datasets [262].
- **SAITS**: Self-Attention-based Imputation for time-series, a model based on using the self-attention mechanism, where it learns to impute missing values from a weighted combination of diagonally-masked self-attention (DMSA) blocks [63].

Implementation

The training process was implemented using TensorFlow, and training was conducted on NVIDIA GeForce RTX 3090 GPUs. We use an exponential decay learning rate schedule for the Adam optimizers.

5.4 Results

5.4.1 Downstream Task

To evaluate the performance of different imputation models, we used the imputed data in a mortality prediction downstream task, on each of the three datasets: PhysioNet 2012, HiRID and eICU. In all models, we first imputed the data and then used the imputed data for evaluation in the respective downstream task. The higher the predictive performance of the model, the better the imputation of the data. In the mortality prediction task for the PhysioNet 2012 dataset, our IGNITE performed the best with an area under the receiving operating curve (AUROC) of 0.834, followed by

TimesNet and transformer AUROC of 0.831 and 0.827, respectively. Models trained on the LOCF-imputed method were ranked the worst with an AUROC of 0.777. The various baseline methods were also evaluated in terms of area under the precision-recall curve (AUPRC), which represents discrimination regarding class imbalance. The results were also consistent on the eICU and HiRID datasets, where IGNITE consistently outperformed all other benchmarks. For instance, IGNITE achieved an AUROC of 0.822 in eICU and 0.968 in HiRID, respectively. A similar trend was observed in the AUPRC results for both datasets. It is worth noting that the eICU imputations generated by IGNITE significantly improved the performance with an increase of more than 10% and 20% when compared to the next best or worst models, respectively. We also implemented Wilcoxon Signed-Ranked test to assess for the statistical significance and show the results.

Table 5.1: Performance for a mortality prediction task using an LSTM model reported in terms of AUROC and AURPC. $*$ indicates that IGNITE is statistically significant ($p < 0.05$) when compared to the corresponding benchmark.

Imputation method	**AUROC**			**AUPRC**		
	PhysioNet 2012	eICU	HiRID	PhysioNet 2012	eICU	HiRID
LOCF	0.777 (0.014)*	0.736 (0.001)*	0.935 (0.007)*	0.389 (0.025)*	0.316 (0.015)*	0.593 (0.031)*
MICE	0.823 (0.009)*	0.765 (0.014)*	0.939 (0.007)*	0.469 (0.025)*	0.289 (0.034)*	0.563 (0.023)*
CATSI	0.782 (0.011)*	0.721 (0.011)*	0.961 (0.004)*	0.405 (0.022)*	0.283 (0.016)*	0.658 (0.028)*
GP-VAE	0.780 (0.013)*	0.798 (0.009)*	0.948 (0.006)*	0.371 (0.024)*	0.397 (0.036)*	0.583 (0.025)*
Transformer	0.831 (0.010)*	0.729 (0.008)*	0.961 (0.004)*	**0.477 (0.028)***	0.249 (0.017)*	0.658 (0.022)*
BRITS	0.818 (0.010)*	0.735 (0.001)*	0.959 (0.005)*	0.452 (0.025)*	0.318 (0.015)*	0.641 (0.027)*
TimesNet	0.827 (0.001)*	0.750 (0.010)*	0.956 (0.004)*	0.352 (0.014)*	0.305 (0.014)*	0.640 (0.022)*
SAITS	0.821 (0.010)*	0.735 (0.011)*	0.961 (0.005)*	0.467 (0.035)*	0.323 (0.016)*	0.655 (0.030)*
IGNITE	**0.834 (0.009)**	**0.822 (0.008)**	**0.968 (0.004)**	0.458 (0.032)	**0.345 (0.014)**	**0.723 (0.030)**

To evaluate the population for varying sample-wise missingness, where values are not observed across time and features, we stratified the dataset into three groups $\leq$ 25%, 25%- 75%, and $\geq$ 75% sample-wise missingness. However, in all three datasets, no patients had $\leq$ 25% sample-wise missingness, indicating the high missingness in the EHR. The high prevalence of missingness in all included datasets resulted in two main subgroups for 25%- 75%, and $\geq$ 75% sample-wise missingness for PhysioNet 2012 and HiRID. On the other hand, more than 95% of the samples in the eICU dataset had more than 75% sample-wise missingness, resulting in one subgroup for eICU. In terms of performance, IGNITE consistently outperformed all baselines in terms of AUROC in all three datasets. Specifically, IGNITE achieved the highest performance of 0.816, followed by BRITS for the 25%- 75% group in the PhysioNet 2012 dataset. For AURPC, IGNITE achieved the second highest performance with 0.419, where the highest performance was achieved by GP-VAE. Similar trends were

Table 5.2: Performance for a Mortality Prediction Task for Patients with at least n% of Features Never Measured (Completely Missing)

% Feature-wise Missingness	**≤ 25%**			**25-75%**			**≥ 75%**		
	PhysioNet 2012	eICU	HiRID	PhysioNet 2012	eICU	HiRID	PhysioNet 2012	eICU	HiRID
Population Description									
Population Size (n)	4434	6213	4513	7604	45,845	2913	NA	2329	NA
Positive Outcome (n, %)	982 (22.7%)	1342 (21.6%)	574 (12.7%)	723 (9.51%)	1614 (3.52%)	157 (5.39%)	NA	131 (5.62%)	NA
AUROC									
LOCF	0.710	0.832	0.913	0.788	0.658	0.919	NA	0.587	NA
MICE	0.769	0.754	0.834	0.703	0.650	0.907	NA	0.651	NA
CATSI	0.719	0.636	0.915	0.757	0.612	0.853	NA	0.664	NA
GP-VAE	0.705	0.817	**0.957**	0.746	0.697	0.947	NA	0.634	NA
Transformer	0.764	0.887	0.924	**0.820**	0.699	**0.973**	NA	0.690	NA
BRITS	0.757	0.866	0.932	0.817	0.682	0.939	NA	0.697	NA
TimesNet	0.717	0.811	0.922	0.745	0.733	0.935	NA	0.719	NA
SAITS	0.759	0.900	0.925	0.798	**0.728**	0.921	NA	0.701	NA
IGNITE	**0.790**	**0.920**	0.952	0.791	0.705	0.961	NA	**0.751**	NA
AUPRC									
LOCF	0.386	0.688	0.724	0.297	0.123	0.565	NA	0.207	NA
MICE	0.543	0.590	0.695	0.273	0.121	0.798	NA	0.201	NA
CATSI	0.444	0.517	0.600	0.250	0.100	0.229	NA	0.210	NA
GP-VAE	0.427	0.643	0.806	0.271	0.159	0.681	NA	0.233	NA
Transformer	0.495	0.734	0.811	0.326	0.235	0.807	NA	0.273	NA
BRITS	0.501	0.724	0.847	0.328	0.219	0.828	NA	0.290	NA
TimesNet	0.454	0.680	0.841	0.336	0.295	0.834	NA	0.280	NA
SAITS	0.514	0.782	0.852	0.290	**0.282**	0.825	NA	**0.367**	NA
IGNITE	**0.561**	**0.807**	**0.870**	**0.346**	0.261	**0.863**	NA	0.321	NA

observed in the eICU and HiRID datasets, with an AUROC achieving 0.959 for the HiRID dataset for samples with ¿ 75% sample-wise missingness. Although SAITS and Transformer showed high performance across all population subsets, the performance deteriorated for the population with low sample-wise missingness. We also observed that mortality outcome was less prevalent in patients with a higher overall sample missingness, 13.4% compared to those with a lower sample missingness 25.1% in the PhysioNet 2012 dataset. Similar trends were also observed in the HiRID dataset, where the prevalence of the outcome decreased from 23.9% to 5.43% for samples with high sample-wise missingness.

5.4.2 Reconstruction Task

Other than measuring the performance of IGNITE in predictive modelling tasks, we also evaluate its ability to recover and reconstruct randomly introduced missing values (i.e., missing at completely random). In the reconstruction experiments, IGNITE consistently outperformed the other baselines across the two metrics, Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE) for each patient. The metrics shown in Table 5.4, are reported in the mean and standard error of the errors across all the patients in the test set. Specifically, IGNITE had an RMSE of 0.100 (0.038), followed by MICE with 0.109 (0.059) for the setting where 10% missingness was introduced in the PhysioNet 2012 data. Similar trends were observed in the 20% and 50% experiments, where IGNITE achieved the lowest reconstruction error and

Table 5.3: Performance for a mortality prediction task for patients with at varying sample-wise missingness. The results are reported in terms of AUROC and AUPRC, respectively.

% Sample-wise Missingness	25-75%			≥ 75%		
	PhysioNet 2012	eICU	HiRID	PhysioNet 2012	eICU	HiRID
Population Description						
Population Size (n)	825	NA	1768	11175	54,370	5709
Positive Outcome (n, %)	207 (25.1%)	NA	422 (23.9%)	1500 (13.4%)	3071 (5.65%)	310 (5.43%)
AUROC						
LOCF	0.762	NA	0.850	0.754	0.622	0.912
MICE	0.786	NA	0.815	0.798	0.604	0.791
CATSI	0.792	NA	0.848	0.743	0.654	0.850
GP-VAE	0.727	NA	0.833	0.754	0.671	0.937
Transformer	0.737	NA	0.855	0.808	0.640	0.935
BRITS	0.804	NA	0.831	0.801	0.680	0.913
TimesNet	0.761	NA	0.837	0.723	0.689	0.906
SAITS	0.770	NA	0.826	0.815	0.662	0.929
IGNITE	**0.816**	NA	**0.861**	**0.826**	**0.747**	**0.959**
AUPRC						
LOCF	0.465	NA	0.716	0.362	0.199	0.582
MICE	0.592	NA	0.649	0.379	0.218	0.629
CATSI	0.542	NA	0.680	0.344	0.147	0.219
GP-VAE	0.477	NA	0.610	0.334	0.258	0.647
Transformer	0.529	NA	0.771	**0.429**	0.244	0.831
BRITS	0.622	NA	0.785	0.421	0.269	0.810
TimesNet	0.500	NA	0.781	0.294	0.254	0.793
SAITS	0.533	NA	0.737	0.420	0.224	0.790
IGNITE	**0.626**	NA	**0.796**	0.419	**0.302**	**0.823**

LOCF achieved the highest error. It is worth noting that the error in IGNITE generally remained stable despite increasing the percentage of missingness, indicating its robustness in reconstructing data with high missingness. To further assess the impact of individualized missingness masks and conditional dependency modeling, we conducted additional experiments that evaluated imputation quality across demographic subgroups. Specifically, we separately analyzed the RMSE and MAE of the imputed values for male and female patients, as shown in Supplementary Table 5. The results indicate that IGNITE effectively reduces the variance in the imputation error between subgroups, demonstrating its robustness in preserving population-specific characteristics. Our analysis also highlights that IGNITE's conditional dependency modeling helps maintain clinically relevant imputation patterns by capturing the underlying relationships between features, even when missingness patterns differ across subgroups. This suggests that the model is effective not only in recovering missing values, but also in ensuring that imputed values align with expected distributions within each demographic group. We showcase the full results for the reconstruction experiments on HiRID and eICU datasets in Supplementary Tables 6 and 7, where

IGNITE consistently outperformed all the baselines.

In addition to its methodological contributions and high performance in reconstruction tasks, IGNITE significantly boosts prediction accuracy by improving how imputed data align with actual patient statuses — a crucial factor in effective clinical decision-making. By accurately imputing missing data in a way that reflects real-time changes in patient health, IGNITE preserves essential temporal trends and patient-specific dynamics within the EHR data. We show example patient imputations with different patient states in sample visualization of the reconstructed values via top-performing models shown in Figure 2. In Figure 2 a and Figure 2 b, we show the imputations generated by the various models for two patients with different types of missingness, i.e., with feature-wise and sample-wise missingness, where IGNITE was the model with the lower error and best at recovering masked values. In Figure 2 b, we see that IGNITE and the other deep learning models generated realistic values. However, LOCF imputation fails as there are no observed values. The other methods' reconstructions are shown in Supplementary Figure 1. Furthermore, in Figure 3, we present a comparison of imputed values using IGNITE against BRITS and MICE, both with strong performance. In this section, we showcase imputed values for three key variables, with additional results provided in Supplementary Figure 2. Our results demonstrate that IGNITE is the only model that imputes urine output within the true range and trend, as seen in a case where the patient's urine output decreases prior to death. In contrast, BRITS and MICE show an increasing trend, far from the ground truth. A similar observation is made for HCO3, where IGNITE is the only model to predict values close to the ground truth. For FiO2, typically flat around 0.5 for patients on mechanical ventilation, IGNITE accurately captures this, providing realistic imputations. These examples highlight that IGNITE not only improves predictive performance but also generates clinically realistic imputations.

5.4.3 Ablation Study

To demonstrate the added value of each of the model's components, we conducted an ablation study on the three datasets The studied components were mainly the time-series conditional component as well as the IMM mask used in IGNITE. As presented in Table 7.2, we observe consistent performance gains in terms of AUROC as new components are added across the three datasets. The highest performance gain was attributed to adding IMM to the model, where the final model achieved 0.835, 0.774 and 0,968 of AUROC for PhysioNet 2012, eICU and HiRID respectively. The performance improvements reported in Table 5.5, though moderate in absolute

Table 5.4: Performance of the various baselines of the reconstruction task on the PhysioNet 2012 dataset. The results are calculated across all the patients in the test set and reported in terms of RMSE and MAE with mean and standard deviation. $*$ indicates that IGNITE is statistically significant ($p < 0.05$) when compared to the corresponding benchmark.

	RMSE			**MAE**		
Introduced Missingness	**10%**	**20%**	**50%**	**10%**	**20%**	**50%**
All Population						
LOCF	0.183 (0.065)*	0.192 (0.051)*	0.202 (0.036)*	0.106 (0.040)*	0.108 (0.034)*	0.115 (0.028)*
MICE	0.109 (0.046)*	0.112 (0.037)	0.122 (0.034) *	0.066 (0.023)	0.066 (0.019)	0.066 (0.017)
CATSI	0.284 (0.067)*	0.305 (0.051)*	0.352 (0.038)*	0.181 (0.051)*	0.196 (0.042)*	0.247 (0.033)*
GP-VAE	0.395 (0.067)*	0.398 (0.521)*	0.398 (0.044)*	0.317 (0.056) *	0.318 (0.045)*	0.317 (0.038)*
Transformer	0.119 (0.059)*	0.128 (0.05)*	0.123 (0.039)*	0.068 (0.028)	0.07 (0.025)*	0.068 (0.021)*
BRITS	0.115 (0.054)*	0.116 (0.044)*	0.118 (0.038)*	0.066 (0.027)	0.066 (0.022)	0.065 (0.020)
TimesNet	0.266 (0.158)*	0.263 (0.045)*	0.266 (0.144)*	0.071 (0.026)*	0.069 (0.021)*	0.079* (0.027)
SAITS	0.122 (0.058)*	0.127 (0.051)*	0.123 (0.038)*	0.071 (0.029)*	0.069 (0.025)*	0.067 (0.021)
IGNITE	**0.100 (0.038)**	**0.103 (0.032)**	**0.104 (0.027)**	**0.062 (0.021)**	**0.063 (0.018)**	**0.063 (0.016)**
Females						
LOCF	0.184 (0.06)	0.190 (0.052)	0.202 (0.037)	0.160 (0.041)	0.107 (0.037)	0.114 (0.031)
MICE	0.108 (0.051)	0.114 (0.044)	0.140 (0.067)	0.065 (0.025)	0.066 (0.021)	0.067 (0.019)
CATSI	0.284 (0.067)	0.353 (0.037)	0.353 (0.038)	0.181 (0.0517)	0.196 (0.041)	0.247 (0.031)
GP-VAE	0.402 (0.060)	0.405 (0.0492)	0.407 (0.041)	0.324 (0.051)	0.325 (0.036)	0.325 (0.036)
Transformer	0.113 (0.058)	0.123 (0.055)	0.120 (0.040)	0.065 (0.028)	0.068 (0.026)	0.0655 (0.023)
BRITS	0.109 (0.054)	0.114 (0.045)	0.114 (0.039)	0.064 (0.027)	0.064 (0.023)	0.063 (0.021)
TimesNet	0.118 (0.045)	0.118 (0.035)	0.122 (0.028)	0.071 (0.026)	0.070 (0.022)	0.071 (0.021)
SAITS	0.116 (0.057)	0.122 (0.052)	0.120 (0.039)	0.067 (0.030)	0.067 (0.026)	0.065 (0.022)
IGNITE	**0.097 (0.0382)**	**0.130 (0.028**)	**0.101 (0.027**)	**0.061 (0.02)**	**0.083 (0.020)**	**0.061 (0.017)**
Males						
LOCF	0.1799 (0.0657)	0.1881 (0.0529)	0.1998 (0.0394)	0.1043 (0.0408)	0.1065 (0.0362)	0.1135 (0.0311)
MICE	0.1029 (0.0480)	0.1089 (0.0412)	0.1364 (0.0685)	0.0627 (0.0231)	0.0634 (0.0198)	0.0659 (0.0187)
CATSI	0.2843 (0.0666)	0.3049 (0.0517)	0.3517 (0.0386)	0.1807 (0.0509)	0.1965 (0.0429)	0.2468 (0.0326)
GP-VAE	0.4063 (0.0634)	0.4055 (0.0510)	0.4056 (0.0413)	0.3272 (0.0531)	0.3244 (0.0438)	0.3238 (0.0366)
Transformer	0.1122 (0.0566)	0.1215 (0.0492)	0.1176 (0.0388)	0.0638 (0.0259)	0.0663 (0.0240)	0.0645 (0.0212)
BRITS	0.1079 (0.0513)	0.1109 (0.0435)	0.1109 (0.0357)	0.0628 (0.0247)	0.0628 (0.0219)	0.0620 (0.0191)
TimesNet	0.1177 (0.0463)	0.1169 (0.0366)	0.1210 (0.0291)	0.0705 (0.0256)	0.0691 (0.0216)	0.0707 (0.0203)
SAITS	0.1150 (0.0550)	0.1196 (0.0505)	0.1186 (0.0372)	0.0673 (0.0270)	0.0649 (0.0238)	0.0641 (0.0205)
IGNITE	**0.0933 (0.0338)**	**0.1259 (0.0268)**	**0.0981 (0.0244)**	**0.0592 (0.0194)**	**0.0813 (0.0190)**	**0.0603 (0.0154)**

terms, are consistent across multiple datasets and evaluation metrics. In real-world clinical settings, even incremental gains in predictive accuracy or imputation quality can translate into better-informed decision-making and improved patient outcomes. Moreover, the improvements are robust across different data conditions, reinforcing the practical value of the proposed approach.

Table 5.5: Performance when ablating various components of the proposed models, reported in terms of AUROC

Ablations	**PhysioNet 2012**	**eICU**	**HiRID**
IGNITE (Dual VAE, No Condition)	0.818	0.746	0.927
IGNITE (Dual VAE, Condition)	0.823	0.754	0.933
IGNITE (Dual VAE, Condition, IMM)	0.830	0.762	0.952
IGNITE (Dual VAE, Condition, IMM, MIT loss)	0.832	0.765	0.965
IGNITE (Dual VAE, Condition, IMM, MIT loss, discriminator)	**0.835**	**0.774**	**0.968**

5.5 Discussion

In this work, we propose a deep end-to-end generative model that synthesizes personalized data in highly sparse and irregularly sampled or even completely missing EHR conditioned on treatments and demographic characteristics, as well as the individual level of missingness patterns. We note that our model, IGNITE, outperformed the baseline models in the full population downstream mortality prediction task in terms of AUROC and AUPRC. Unlike previous works, we proposed a new framework for the evaluation of imputation models trained on various types of missingness across features and samples in time-series data. In the proposed downstream task evaluation, IGNITE showed consistent robustness despite varying the percentage of feature-wise and sample-wise missingness. To ensure a clear interpretation of our results, we emphasize that the datasets analyzed in Table 2 and Table 3 are not subsets of each other but are designed for distinct evaluation purposes. Table 2 presents the standard train-validation-test split used to evaluate overall model performance, while Table 3 examines model robustness under different levels of missingness. As these experiments use different data partitions and evaluation criteria, direct comparisons between their results should be interpreted within their respective contexts. By augmenting our training with a novel individualized missingness mask, we believe that IGNITE captured better dynamics and underlying indicators of the patient's health status. Unlike traditional imputation methods that often assume a specific missingness mechanism, IGNITE is designed to be adaptable to various types of missing data scenarios commonly encountered in healthcare datasets. This adaptability improves the utility of the model in a wide range of clinical applications beyond the datasets used in this study.

This paper has several strengths. First, this is the first work to introduce a novel missingness mask that represents measurement patterns and frequencies in an individualized way. The proposed IMM mask showed high utility in imputation tasks, we believe such masks can be potentially utilized in other diverse time-series tasks such as prediction [97], detection [24], clustering [6] and more. Second, the proposed work is the first to investigate feature-wise missingness and sample-wise missingness in observed healthcare records, bringing a new understanding of missingness patterns in real-world data. We found a correlation between the prevalence of positive outcomes and sampling frequency by investigating the percentage of outcome prevalence across populations with features that were never observed and those with different overall

sample missingness. There was a significantly higher percentage of positive (mortality) outcomes in patients with $\leq$25% of features never observed across all three datasets, where the prevalence of outcomes increased more than double compared to those with higher feature-wise missingness. Similar trends were also observed in sample-wise missingness, suggesting that the missingness pattern is personalized and could inform the underlying patient's health status for various personalized medical applications. Imputation methods that consider individual missingness patterns, such as IGNITE, could lead to more accurate predictions and tailored treatments. Moreover, our feature-wise missingness experiments showcased IGNITE's power to generate features that were never observed in patients. In traditional statistical approaches, many patients with such never-observed features are generally omitted/removed, leading to a waste of data or a reduction in model performance. However, with IGNITE, such features can be generated based on other readily available information about an individual. We believe that a flexible generative approach proposed by IGNITE can open doors for a new way of rethinking missingness imputation as a generative task, making various new applications utilizing generative models.

Another strength of our proposed work is that IGNITE is designed to generate imputations conditioned on individualized patient observed and missingness patterns and dynamics, making the generated imputations of high quality and more aligned with personalized medicine applications. Generating digital twins to drive personalized medicine applications requires understanding and representing patient data beyond observed values, which IGNITE incorporates into its generative process. Lastly, IGNITE was trained and tested across three large-scale EHRs, outperforming state-of-the-art imputation models across a series of experiments. Other experiments, such as the reconstruction experiments, demonstrate IGNITE's robustness and ability to recover original patient values even when random missingness is introduced to patient records. This robust performance across datasets and experimental settings showcases IGNITE's robustness in learning individualized representations and generating high-quality imputations.

Limitations

Overall, this study has some limitations. IGNITE and the other imputation models were tested only on retrospective ICU datasets. In future work, we plan to investigate IGNITE's ability to learn individualized patient representations with higher sparsity and missingness patterns in primary care and wearable data. Furthermore, while our work expands on the evaluation metrics proposed in the literature, IGNITE was only

evaluated in prediction tasks involving mortality, as limited by the available datasets such as Physionet 2012. In future work, we aim to investigate the impact of imputations on various tasks in healthcare settings, such as treatment recommendation and phenotyping. Another area of future work involves the theoretical analysis of the various missingness patterns in the data to better understand the strengths and limitations of the plethora of time-series techniques.

In addition, the datasets selected for the evaluation of the IGNITE model were chosen based on their open-access nature to ensure transparency and reproducibility. These ICU datasets are widely used in healthcare analytics research, allowing for benchmarking against well-established methods in the field. Although the current study focuses on ICU datasets due to their availability and the richness of their clinical data, the principles underlying IGNITE are broadly applicable to other types of EHRs, including hospital-based and primary care settings. While ICU data is characterized by high-frequency monitoring, hospital ward data is also collected at regular intervals, making it structurally similar in terms of time-series imputation challenges. In contrast, primary care EHRs often have lower sampling rates, with visits occurring months apart. Despite these differences, challenges such as missing data, temporal dependencies, and patient-specific variability persist across all EHR settings, reinforcing the need for robust imputation methods like IGNITE. Although longitudinal EHR data in primary care settings spans a longer timeframe, the number of recorded observations per patient remains sparse—often limited to a few entries per year. Since IGNITE models patient trajectories and reconstructs meaningful missing values, its methodological framework remains valid across both high-frequency hospital settings and lower-frequency primary care records.

Furthermore, endpoints such as unplanned hospital admissions or all-cause mortality can be incorporated into the analysis, ensuring the model remains relevant for long-term patient outcome prediction. It is important to note that IGNITE is not designed for wearable sensor data, which involves continuous, high-resolution monitoring and presents fundamentally different modeling challenges. Instead, IGNITE is specifically optimized for structured clinical datasets, where missingness occurs within patient records rather than as a result of real-time streaming gaps. Future research will focus on applying IGNITE to additional datasets from diverse medical settings to further demonstrate its broader applicability and robustness across different healthcare environments. Expanding the evaluation of IGNITE will not only validate its generalizability but also address unique challenges presented in varying clinical contexts, supporting its reproducibility and potential for widespread adoption in clinical

practice. Furthermore, although several works investigated theoretical missingness assumptions such as MCAR, MAR and MNAR [54], such missingness mechanisms are hard to prove and remain far from patterns observed in real-world health data. While IGNITE effectively learns conditional dependencies to reconstruct missing values, it does not assume or infer the underlying reasons for why data is missing. This distinction is important, as handling non-random missingness remains a broader challenge across all imputation methods in clinical data.

Another potential area of future work is to include more benchmarks that might be relevant in medical statistics such as [282, 238]. For the sake of this work, we selected traditional statistical methods such as MICE and LOCF alongside advanced deep learning models, to provide a comprehensive perspective on imputation effectiveness. This selection strategy ensures a robust comparison across different methodological approaches, demonstrating the advances deep learning offers over traditional benchmarks, particularly when learning individualized imputations. The hyperparameters for IGNITE were tuned to ensure that it is robustly configured to each dataset's challenges, enhancing its generalizability and clinical utility. While the chosen configuration reflects optimal empirical performance, we acknowledge that future work could benefit from a formal sensitivity analysis of the loss components and hyperparameter interactions. Such an analysis would help isolate the contribution of each objective term and inform principled weighting strategies in multi-task optimization frameworks. Lastly, while deep learning models like IGNITE require more computational resources during training compared to traditional methods, they significantly expedite the inference process. In our analysis, the improvements in data quality and model performance using IGNITE are statistically significant, validating the trade-off between increased training time and enhanced operational efficiency in practical applications.

Overall, this study has some limitations. IGNITE and the other imputation models were tested only on retrospective ICU datasets. In future work, we plan to investigate IGNITE's ability to learn individualized patient representations with higher sparsity and missingness patterns in primary care and wearable data. Furthermore, while our work expands on the evaluation metrics proposed in the literature, IGNITE was only evaluated in prediction tasks involving mortality, as limited by the available datasets such as Physionet 2012. In future work, we aim to investigate the impact of imputations on various tasks in healthcare settings, such as treatment recommendation and phenotyping. Another area of future work involves the theoretical analysis of the various missingness patterns in the data to better understand the strengths and

limitations of the plethora of time-series techniques. In addition, the datasets selected for the evaluation of the IGNITE model were chosen based on their open-access nature to ensure that our results are transparent and reproducible. These ICU datasets are commonly used in healthcare analytics research, allowing us to benchmark our results against well-established methods in the field. Although the current study focuses on ICU datasets due to their availability and richness of their clinical data, we recognize that this may limit the perceived applicability of our method. However, it is important to note that the principles underlying the IGNITE model are broadly applicable to other types of clinical data. Future research will aim to apply IGNITE to additional datasets from diverse medical settings, demonstrating its broader applicability and exploring its potential in other areas of clinical practice. This expansion will also allow us to test the model's robustness across different healthcare environments, addressing any unique challenges they present, and supporting the reproducibility of our findings.

Another potential area of future work is to include more benchmarks that might be relevant in medical statistics such as [282, 238]. For the sake of this work, we selected traditional statistical methods such as MICE and LOCF alongside advanced deep learning models, to provide a comprehensive perspective on imputation effectiveness. This selection strategy ensures a robust comparison across different methodological approaches, demonstrating the advances deep learning offers over traditional benchmarks, particularly when learning individualized imputations. Lastly, while deep learning models like IGNITE require more computational resources during training compared to traditional methods, they significantly expedite the inference process. In our analysis, the improvements in data quality and model performance using IGNITE are statistically significant, validating the trade-off between increased training time and enhanced operational efficiency in practical applications.

5.6 Relevant Publications

- **Ghosheh, G.**, Li, J., Zhu, T. (2023). A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Computing Surveys*, 56(4), Article 63.

- **Ghosheh, G. O.**, Li, J., & Zhu, T. (2024). IGNITE: Individualized GeNeration of Imputations in Time-series Electronic Health Records. arXiv preprint arXiv:2401.04402. NPJ Digital Medicine [Under Review]

Chapter 6

Relaxing Sequential Ignorability in ITE Estimation from Time-series EHRs

The preceding chapter introduced individualized imputation in time-series data to improve data fidelity. We now shift focus to causal inference and explore how missingness mechanisms can influence treatment effect estimation. Rather than reiterating the challenges of confounding and observational bias, this chapter proposes SI-Mask, a model that dynamically adapts to missingness patterns while preserving causal identifiability. It advances the prior work by bridging imputation with individualized treatment effect (ITE) modeling in the presence of temporal gaps.

6.1 Introduction

Estimating the individual treatment effect (ITE) in time-series data presents significant challenges, especially when dealing with missing data, a common issue in electronic health records (EHRs). Missingness in EHRs arises from various sources, such as skipped documentation, system errors, or unrecorded patient measurements[87], introducing time-varying confounders related to missingness, which are missing variables that influence both treatment assignment and outcome, leading to potential biases in treatment effects estimations.

A fundamental assumption for ITE in time-series data is sequential ignorability, which asserts that given the observed covariates, future potential outcomes are independent of treatment assignments [206]. However, the violation of this assumption by confounding data patterns because of missing data often leads to biased estimates and unreliable inferences. Previous works in ITE estimation for time-series data focus

on scenarios where they account for observed confounding and assume that there are no hidden confounders in a sequentially ignorable environment. These works include Counterfactual Recurrent Networks (CRN) [20], Recurrent Marginal Structural Models (RMSN) [154], and G-Net [151], which assume that all relevant confounders are observed and often use imputation techniques such as the Last Observation Carried Forward (LOCF). However, these methods apply strict assumptions related to mechanisms and the nature of confounding and fail to account for missingness, limiting their effectiveness in real-world applications.

In this work, we introduce the concept of partially hidden confounders, which are variables that influence both the treatment and the outcome, but are only partially observed or measured in the dataset. Some individuals have observed values for the confounder, while others have missing or unmeasured values. Partially hidden confounders can also introduce bias, especially if the missingness is related to both the treatment and the outcome. For example, in a study investigating the effect of a new drug/medication ($\mathbf{A}$, treatment) on improving recovery rates from chronic disease ($\mathbf{Y}$, outcome), socioeconomic status ($\mathbf{Z}$, partially hidden confounder) could be a significant factor. Socioeconomic status can influence both the likelihood of receiving the new medication and overall health and recovery rates. However, due to incomplete surveys or non-disclosure, socioeconomic status data might only be available for some patients over some time steps, leading to partial observation. This partial observation can introduce bias in the estimated effects of the medication if not addressed properly. The rest of this work focuses on this type of confounder.

To overcome these limitations, we introduce a novel framework, the Sequentially Ignorable Mask (SI-mask), designed to relax sequential ignorability specifically in the context of missingness-related confounding. This framework employs an adversarial training approach to assign weights to each data point in the time-series covariates, aiming to balance the impact of confounding factors related to missingness. By ensuring that the resultant mask assigns weights to data points that make them predictive of outcomes but not of treatment assignments, we effectively deconfound the time-varying covariates, offering a robust solution for causal inference even with incomplete data. Our contributions are as follows:

- **Adversarial Training for Deconfounding**: - Our framework employs adversarial training, where two neural networks, one for treatment prediction and one for outcome prediction, are trained simultaneously. The adversarial approach involves a treatment model (the adversary) that predicts treatment assignments

and an outcome model (the primary model) that predicts outcomes. The training objective is to ensure that the representations learned by the outcome model are predictive of the outcome but not of the treatment, effectively balancing out the confounding factors.

- **Sequential Ignorability Mask**: We propose a novel Sequential Ignorability mask (i.e., SI-mask) that directly addresses the confounding effects due to missingness in time-series data. This mask relaxes the traditional assumption of sequential ignorability in ITE by introducing weights assigned to each time-varying data point to mitigate the bias introduced by missing values. The mask is constructed in such a way that it is informative for the prediction of the outcome while remaining uninformative for the prediction of the treatment. The SI-mask generates individualized weights for each patient's time-series data, ensuring personalized deconfounding by accounting for the missingness patterns in the patient's data.
- **Flexible Integration with Outcome Estimator Models**: The SI-mask framework is highly flexible and can be integrated into existing ITE outcome estimators without modifications to their architecture. The SI-mask framework improves the robustness and performance of ITE estimators by deconfounding the data, thereby addressing the biases introduced by missing values.

This work addresses a critical gap in the literature and sets the stage for more accurate and reliable causal inference in various domains, thus contributing to the broader goal of making data-driven decisions more effective and trustworthy.

6.2 Related Work

Addressing time-varying confounding in longitudinal studies is a key challenge in causal inference. One widely used statistical approach is the inverse probability of treatment weighting (IPTW). IPTW redistributes the population by applying weights to individuals to eliminate bias introduced by time-varying confounders [261]. These weights are calculated as the inverse of the probability of receiving the treatment given an individual's covariate history up to each time point. By constructing a pseudo-population where the distribution of confounders is independent of treatment assignment, IPTW ensures that previous treatment and covariate history do not influence current treatment assignment [12]. This technique is particularly useful in

longitudinal studies, where confounders evolve over time and affect both treatment and outcome.

Marginal Structural Models and Extensions: The application of IPTW in time-series data is facilitated by marginal structural models (MSMs) [261]. MSMs are designed to handle time-dependent confounding by creating weights that account for the history of time-varying covariates and their effect on treatment assignment. However, MSMs are based on correctly specified treatment models and are sensitive to extreme weights, especially in high-dimensional settings. Recurrent marginal structural networks (RMSN) address some of these limitations by incorporating recurrent neural networks (RNNs) to estimate treatment probabilities and counterfactual outcomes [154]. RMSNs demonstrate the potential of deep learning to improve traditional IPTW-based approaches, but remain limited in their handling of missing data.

G-computation : G-computation, or the G-formula, is another flexible method for estimating causal effects in the presence of time-varying confounding [203]. It models the entire data-generating process, including relationships between covariates, treatments, and outcomes. While effective when models are correctly specified, G-computation can be computationally intensive and requires large sample sizes [203]. Furthermore, its reliance on predefined models for relationships among covariates, treatments, and outcomes makes it less adaptable to highly complex or nonlinear settings.

Balanced Representation Models : Deep learning approaches have emerged as powerful tools for treatment effect estimation by creating balanced representations of the treatment and control groups. Early methods, such as those proposed by [120], focused on static settings. Recent advances, such as Counterfactual Recurrent Networks (CRN) [20] and CausalTransformer [176], have extended these methods to time-series data. CRN utilizes adversarial training within a sequence-to-sequence framework to learn representations predictive of outcomes rather than treatment assignments. CausalTransformer leverages transformer-based architectures with treatment-invariant representations to handle time-varying variables, treatments, and outcomes. These methods effectively mitigate confounding but treat missingness as a preprocessing problem rather than integrating it directly into the causal inference framework.

Adversarial Training for Balancing Representations : Adversarial approaches, such as DeepMatch [124] and adversarial balancing [194], focus on balancing covariates to create treatment-invariant representations. These methods have proven effective in static settings, but face challenges in adapting to time-varying contexts where confounders evolve dynamically. While DeepMatch and adversarial balancing optimize for static covariate balance, they do not address missingness or the interaction between time-varying confounders and treatment assignments. SI-Mask extends these principles to time-series data by embedding adversarial training within a dynamic masking framework. This allows SI-Mask to handle evolving confounders and missingness simultaneously, which remains a gap in prior work.

Missingness and Hidden Confounders : Although causal inference frameworks such as Time-Series Deconfounder [19], Deep Sequential Weighting (Deep-SW) [158], and Sequential Deconfounder [100] address hidden confounders, they do not explicitly address missingness-related confounding. Deep-SW combines IPTW with RNNs to handle complex data structures, while Time-Series Deconfounder employs Gaussian Processes to infer substitutes for hidden confounders. Sequential Deconfounder extends these ideas by using RNNs to infer substitutes for confounders at each time step. These approaches provide valuable tools for handling hidden confounders, but leave the issue of missingness-related confounding largely unaddressed.

SI-Mask Framework : The Sequentially Ignorable Mask (SI-Mask) framework bridges these gaps by introducing a dynamic masking mechanism to address biases from missingness and time-varying confounders simultaneously. Unlike traditional approaches based on IPTW or MSM, SI-Mask integrates adversarial training with a learned masking process to dynamically weight covariates during training. This allows SI-Mask to adapt to missingness patterns and mitigate confounding in real time. Unlike balanced representation models that rely on imputation as a preprocessing step, SI-Mask embeds the masking mechanism directly within the training process, reducing the risk of biases introduced by imputation. Furthermore, by leveraging adversarial training, SI-Mask ensures covariate balance between treatment groups while maintaining predictive power for outcomes. This makes it particularly suited for high-dimensional, time-series datasets, where missingness and confounding interact in complex ways.

6.3 Problem Setting

In this section, we formalize the problem of ITE estimation for time-series data and introduce the challenge of missingness, which our proposed work aims to address.

Time-Series ITE Estimation

Consider a dataset comprising N patients, each observed in t time steps with d features per time step. The covariates for patient i at time t are denoted as $\mathbf{X}_{i,t} \in \mathbb{R}^D$, forming the complete covariate matrix $\mathbf{X} \in \mathbb{R}^{N \times T \times D}$. Each patient receives a sequence of treatments $\mathbf{A}_{i,t} \in \{0, 1\}$, where 0 denotes no treatment and 1 denotes treatment. The outcomes for each patient are represented by $\mathbf{Y}_{i,t} \in \mathbb{R}$. The objective of the ITE estimation is to predict the potential outcomes $\mathbf{Y}^0_{i,t}$ and $\mathbf{Y}^1_{i,t}$ for each patient i at each time step t, under the control conditions (no treatment) and the treatment conditions, respectively. The ITE for patient i at time t is defined as:

$$\tau_{i,t} = \mathbf{Y}^1_{i,t} - \mathbf{Y}^0_{i,t} \tag{6.1}$$

To achieve this, we employ an ITE estimator $\mathcal{E}$ that takes as input the covariates $\mathbf{X}_{i,t}$ and predicts the potential outcomes $\mathbf{Y}_{i,t}$.

Missingness in EHR Time-Series Data

In EHR time-series data, missing values are common due to various reasons such as irregular sampling, data entry errors, or missing patient appointments. Let $\mathbf{O}_{i,t} \in \{0, 1\}^d$ be an observation mask that indicates whether the covariate $\mathbf{X}_{i,t,d}$ is observed (1) or missing (0). The presence of missing values can introduce bias into the ITE estimation if not addressed properly.

The observed covariates can be expressed as:

$$\tilde{\mathbf{X}}_{i,t,d} = \begin{cases} \mathbf{X}_{i,t,d} & \text{if } \mathbf{O}_{i,t,d} = 1 \\ \text{NaN} & \text{if } \mathbf{O}_{i,t,d} = 0 \end{cases} \tag{6.2}$$

Challenge of Missingness and Confounding

Missing data can lead to confounding bias if the mechanism behind the missingness is related to treatments or outcomes. For example, if patients with no serious conditions are more likely to have missing data, any naive imputation or analysis could produce biased ITE estimates. Traditional methods such as Last Observation Carried Forward (LOCF) or simple imputation do not account for the potential confounding introduced by missingness, leading to biased estimates.

Causal Inference Assumptions

Building on the framework for prospective outcomes by [203] and its extension to dynamically changing outcomes and treatments over time[204], our study takes advantage of the potential outcomes framework, as commonly used in recent research [20, 176]. To effectively identify a counterfactual outcome distribution over time, specifically the average τ-step-ahead potential outcome conditioned on observed history, we adhere to three pivotal assumptions:

- **Consistency**: Under this assumption, if the treatment sequence up to time t for any patient is $\overline{\mathbf{A}}_t = \overline{\mathbf{a}}_t$, then the potential outcome $\mathbf{Y}_{t+1}[\overline{\mathbf{a}}_t]$ is identical to the observed outcome $\mathbf{Y}_{t+1}$. This implies that, given the history of treatment $\overline{\mathbf{A}}_t = \overline{\mathbf{a}}_t$, the observed outcome mirrors the outcome that would have occurred under that specific treatment sequence.

- **Sequential Overlap**: This assumption ensures that for any possible history $\overline{\mathbf{H}}_t$ with a positive probability of occurring, there remains a non-zero probability of receiving any specific treatment sequence $\mathbf{A}_t = \mathbf{a}_t$. Formally, if $\mathbb{P}(\overline{\mathbf{H}}_t = \overline{\mathbf{h}}_t) > 0$, then $0 < \mathbb{P}(\mathbf{A}_t = \mathbf{a}_t \mid \overline{\mathbf{H}}_t) < 1$. This ensures that every treatment option is possible at any time given the patient's history.

- **Sequential Ignorability**: This critical assumption posits that the choice of treatment at any time t, represented as $\mathbf{A}_t$, does not depend on future outcomes beyond the history observed up to that point, $\overline{\mathbf{H}}_t$. This implies that $\mathbf{A}_t \perp \mathbf{Y}_{t+1}[\mathbf{a}_t] \mid \overline{\mathbf{H}}_t$ for all $\mathbf{a}_t$, which means that there are no hidden biases or unmeasured confounders that simultaneously influence treatment and outcome. Often referred to as sequential exchangeability or the absence of unobserved confounders, this assumption highlights that all factors that influence treatment decisions and outcomes are captured or observed.

Relaxing the Sequential Ignorability Assumption with the SI-Mask Framework

The Sequential Ignorability assumption states that, conditional on the observed covariates, the treatment assignment is independent of the potential outcomes. However, when missingness is present in the covariates, this assumption can be violated, resulting in a biased ITE estimation. It is important to distinguish this assumption

from the sequential ignorability used in mediation analysis, which focuses on the decomposition of natural direct and indirect effects. To ensure clarity, we explicitly differentiate between these contexts within our theoretical framework.

The SI-Mask framework addresses the challenge of biased estimation due to missing data by learning a mask, $\mathbf{SI-Mask}_{i,t} \in \mathbb{R}^D$, which dynamically adjusts the weights of the observed covariates. The adjusted covariates are optimized to predict the outcome while minimizing their predictiveness of treatment assignment. This process effectively ensures that the Sequential Ignorability assumption holds even in the presence of missing data, thereby reducing confounding bias in the ITE estimation process. The SI-Mask is constructed through adversarial training, balancing two objectives: maximizing the predictiveness of the outcome and minimizing the predictiveness of the treatment. This mechanism enables the framework to mitigate the biases introduced by missing data dynamically.

It is important to clarify that the title "Relaxing Sequential Ignorability" refers to the framework's ability to handle covariates with missingness directly, without the need to impute observations. The assumption of Sequential Ignorability, defined as the conditional independence of potential outcomes from treatment assignments given the observed covariates, remains central to the framework. However, our approach does not address scenarios involving unobserved confounders of the Sequential Ignorability assumption, which would necessitate fundamentally different methodologies.

The next section provides a detailed explanation of the adversarial training methodology used to generate the SI-Mask and its integration with existing ITE estimation models.

6.4 Theoretical Assumptions & Identifiability

In this section, we discuss the identifiability of the causal estimand in the presence of time-varying confounding and missing data. We provide theoretical insights into how the proposed SI-Mask framework ensures identifiability under certain assumptions.

6.4.1 Identifiability Challenges with Missing Data

Estimating causal effects from observational data with missing covariates poses significant challenges to identifiability. Without assumptions about the missingness mechanism, the causal effect is generally not identifiable. This is because missing data can induce unobserved confounding, making it impossible to distinguish between the effects of treatment and the effects of confounders that are not fully observed.

Missingness Mechanism Assumptions: Our framework does not assume a specific missingness mechanism, such as Missing Completely at Random (MCAR) or Missing At Random (MAR). Instead, we operate under the assumption that the observed data, combined with the SI-Mask adjustments, can mitigate biases introduced by missingness. Although this broader assumption makes the framework more flexible, it does not address Missing Not At Random (MNAR) scenarios where missingness depends on unobserved variables.

6.4.2 Assumptions for Identifiability

To address the identifiability issue in the presence of missing data and time-varying confounding, we rely on the following assumptions:

- **Sequential Ignorability with Observed Data**: We assume that, conditional on the observed covariates and past treatment history, the treatment assignment at each time point is independent of future potential outcomes. Formally, for all t:

$$Y_{t+1}(a) \perp A_t \mid \bar{X}_t, \bar{A}_{t-1}, \tag{6.3}$$

  where $Y_{t+1}(a)$ is the potential outcome at time $t+1$ under treatment sequence a, $\bar{X}_t$ denotes the history of observed covariates up to time t, and $\bar{A}_{t-1}$ denotes the treatment history up to time $t-1$.

- **Positivity (Overlap)**: There is a non-zero probability of receiving each treatment given the history of observed covariates. Formally, for all possible treatment assignments a_t and observed histories:

$$0 < P(A_t = a_t \mid \bar{X}_t, \bar{A}_{t-1}) < 1. \tag{6.4}$$

Unverifiability of Missingness Mechanisms: It is important to note that the missingness mechanism (e.g., MCAR, MAR, or MNAR) cannot be empirically verified from the observed data alone. The SI-Mask framework operates under the assumption that learned adjustments on the observed data can mitigate biases introduced by missingness, without requiring verification of the missingness mechanism [157].

6.5 Methodology

6.5.1 SI-Mask Generation

We propose a novel framework for generating treatment-invariant representations through adversarial training, aimed at domain adaptation in time-series data. Our approach, referred to as *Adversarial Model*, integrates two primary components: the Treatment Model and the Outcome Model. These models collaboratively generate SI-Masks, which are subsequently utilized to deconfound the input data and produce robust, treatment-invariant representations with respect to time-series missingness. We show an overview of the proposed approach in Figure 6.1.

6.5.2 Dynamic Masking Mechanism

The SI-Mask framework introduces a dynamic masking mechanism to address missingness and time-varying confounding simultaneously. This mechanism generates weights W_t for observed covariates based on their relevance to the outcome of interest. The masking process is integrated into the adversarial training pipeline as follows: This mechanism differs from previous approaches [124] by embedding dynamic masking directly into the training process, rather than relying on static weighting or preprocessing steps such as imputation.

6.5.3 Proposed Model

Treatment Model: The Treatment Model, denoted as $G_a(X;\theta_a)$, consists of an LSTM layer followed by dense layers. The final dense layer, denoted as output_layer, produces a probability distribution over the treatment classes:

$$G_a(X;\theta_a) = \sigma\big(\mathbf{W}_a \cdot \mathrm{ReLU}(\mathbf{U}_a \cdot \mathrm{LSTM}(X) + \mathbf{b}_a) + \mathbf{c}_a\big) \tag{6.5}$$

Outcome Model: The Outcome Model, denoted as $G_y(X;\theta_y)$, mirrors the architecture of the Treatment Model, with an LSTM layer and dense layers. The final layer predicts outcome values, constrained to the $[0,1]$ range using a sigmoid activation function:

$$G_y(X;\theta_y) = \sigma\big(\mathbf{W}_y \cdot \mathrm{ReLU}(\mathbf{U}_y \cdot \mathrm{LSTM}(X) + \mathbf{b}_y) + \mathbf{c}_y\big) \tag{6.6}$$

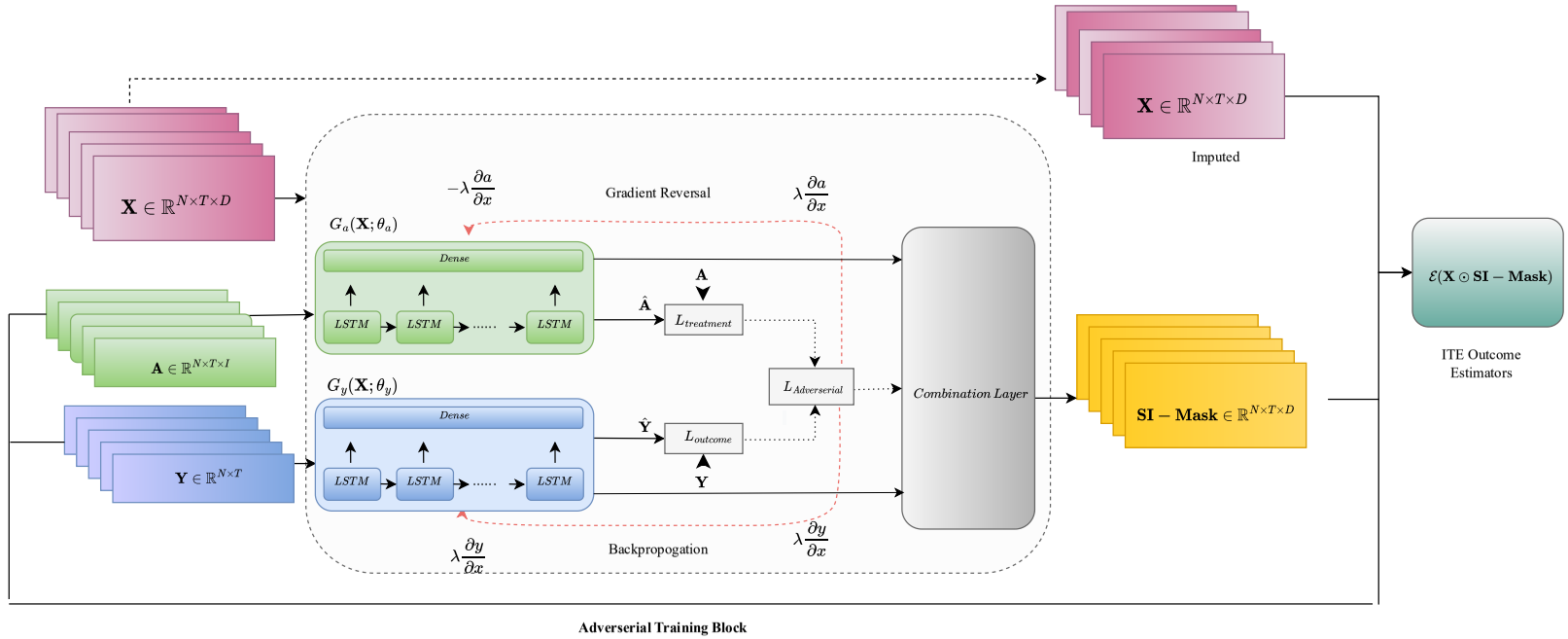


Figure 6.1: This figure illustrates the SI-Mask framework for estimating Individual Treatment Effects (ITE) in the presence of highly missing time-series data. The framework takes as input population records X, treatment assignments A, and observed outcomes Y. The Adversarial Training Block includes treatment and outcome predictors denoted as $G_a(X; \theta_a)$ and $G_y(X; \theta_y)$, respectively. These models are optimized jointly using adversarial loss. Final ITE estimation is performed by feeding the imputed covariates, obtained via SI-Mask, into treatment-specific ITE estimators $\mathcal{E}$.

Combination Layer: To integrate the outputs of the Treatment Model and Outcome Model, we introduce a trainable combination weight α. The combination layer linearly combines the outputs of the two models:

$$\text{Combined Rep} = \alpha \cdot G_a(X; \theta_a) + (1 - \alpha) \cdot G_y(X; \theta_y) \tag{6.7}$$

Adversarial Training Procedure

Gradient Reversal Layer: To implement adversarial training, we use a gradient reversal layer in the treatment model [78]. During the forward pass, the gradient reversal layer acts as an identity function, passing the input directly to the next layer. However, during the backward pass, this layer multiplies the gradient by a negative scalar, effectively reversing the gradient. This reversal ensures that the treatment model is trained to ensure that the SI-mask is not predictive of the treatment assignment.

Adversarial Loss: The adversarial loss is defined as the difference between the prediction loss of the outcome and the prediction loss of the treatment. Specifically, the outcome loss is calculated using the mean squared error (MSE) between the predicted and true outcomes, while the treatment loss is computed using categorical cross-entropy (CCE) between the predicted and true treatments. Let $\hat{y}_a$ and y_a be the predicted and true treatments, and $\hat{y}_y$ and y_y be the predicted and true outcomes. The losses are defined as follows:

$$\mathcal{L}_{\text{outcome}} = \frac{1}{N} \sum_{i=1}^{N} (y_{O,i} - \hat{y}_{O,i})^2 \tag{6.8}$$

$$\mathcal{L}_{\text{treatment}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{T,i,j} \log(\hat{y}_{T,i,j}) \tag{6.9}$$

The adversarial loss is given by:

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{outcome}} - \lambda \mathcal{L}_{\text{treatment}} \tag{6.10}$$

where λ is a hyperparameter that controls the trade-off between treatment and outcome losses.

The SI-Mask generation process involves the following steps, as shown in the algorithm 3.

Algorithm 3 SI-Mask Generation Process

Input:
  X : Input data
  θ_a : Parameters of the Treatment Model
  θ_y : Parameters of the Outcome Model
  λ : Weighting factor for treatment loss
  α : Combination weight for parameter updates
  LR : Learning rate for optimizer
Begin
  Forward Pass
  $G_a(X;\theta_a) = \sigma(\mathbf{W}_a \cdot \text{ReLU}(\mathbf{U}_a \cdot \text{LSTM}(X) + \mathbf{b}_a) + \mathbf{c}_a)$
  $G_y(X;\theta_y) = \sigma(\mathbf{W}_y \cdot \text{ReLU}(\mathbf{U}_y \cdot \text{LSTM}(X) + \mathbf{b}_y) + \mathbf{c}_y)$
  Apply GradientReversalLayer to $Treatment\_yutput$
  Loss Calculation
  $L_{treatment} \leftarrow \text{CCE}(y_a, G_a(X;\theta_a))$
  $L_{outcome} \leftarrow \text{MSE}(y_y, G_y(X;\theta_y)$
Adversarial Loss Calculation
  $L_{adv} \leftarrow L_{outcome} - \lambda \cdot L_{treatment}$
  Gradient Calculation and Update
  Compute gradients for $\theta_a, \theta_y, \alpha$ using L_{adv}
  Update θ_a, θ_y using Adam optimizer with gradients and LR
SI-Mask Application
  $CombinationLayer \leftarrow \alpha \cdot G_a(X;\theta_a) + (1-\alpha) \cdot G_y(X;\theta_y)$
  Normalize SI_Mask to $[0, 1]$ range
  return SI_Mask
End

Integration with ITE Estimators

The SI-mask framework can be seamlessly integrated with existing ITE estimators to enhance their performance by mitigating the bias introduced by missing data. Let $\mathcal{E}$ denote an ITE estimator that takes as input the weighted covariates $\mathbf{X} \odot \mathbf{M}$, where $\odot$ denotes element-wise multiplication. The ITE estimator then predicts the potential outcomes $\hat{\mathbf{Y}}^0$ and $\hat{\mathbf{Y}}^1$ under control and treatment conditions, respectively. The integration process involves three main steps: first, mask generation, where the treatment model and outcome model are trained using adversarial training to generate the SI-mask $\mathbf{M}$; second, weight adjustment, where the SI-mask is applied to the covariates $\mathbf{X}$ to obtain the weighted covariates $\mathbf{X} \odot \mathbf{SI} - \mathbf{M}$; and third, ITE estimation, where the weighted covariates $\mathbf{X} \odot \mathbf{M}$ are input into the ITE estimator $\mathcal{E}$ to predict potential outcomes. The predicted ITE for each patient i is given by $\hat{\tau}_i = \mathcal{E}(\mathbf{X}_i \odot \mathbf{M}_i)$. By integrating the SI-mask with the ITE estimators, we enhance their ability to accurately predict individual treatment effects by reducing the confounding bias associated with missing data. This approach is flexible and can be applied to various ITE estimation frameworks, making it a valuable addition to the field of causal inference in time-series data.

Missing values in the covariates were imputed using the Last Observation Carried Forward (LOCF) method, a widely used imputation technique in time-series analysis. This approach replaces missing values with the most recent non-missing value, ensuring that the temporal structure of the data is preserved. The observation mask M_t, which indicates the location of missing values, was included as an additional input to the model to improve its ability to account for missingness during training.

Implementation

The SI-Mask generation process was implemented using TensorFlow, and training was conducted on NVIDIA GeForce RTX 3090 GPUs. We use an exponential decay learning rate schedule for the Adam optimizers. Training proceeds for 100 epochs with a batch size of 32. The regularization weight α was dynamically adjusted during training by linearly increasing it from 0 to a maximum value of 1 over the first 20 epochs. This annealing schedule was used to stabilize early training and progressively enforce the regularization term.

6.5.4 Dataset Description

Simulated Data

The tumor growth (TG) simulator proposed by Geng et al. [**?**] simulates tumor volume Y_{t+1} at time $t+1$ days after cancer diagnosis. This results in a univariate outcome ($d_y = 1$). The simulator includes two binary treatment options: (i) radiotherapy (A_a^r) and (ii) chemotherapy (A_a^c), each having distinct effects on tumor growth. Specifically, (i) radiotherapy exerts an immediate effect $d(t)$ on the next tumor volume, while (ii) chemotherapy introduces an effect $C(t)$ that decays exponentially over time. The evolution of tumor volume is governed by the following equation:

$$Y_{t+1} = \left(1 + \rho \log\left(\frac{K}{Y_a}\right) - \beta_c C_a - (\alpha_r d_a + \beta_r d_a^2) + \epsilon_a\right) Y_a,$$

where ρ, K, β_c, α_r, and β_r are simulation parameters, and $\epsilon_a \sim N(0, 0.012)$ represents independent noise. Parameters β_c, α_r, and β_r capture individual patient responses, sampled from a mixture of truncated normal distributions with three components. In addition, the mixture component indices are treated as static covariates ($d_v = 1$).

Time-varying confounding is introduced through a biased treatment assignment process applied to both treatments, modeled as follows:

$$A_a^c, A_a^r \sim \text{Bernoulli}\left(\sigma\left(\gamma \frac{D_{\max}}{\bar{D}_{15}(\bar{Y}_{t-1}) - D_{\max}/2}\right)\right),$$

where $\sigma(\cdot)$ denotes the sigmoid function, $D_{\max}$ is the maximum tumor diameter, and $\bar{D}_{15}(\bar{Y}_{t-1})$ represents the average tumor diameter during the last 15 days. The parameter γ controls the degree of confounding, where increasing γ introduces a greater bias in assignment.

In our simulation, the two binary treatments are combined into a single categorical variable representing four possible treatment states:

$$\{(A_a^c = 0, A_a^r = 0), (A_a^c = 1, A_a^r = 0), (A_a^c = 0, A_a^r = 1), (A_a^c = 1, A_a^r = 1)\}.$$

To increase the complexity of the task while maintaining a simplified structure, we simulate missing data under a missing completely at random (MCAR) assumption. Under MCAR, the probability that a data point is missing is independent of both observed and unobserved variables, including treatment assignment. For evaluation, the models are retrained on five simulated datasets using different random seeds.

We report the average root mean square error (RMSE) in the holdout test set. In addition, a normalized RMSE is computed by scaling the error using the maximum tumor volume, $V_{\max} = 1150$ cm^3.

Real Data

The evaluation of ITE estimators in real-world data was performed using the Medical Information Mart for Intensive Care (MIMIC-III) dataset, a renowned dataset from the United States detailing the patient journeys of the intensive care unit [120]. From MIMIC-III, we meticulously extracted 25 time-varying covariates, which include vital signs and laboratory results. Our primary outcome of interest was diastolic blood pressure, with a specific focus on two treatments: vasopressors and mechanical ventilation. The administration of these treatments over time was carefully recorded. Predicting blood pressure is of the utmost importance in critical care medicine, particularly to prevent severe complications such as tissue hypoperfusion. The challenge lies in the fact that the use of vasopressors is intricately linked to both previous and current blood pressure levels, acting as a time-varying confounder. Optimizing vasopressor administration to achieve desired outcomes in blood pressure remains a complex task that requires further understanding. We selected patients with ICU stays of at least 30 hours. For these patients, we limited the stay in the ICU to 60 hours. We divided this cohort into training, validation, and test subsets with proportions of 70%, 15%, and 15%, respectively. The implementation was adjusted according to the projection horizon τ.

(i) For one-step-ahead predictions, all the trajectories of the test set were used.

(ii) For the τ-step-ahead predictions where $\tau \geq 2$, we followed this approach: Let $\tau_{\max}$, which is at least τ, be the maximum projection horizon, set at 5 in our study. We extracted all sub-trajectories with a length of at least $\tau_{\max} + 1$ using a rolling origin, removing vital signs from time steps 1 to $T(i) - \tau_{\max} + 1$. Predictions were made while preventing future insights by masking. Performance results were reported only for predictions in the τ step ahead.

The covariates that are included in the time-varying MIMIC III dataset are as follows: heart rate, red blood cell count, sodium, mean blood pressure, systemic vascular resistance, glucose, chloride urine, hematocrit, positive end-expiratory pressure set, respiratory rate, prothrombin time (PT), cholesterol, hemoglobin, creatinine, blood

urea nitrogen, bicarbonate, ionized calcium, partial pressure of carbon dioxide, magnesium, anion gap, phosphorus, platelets, and calcium urine. Experimental results on another dataset will be provided in the Appendix.

6.5.5 Benchmarks

In this work, we evaluate our proposed approach in three commonly used ITE outcome estimation models. We specifically benchmark against, Marginal Structural Models (MSMs) [205], Recurrent Marginal Structural Networks (RMSNs) [154], Counterfactual Recurrent Network (CRN) [20] and Causal Transformer [176].

We have developed integrated SI-Mask variants for each of the ITE outcome estimator models. Our methodology improves ITE outcome estimation by integrating SI-Mask with various models, assigning deconfounded weights to account for missingness. These variants include SI-MSM, SI-RMSN, SI-CRN, and SI-Causal Transformer. We compare the performance of every model with its corresponding SI pair. Throughout this chapter, the term Individual Treatment Effect (ITE) refers to the difference in potential outcomes under treatment and control for a given individual—consistent with standard usage in econometrics and causal inference. On synthetic datasets, we validate ITE estimates directly using RMSE because ground-truth counterfactuals are known. In real-world datasets like MIMIC-III, where counterfactuals are not observable, model performance is evaluated indirectly via prediction accuracy.

6.6 Results

This section presents the experimental evaluation of the SI-Mask framework on both simulated and real-world datasets. The results are organized around two primary objectives: (1) assessing the robustness of SI-Mask under varying degrees of confounding and missingness, and (2) demonstrating its performance across a diverse set of baseline models commonly used in treatment effect estimation tasks.

Evaluation on Simulated Data

Figure 6.2 presents the results of experiments on simulated tumor growth data, showing normalized RMSE values as mean $\pm$ standard deviation across five runs for varying levels of confounding ($\gamma = 1$ to $\gamma = 4$). The figure highlights a consistent and significant performance improvement of SI variants over their standard counterparts across all models: MSM, RMSN, CRN, and CausalTransformer. This improvement becomes

more pronounced as the level of confounding increases, demonstrating the robustness of SI variants in handling time-varying confounding and missingness in the data.

For example, in the case of RMSN, the SI variant exhibits a consistent reduction in RMSE at all levels of confounding. At $\gamma = 4$, SI-RMSN achieved an RMSE of 0.80 ± 0.11, compared to 1.18 ± 0.11 for RMSN, highlighting a substantial improvement in prediction accuracy. Similarly, the SI-CausalTransformer variant maintained a more stable and accurate performance with an RMSE of 1.01 ± 0.25 at $\gamma = 4$, in contrast to its standard counterpart's higher variability and error.

The SI variants of the MSM and CRN models also show consistent improvements across confounding strengths. SI-MSM significantly reduced RMSE at higher γ values, indicating enhanced stability in estimating treatment effects under high-bias scenarios. Likewise, SI-CRN demonstrated superior performance, achieving a lower RMSE of 0.86 ± 0.18 at $\gamma = 4$, compared to 1.20 ± 0.08 for the standard CRN.

These results underline the adaptability and robustness of SI-Mask across different architectures, confirming that the framework generalizes well to various causal modeling approaches. This makes SI-Mask a valuable enhancement for treatment effect estimation pipelines in contexts with high data missingness and confounding — such as retrospective EHR studies and longitudinal health registries.

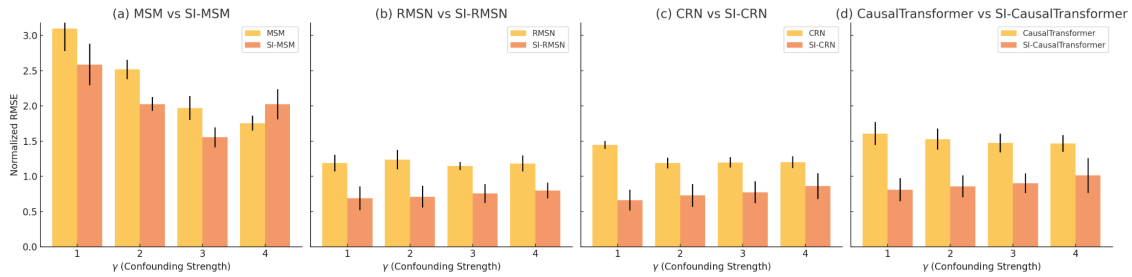


Figure 6.2: Comparison of models across varying levels of confounding for simulated tumor growth data. The normalized RMSE (Root Mean Square Error) is shown for each model with error bars indicating the standard deviation.

Evaluation on Real Data

We further evaluate SI-Mask using the MIMIC-III dataset, a real-world ICU dataset with inherent missingness and time-dependent clinical variables. Table 6.1 summarizes the performance of each model across prediction horizons ($\tau = 1$ to $\tau = 5$). Once again, SI-Mask consistently improves RMSE across all model variants and time steps.

The SI-CRN model achieves the lowest RMSE at $\tau = 1$ (9.834 ± 0.216), outperforming the standard CRN and maintaining this lead across all time steps. SI-RMSN also improves upon RMSN at every time point, with a lowest RMSE of 10.406 ± 0.3608 at $\tau = 1$, indicating that the SI enhancement improves generalization even in complex clinical settings.

The SI-CausalTransformer demonstrates strong robustness across longer horizons, achieving RMSE values of 11.94 ± 0.25 at $\tau = 5$, outperforming its baseline despite the model's complexity and the difficulty of longer-term prediction.

While some RMSE differences appear modest, the improvements are consistent across runs, models, and prediction intervals. This consistency is a key strength of SI-Mask: it enhances both average performance and reliability. These results reaffirm the utility of SI-Mask in healthcare applications, particularly where model robustness under missingness is essential.

Table 6.1: Results for experiments with real-world medical data (MIMIC-III). Shown: Normalized RMSE as mean ± standard deviation over five runs.

Model	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
MSM	11.748 ± 0.266	11.250 ± 0.469	11.634 ± 0.561	11.724 ± 0.532	12.154 ± 0.504
SI-MSM	$\mathbf{11.651 \pm 0.198}$	$\mathbf{11.122 \pm 0.134}$	$\mathbf{11.406 \pm 0.494}$	$\mathbf{11.692 \pm 0.534}$	$\mathbf{12.040 \pm 0.544}$
RMSN	10.536 ± 0.3722	11.150 ± 0.569	11.434 ± 0.561	11.724 ± 0.532	12.054 ± 0.504
SI-RMSN	$\mathbf{10.406 \pm 0.3608}$	$\mathbf{11.022 \pm 0.434}$	$\mathbf{11.306 \pm 0.494}$	$\mathbf{11.592 \pm 0.534}$	$\mathbf{11.930 \pm 0.544}$
CRN	$\mathbf{9.829 \pm 0.219}$	10.301 ± 0.210	10.580 ± 0.192	10.814 ± 0.200	11.052 ± 0.230
SI-CRN	9.834 ± 0.216	$\mathbf{10.298 \pm 0.205}$	$\mathbf{10.574 \pm 0.196}$	$\mathbf{10.806 \pm 0.202}$	$\mathbf{11.042 \pm 0.226}$
CausalTransformer	10.960 ± 0.382	11.430 ± 0.368	11.714 ± 0.340	11.896 ± 0.336	12.046 ± 0.326
SI-CausalTransformer	$\mathbf{10.818 \pm 0.245}$	$\mathbf{11.290 \pm 0.236}$	$\mathbf{11.593 \pm 0.227}$	$\mathbf{11.780 \pm 0.230}$	$\mathbf{11.940 \pm 0.250}$

Although formal statistical significance tests were not conducted, the improvements observed with SI-Mask are consistent across all datasets, models, and time horizons. This uniform performance suggests that the gains are systematic rather than due to random variation. Future work may incorporate statistical testing (e.g., bootstrap-based comparison) to further validate the robustness of these improvements.

6.7 Discussion

This work contributes to both the academic literature and real-world applications by addressing important gaps in the estimation of TE under conditions where traditional assumptions, such as Sequential Ignorability, may not hold. The introduction of the SI-Mask framework represents an incremental advancement that improves the

robustness and accuracy of ITE estimation in time-series data with missing values. By dynamically accounting for missingness-related confounders, the SI-Mask provides a flexible and adaptive approach that facilitates a more accurate estimation of treatment effects. The use of adversarial training to generate individualized masks for each data point is a novel approach that enhances the ability to deconfound data in complex settings.

While the SI-Mask framework demonstrates effectiveness in mitigating biases introduced by observed confounders, it is important to acknowledge certain limitations. The framework assumes Sequential Ignorability, which means that treatment assignments are independent of potential outcomes conditional on observed covariates. However, the framework does not address scenarios where the missingness mechanism depends on unobserved variables, such as Missing Not At Random (MNAR). This limitation could impact the framework's performance in cases involving severe missingness or high-dimensional covariates without sufficient observed data. Additionally, the use of LOCF for imputing missing values, while computationally efficient, may not adequately capture complex temporal dependencies in datasets with intricate missingness patterns. Furthermore, the computational complexity of the adversarial training process, which involves balancing multiple objectives simultaneously, presents challenges for scalability in large-scale applications.

Future research should focus on addressing these limitations and exploring new directions to extend the framework's capabilities. First, validating the SI-Mask framework across diverse datasets and application domains, such as economics and social sciences, will help ensure its robustness and generalizability. Second, integrating SI-Mask with advanced causal inference techniques, such as structural causal models or Bayesian approaches, could enable it to address scenarios involving MNAR or unobserved confounding. Third, exploring alternative imputation methods, such as model-based approaches, deep generative models [86, 63], could enhance its ability to capture non-linear and high-dimensional dependencies in data with complex missingness patterns. Fourth, optimizing the computational efficiency of the adversarial training process through techniques like multi-task optimization or distributed training could make the framework more scalable and applicable to large-scale real-world datasets. Finally, future work could focus on extending the SI-Mask framework to address dynamic treatment regimes and reinforcement learning scenarios, where sequential decision-making plays a critical role. Investigating the potential of the framework to handle real-time data streams and interactive systems could further broaden its applicability across various domains.

Overall, the introduction of the SI-Mask framework represents a significant step forward in the field of causal inference. By addressing real-world challenges associated with time-varying confounding and missing data, the framework provides valuable insights that can improve the reliability of data-driven decisions across various fields. Although promising, this work also opens new avenues for future research, encouraging further exploration and refinement of the proposed methods to better address the complexities of real-world data and provide more reliable causal insights.

6.8 Relevant Publications

- **Ghosheh, G.**, Gogl. M., Zhu, T. (2025). A Perspective on Individualized Treatment Effects Estimation from Time-Series Data. Journal of the American Medical Informatics Association, 2025; https://doi.org/10.1093/jamia/ocae323
- **Ghosheh, G.**, Zhu, T. (2025). Relaxing Sequential Ignorability: A Flexible Framework for Accounting for Missingness-Related Confounding in Time-Series ITE Estimation. KDD 2025 [Under Review].
- **Ghosheh, G. O.**, Li, J., & Zhu, T. (2025). Understanding Missingness in Time-series Electronic Health Records for Individualized Representation. arXiv preprint arXiv:2402.15730. NPJ Artificial Intelligence [Under Review]

Chapter 7

Multi-objective ITE Estimation from Tabular EHRs

Having tackled ITE estimation from time-series data in Chapter 6, we now extend the investigation to multi-objective ITE estimation in tabular EHRs. Unlike prior chapters that focused on single-outcome modeling, this chapter considers the clinical reality where multiple outcomes must be evaluated simultaneously. Without repeating the foundational challenges of ITE from observational data, we introduce a framework that supports nuanced treatment decision-making by modeling interdependent health outcomes concurrently.

7.1 Introduction

In recent years, the widespread adoption of electronic health records (EHRs) has generated a large collection of routinely documented patient data, including clinical context, treatments, and outcomes [81]. This abundance of real-world information allows researchers to go beyond population-level analyses and estimate the *individual treatment effect* (ITE), which predicts the impact of a particular intervention on a specific patient. Such individualized insights are crucial for refining clinical decision-making, improving patient outcomes, and advancing the broader goals of precision medicine.

Despite this promise, using observational EHR data to estimate ITE presents several methodological challenges. In particular, the issue of *confounding* arises when treatment assignments are correlated with patient features [210]. Patients who show similar characteristics in clinical practice frequently receive similar treatments, guided by established protocols or the expertise of the clinicians [26]. Consequently, any differences observed in results across treatment groups cannot be attributed solely to

treatment. The *potential outcome framework* [210], also called the Rubin-Neyman causal model, addresses this challenge by conceptualizing what the outcome for each patient *would have been* under every possible treatment. However, in reality, only one outcome is observed per patient; Unobserved outcomes, known as *counterfactuals*, form the central problem in causal inference [113, 210].

Various statistical and machine learning methods have been proposed to estimate these unobserved counterfactual outcomes from observational data. These include simple regression-based approaches matching methods such as propensity score matching (PSM) [209], and doubly robust estimators that combine outcome regression with propensity score weighting [148]. Recently, deep learning has shown promise in extracting latent representations that can better capture the complex relationships between patient features, treatments, and outcomes [227, 221]. However, many existing deep learning models incorporate treatment merely as a single feature or train separate classifiers for each type of treatment [55]. This limits their ability to exploit the nuanced, treatment-specific interactions inherent in the data.

Despite recent advances in ITE estimation, most existing methods focus on modeling only one outcome of interest [220, 55, 141]. However, in real-world clinical practice, physicians routinely weigh multiple, often interdependent outcomes when formulating treatment plans. For example, an antihypertensive drug may successfully lower blood pressure but simultaneously influence kidney function, or the incidence of cardiovascular events [8]. Ignoring these interconnected effects can lead to incomplete or even misleading conclusions regarding the overall efficacy of a treatment, potentially hindering efforts to achieve truly personalized care.

To address this limitation, we propose **MOITE (Multi-Objective Individual Treatment Effect)**, a novel framework designed to estimate ITE *across multiple clinical outcomes* simultaneously. We apply MOITE to data drawn from the *Clinical Practice Research Datalink* (CPRD), a large-scale primary care database in the United Kingdom that encompasses millions of longitudinal patient records and offers extensive coverage of diverse patient populations [61]. By jointly modeling a network of interdependent endpoints, MOITE seeks to capture the broader spectrum of therapeutic effects. This multi-outcome perspective not only provides more comprehensive insight into treatment benefits and risks, but also enables clinicians to tailor interventions that align with patient-specific priorities and health profiles.

Novelty and Scope: Unlike conventional single-outcome ITE estimation methods, MOITE introduces an integrated approach that accounts for multiple potentially correlated clinical endpoints. This holistic view allows for improved accuracy in ITE

predictions and facilitates more informed treatment decisions, reflecting the real-world complexity where clinicians balance multiple risk factors and potential benefits. Furthermore, by harnessing the rich longitudinal data available in CPRD, our framework is particularly well suited to large-scale population-level analyses, offering robust evidence to support precision medicine strategies. Key contributions of this work include the following:

1. *Multi-Outcome Modeling:* We introduce MOITE, a unified framework that estimates treatment effects for several clinically relevant outcomes concurrently, reflecting the multidimensional nature of patient health.

2. *Extensive Empirical Evaluation in real population:* We rigorously assess MOITE on a large-scale CPRD-derived EHR dataset, demonstrating its ability to outperform single-outcome baseline methods in both predictive accuracy and treatment effect estimation. In addition, we compare the results of the suggested MOITE model with clinical trials in which it demonstrated strong performance in replicating the RCT findings.

3. *Enhanced Clinical Utility:* By highlighting the interplay among multiple outcomes, our approach supports more informed and holistic clinical decision-making, a cornerstone of modern precision medicine.

Taken together, our findings underscore the potential for multi-outcome ITE estimation models to transform the landscape of digital medicine, paving the way for more nuanced and effective personalized interventions.

7.2 Related Work

Individual Treatment Effect Estimation

Estimating the ITE has been a long-standing challenge in statistics and machine learning. Traditional methods, such as propensity score matching [208, 209] and outcome regression [113], have been widely used. These methods aim to balance the covariates between the treatment and control groups or model the outcome as a function of the treatment and covariates. More recently, machine learning techniques have been applied to ITE estimation, showing improved performance.

Tree-based models, such as causal forests [252] and BART [108], have been proposed to estimate heterogeneous treatment effects. These models capture non-linear relationships and complex interactions between covariates and outcomes. Neural

networks have also been utilized for ITE estimation, including counterfactual regression with neural networks and treatment-agnostic representation networks (TARNet) [222, 221]. TARNet serves as a key baseline in our work due to its effectiveness in learning representations for causal inference. Furthermore, domain-relevant baselines such as treatment networks (TNet) and doubly robust networks (DR-Net) [128] are included, as they represent notable advancements in causal machine learning by integrating observational and counterfactual components.

Although these methods often demonstrate superior performance compared to traditional techniques, they mainly focus on single-outcome estimation and treat each outcome independently. In reality, clinical outcomes are often interconnected and influence each other. For example, in the context of hypertension treatment, improvements in blood pressure can impact cardiovascular health and vice versa. Therefore, considering the interdependence of multiple outcomes is crucial for more accurate and comprehensive ITE estimation.

In this work, we build upon these ideas and propose a novel multi-outcome ITE estimation model, MOITE, specifically designed for tabular EHR data. MOITE integrates ideas from representation learning (TARNet and DR-Net) and multi-task architectures to capture the interdependence of clinical outcomes. Unlike existing baselines, MOITE simultaneously estimates treatment effects for multiple outcomes, sharing information across outcomes to improve prediction accuracy. By addressing the complexities of primary care data in the real world, MOITE provides personalized treatment effect estimates that align with clinical decision-making and account for the interconnected nature of outcomes in observational healthcare datasets.

ITE vs Supervised Learning Models

ITE models fundamentally differ from traditional supervised learning models due to their primary focus on causal inference rather than outcome prediction. In normal supervised learning, the objective is to learn a mapping between input features (X) and a single observed outcome (Y) by minimizing the prediction error, under the assumption that the data are independently and identically distributed (i.i.d.). These models are concerned only with accurately predicting Y based on observed data, without considering the effects of the intervention or hypothetical scenarios.

In contrast, ITE models aim to estimate the causal effect of a treatment by predicting the difference in potential outcomes for an individual in two counterfactual scenarios: one where the treatment is applied ($Y(1)$) and another where it is not ($Y(0)$). This requires ITE models to infer both the *factual outcome* (observed in the

data) and the *counterfactual outcome* (unobserved and hypothetical). The treatment effect is quantified as:

$$\text{ITE} = Y(1) - Y(0),$$

which represents the effect at the individual level of the treatment.

To achieve this, ITE models explicitly incorporate the treatment assignment indicator (T) as an input feature and adjust for confounding variables to address biases introduced by non-random treatment allocation. These models estimate the conditional probabilities $P(Y, X, T)$ for the treated and untreated groups, allowing the estimation of causal effects. Unlike normal supervised learning, ITE models leverage architectures designed to balance treated and control groups, such as combining shared feature representations with task-specific layers to model both $Y(1)$ and $Y(0)$ accurately.

By modeling counterfactuals and addressing confounding, ITE models are well-suited for causal inference tasks. In contrast, traditional supervised learning models lack the framework to account for treatment bias or estimate causal effects, rendering them insufficient for applications requiring intervention evaluation or policy optimization.

Multi-Task Learning

Our work is closely related to the field of multi-task learning, where multiple tasks are learned together utilizing shared information and representations [32]. By learning tasks in parallel, multi-task models can benefit from improved generalization performance, reduced overfitting, and enhanced sample efficiency [283]. In the context of ITE estimation, multi-task learning can be particularly advantageous due to the inherent relationships and dependencies that exist between different outcomes or tasks. In our proposed model, MOITE, we adopt a multi-task learning framework to estimate the ITE for multiple outcomes simultaneously. MOITE consists of shared representation layers that capture general patterns across outcomes, as well as outcome-specific layers that learn to predict the ITE for each specific outcome. By sharing information and representations across outcomes, MOITE aims to improve the accuracy and practicality of ITE estimation in real-world clinical settings.

7.3 Methodology

7.3.1 Proposed Model

We propose a multi-outcome ITE estimation model, named MOITE, designed to capture the interdependence of clinical outcomes. MOITE is a multi-task learning framework that consists of shared and outcome-specific layers. The shared layers capture general representations from the input features, while the outcome-specific layers learn to predict the ITE for each specific outcome. By sharing information across outcomes, MOITE can leverage the relationships between them and improve estimation performance.

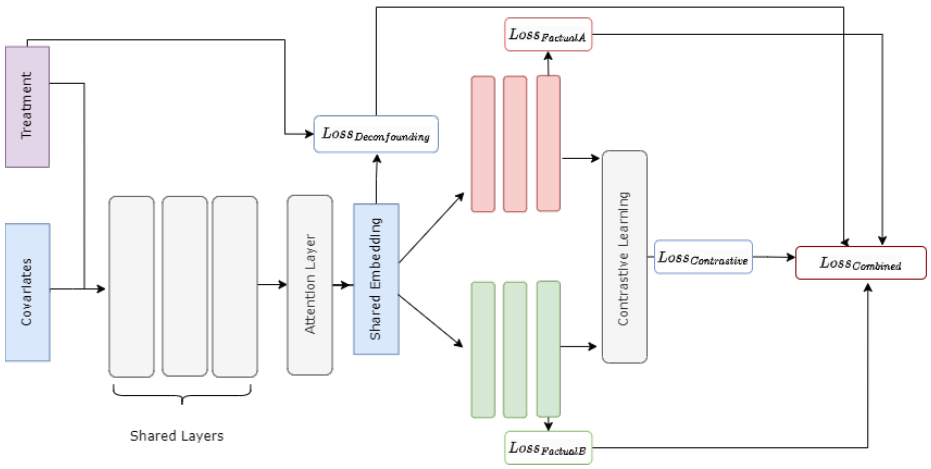


Figure 7.1: The overview figure illustrates the multi-objective training framework for Individualized Treatment Effect estimation. The model processes input features $X \in \mathbb{R}^{N \times D}$ through shared layers to learn general representations, while treatment assignments $A \in \mathbb{R}^{N \times 1}$ direct the flow into outcome-specific layers. These layers specialize in predicting individualized outcomes (e.g., Falls, Acute Kidney Injury (AKI)) and corresponding loss terms ($L_{\text{Factual A}}, L_{\text{Factual B}}$). A combined multi-task loss (L_{Combined}) ensures unified optimization across all outcomes, incorporating deconfounding contrastive losses to balance treatment and control group representations and align feature distributions. Through backpropagation, shared and outcome-specific parameters are iteratively refined, enhancing individualized predictions and treatment effect estimates. This framework facilitates personalized care by generating data-driven predictions tailored to specific treatment scenarios.

Mathematical Framework

Problem Definition

Given a dataset $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^d$ represents the covariates for the i-th individual, $A_i \in \{0, 1\}$ denotes the assignment of binary treatment and $Y_i = [Y_i^1, Y_i^2, \ldots, Y_i^M] \in \mathbb{R}^M$ corresponds to M observed outcomes, the objective of the Multi-Outcome Individual Treatment Effects (MOITE) model is to estimate the individualized treatment effect for each outcome $m \in \{1, \ldots, M\}$:

$$\text{ITE}_i^m = \mathbb{E}[Y_i^m \mid A_i = 1, X_i] - \mathbb{E}[Y_i^m \mid A_i = 0, X_i],$$

where $\mathbb{E}[Y_i^m \mid A_i, X_i]$ represents the potential outcome under treatment A_i for the m-th outcome.

Assumptions

When working with real-world data to estimate treatment effects, we face a challenge: we can only observe what actually happened, not what would have happened if a different treatment was given. The potential outcomes framework [210] helps address this, but for it to work, we need to make certain assumptions. These assumptions ensure that we can compare treated and untreated individuals in a way that provides meaningful insights.

For each individual i, we observe their characteristics $\mathbf{x}_i$, a treatment assignment $A_i \in \{0, 1\}$ (where 1 means that they received the treatment and 0 means they did not), and an outcome Y_i. However, we only see one version of their outcome, either the treated or untreated outcome, not both. To estimate what the effect of the treatment would have been, we rely on three key assumptions:

1. **Consistency (Each Person Has a Well-Defined Outcome).**

   This assumption means that if a person actually received a treatment, their observed outcome is the same as the potential outcome under that treatment. In other words:

   $$Y_i = Y_i(a) \quad \text{if they actually received } A_i = a. \tag{7.1}$$

   Simply put, if someone took a blood pressure medication, the recorded effect on their blood pressure should be exactly what the drug would have done for them, no unexpected variations due to external influences.

2. **Overlap (Everyone Had a Fair Chance of Getting Either Treatment).**

   This ensures that for every type of individual in the dataset, there was at least some chance of receiving the treatment and some chance of not receiving it. This is written as:

$$0 < P(A_i = 1 \mid \mathbf{x}_i) < 1. \tag{7.2}$$

   This means that we do not have groups of people who always receive treatment or who always do not receive treatment according to their characteristics. If we did, we would not be able to compare the outcomes between treated and untreated groups fairly. For example, suppose that we are studying the effects of a cholesterol drug and that all patients over 60 automatically receive it while younger patients never do. This would violate the overlap assumption because there would be no younger people in the treated group and no older people in the untreated group, meaning that we could not estimate what treatment would have done for the younger group.

3. **Ignorability (No Hidden Factors Driving Both Treatment and Outcome).**

   This assumption says that once we account for all observed characteristics $\mathbf{x}_i$, treatment assignment is like a coin flip: It does not depend on hidden factors that also affect the outcome:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp A_i \mid \mathbf{x}_i. \tag{7.3}$$

   In simple terms, there are no unmeasured variables that influence both who is treated and what their outcome will be. If this assumption is violated, it means that some external factor is skewing the treatment assignment, making the treated and untreated groups fundamentally different in ways that we cannot see in the data. Suppose that we are studying how an exercise program affects weight loss, but the dataset does not include motivation levels. If more motivated people are both more likely to enroll in the program and more likely to lose weight regardless of the program, we could falsely attribute weight loss to the program when it was actually due to motivation.

These three assumptions, *Consistency*, *Overlap*, and *Ignorability*, form the basis for estimating treatment effects using observational data. If any of them are violated, we risk drawing incorrect conclusions about how treatments work in the real world. These assumptions—consistency, positivity, and sequential ignorability—are closely related to those introduced in Chapter 6. While Chapter 6 focused on the assumptions required for potential outcomes in the context of temporal missingness, here we frame them in the setting of counterfactual treatment assignment under structured interventions. The core ideas remain the same (e.g., no unmeasured confounding and sufficient variability in treatment), but their operationalization differs due to model structure and setting. While these assumptions are fundamental to causal inference, they are generally untestable directly in real-world observational data. In practice, we rely on domain knowledge, proxy variables, and diagnostic tools (e.g., covariate balance or overlap checks) to evaluate plausibility.

7.3.2 Model Architecture and Motivation

The proposed MOITE model represents a significant advancement in the estimation of individualized treatment effects when multiple interrelated outcomes are of interest. In many real-world scenarios, particularly in healthcare, patients often exhibit a spectrum of comorbidities or outcomes that do not manifest in isolation. For example, a patient with hypertension might also have an increased risk of diabetes or cardiovascular events. Traditional ITE methods that focus on one outcome at a time overlook the shared risk factors and complex relationships across these outcomes, leading to suboptimal predictive performance and weaker generalization.

The overarching goal of treatment effect estimation is to predict how an intervention will affect a specific patient. However, in clinical practice, the success of an intervention is rarely measured by a single endpoint; instead, clinicians and policy makers consider multiple indicators of patient well-being. Learning individualized treatment effects across multiple outcomes allows practitioners to make more informed decisions, balancing therapeutic benefits against potential risks or side effects. By combining knowledge across related outcomes, MOITE provides a holistic view of the patient's likely response to treatment. To that end, the MOITE model employs a multi-task learning framework enhanced with *attention mechanisms*, *contrastive learning*, and *domain adaptation* strategies such as Maximum Mean Discrepancy (MMD). This synergy of techniques allows for the concurrent estimation of treatment effects on several binary outcomes, ensuring that latent representations are robust, balanced, and reflective of shared patterns among outcomes.

The MOITE architecture consists of three primary components: (1) a shared representation network that processes patient covariates and treatment assignment to learn a common latent space, (2) an attention mechanism that adaptively weights the importance of features based on their relevance, and (3) outcome-specific branches that focus on estimating potential outcomes under different treatment assignments. Furthermore, the model is equipped with loss terms to balance factual predictions, encourage alignment of treated and control distributions, and enforce consistency between different tasks.

Shared Representation Network: The shared representation network learns a common feature space across all outcomes by processing both patient covariates X_i (which may include demographic, clinical, or genetic information) and the treatment assignment A_i. This shared representation $\Phi(X_i, A_i)$ is modeled using a deep neural network with multiple fully connected layers, batch normalization, and dropout, parameterized by θ_{shared}:

$$\Phi(X_i, A_i) = f_{\text{shared}}(X_i, A_i; \theta_{\text{shared}}). \tag{7.4}$$

By pooling knowledge across tasks, the shared network captures global patterns relevant to multiple outcomes. This helps improve data efficiency, especially in smaller datasets, and reduces overfitting by forcing the model to focus on signals that generalize across multiple endpoints.

Attention Layer: Although the shared representation network captures useful features, it may also include irrelevant or noisy information. To focus on the most important features, an attention mechanism is applied to refine the shared representation. This mechanism assigns different importance weights to different dimensions of the learned representation.

The attention layer operates as follows:

$$\alpha(X_i, A_i)\& = \text{softmax}\big(W_2 \tanh(W_1 \Phi(X_i, A_i))\big), \tilde{\Phi}(X_i, A_i)\& = \alpha(X_i, A_i) \odot \Phi(X_i, A_i). \tag{7.5}$$

Here, W_1 and W_2 are learnable weight matrices that transform the shared representation, while the function $\tanh(\cdot)$ introduces non-linearity to enhance feature interactions. The *softmax* operation ensures that the attention scores $\alpha(X_i, A_i)$ sum to 1 across feature dimensions, making them interpretable as probabilities. Finally,

the re-weighted representation $\tilde{\Phi}(X_i, A_i)$ is obtained by applying the attention scores element-wise ($\odot$) to the original shared representation, thereby emphasizing the most relevant features while suppressing less informative ones.

Outcome-Specific Task Branches: After obtaining the attention-refined representation $\tilde{\Phi}$, each outcome $m \in 1, \ldots, M$ is modeled using a separate neural network head. These outcome-specific branches, parameterized by $\theta^m_{\text{outcome}}$, predict the potential outcome under both treatment ($A = 1$) and control ($A = 0$):

$$\hat{Y}i^m(A) = f\text{outcome}^m(\tilde{\Phi}(X_i, A_i), A; \theta^m_{\text{outcome}}). \tag{7.6}$$

This design decouples the final stage of prediction for each outcome, allowing unique behaviors of individual outcomes to be captured while still leveraging the shared knowledge contained in $\tilde{\Phi}$. As a result, the model can better handle potential heterogeneity in outcome response, reflecting the reality that certain treatments may benefit one outcome more than another.

Loss Function and Regularization

The MOITE model's training objective encourages accurate factual prediction, balanced treatment assignment in the latent space, and consistency across multiple outcomes. The total loss is composed of three terms, described below.

1. Factual Prediction Loss :

Because each outcome is binary, we employ the binary cross-entropy (BCE) loss to align predictions with observed outcomes:

$$\mathcal{L}_{\text{factual}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\left(Y_i^m \log(\hat{Y}_i^m) + (1 - Y_i^m)\log(1 - \hat{Y}_i^m)\right) \tag{7.7}$$

In this equation, N refers to the total number of patients, and M denotes the number of binary outcomes (tasks). The variable Y_i^m is the observed value of outcome m for patient i, while $\hat{Y}_i^m$ represents the predicted probability of that outcome under the treatment actually received. Minimizing $\mathcal{L}_{\text{factual}}$ ensures the model is accurately predicting each observed outcome across all tasks.

2. Counterfactual Regularization (MMD) :

To mitigate bias introduced by non-random treatment assignment, we incorporate a regularization term based on Maximum Mean Discrepancy (MMD). This penalizes differences between the distributions of latent representations in the treated and untreated groups:

$$\mathcal{L}_{\text{MMD}} = \left\| \mathbb{E}_{T_i=1}[\tilde{\Phi}(X_i)] - \mathbb{E}_{T_i=0}[\tilde{\Phi}(X_i)] \right\|^2 \tag{7.8}$$

Here, $\tilde{\Phi}(X_i)$ denotes the attention-weighted latent representation of patient i, and T_i is the treatment assignment indicator (1 for treated, 0 for control). The operator $\mathbb{E}_{T_i=a}[\cdot]$ denotes the average (expectation) computed over all patients who received treatment a. This loss encourages the model to learn a shared space where treated and control groups are balanced, supporting more reliable counterfactual predictions.

3. Multi-Task Contrastive Loss :

To ensure consistency between outcome-specific embeddings, the model also minimizes a contrastive loss across all outcome heads:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{n=m+1}^{M} \|\Phi^m(X_i, T_i) - \Phi^n(X_i, T_i)\|^2 \tag{7.9}$$

In this formulation, $\Phi^m(X_i, T_i)$ is the representation of patient i derived from the outcome-specific branch for task m, and $\Phi^n(X_i, T_i)$ is the corresponding embedding for task n. The summation iterates over all unique pairs of outcomes. This loss penalizes large differences between outcome-specific embeddings for the same patient, promoting consistency and alignment in the latent space while still allowing task-specific nuances.

Overall Loss Function :

The total loss is the weighted sum of the three components:

$$\mathcal{L} = \mathcal{L}_{\text{factual}} + \alpha \mathcal{L}_{\text{MMD}} + \beta \mathcal{L}_{\text{contrast}} \tag{7.10}$$

Here, the hyperparameter α controls the influence of the MMD regularization term, while β modulates the strength of the contrastive loss. These weights allow practitioners to adjust the emphasis between factual prediction accuracy, treatment group balancing, and multi-task coherence. The model is trained by minimizing this composite loss using the Adam optimizer [285].

Algorithm 4 Multi-Objective ITE Estimation Framework for Tabular EHRs

Require: Input data $\mathcal{D} = \{X, A, Y_1, \ldots, Y_M\}$, where X are covariates, A is treatment assignment, and Y_m are outcome labels

Ensure: Trained ITE model f_{ITE}, predicted outcomes $\hat{Y}_m$, estimated ITEs $\hat{\tau}_m$

Preprocessing:
    Normalize covariates X
    Encode treatments A and outcomes $Y_1, \ldots, Y_M$
    Split data into $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{test}}$
Model Initialization:
    Define shared feature encoder f_{shared}
    For each outcome $m = 1, \ldots, M$, define outcome-specific head f_m
    Initialize parameters $\theta_{\text{shared}}, \theta_1, \ldots, \theta_M$
    while *not converged* **do**
        Sample minibatch $(X_{\text{batch}}, A_{\text{batch}}, Y^1_{\text{batch}}, \ldots, Y^M_{\text{batch}})$
        Compute shared representation: $H \leftarrow f_{\text{shared}}(X_{\text{batch}})$ **for** $m = 1$ *to* M **do**
            Compute predictions: $\hat{Y}_m \leftarrow f_m(H, A_{\text{batch}})$
        Compute task loss $\mathcal{L}_m$ (e.g., cross-entropy or MSE)
        Compute total loss: $\mathcal{L}_{\text{total}} \leftarrow \sum_{m=1}^{M} \mathcal{L}_m + \lambda \cdot \mathcal{L}_{\text{contrastive}}$
        Update parameters $\theta_{\text{shared}}, \theta_1, \ldots, \theta_M$ via backpropagation
Evaluation:
    Predict $\hat{Y}^0_m$ and $\hat{Y}^1_m$ for each patient under control and treatment
    Compute ITE: $\hat{\tau}_m \leftarrow \hat{Y}^1_m - \hat{Y}^0_m$
    Evaluate performance using AUROC, AUPRC, risk ratios, etc.
    return f_{ITE}, estimated ITEs

Implementation

The code was implemented using Python version 3.7 and PyTorch version 1.9. We use an exponential decay learning rate schedule for the Adam optimizers. Training proceeds for 100 epochs with a batch size of 32. The combination weight α is initialized to 0.5 and dynamically updated during training.

7.3.3 Baselines

7.3.4 Dataset Description

The STRATIFY study (STRAtifying Treatments in Multimorbid Frail Elderly) is an initiative that leverages the Clinical Practice Research Datalink (CPRD) to investigate the effectiveness and safety of treatments in older adults with multiple comorbidities [61]. In this work, we specifically focus on the antihypertensive subset of STRATIFY, which includes patients aged 40 years and older with at least one systolic blood pressure reading of 130 mmHg or higher, indicating the potential need for antihypertensive therapy. The dataset comprises patient demographics (e.g., age, sex, ethnicity), clinical measurements (e.g., blood pressure, body mass index), laboratory results (e.g., cholesterol levels), prescribed antihypertensives, and relevant clinical outcomes (e.g., acute kidney injury, falls, fractures, gout). The index date for each patient is defined as 12 months after their first qualifying blood pressure measurement and the outcomes are tracked for up to 10 years following this index date. The records span from 1 January 1998 to 31 December 2018 and include only data from "up-to-standard" CPRD practices, ensuring high-quality data for analyzing how antihypertensive treatments affect frail and multimorbid older patients in real-world settings. We show the characteristics of the patients across the treatment and control groups in section 3.7.

We compare the performance of MOITE with several state-of-the-art baselines implemented in [55] for the ITE estimation, all of which are for single outcome estimation:

- **T-Learner**: TNet applies the T-Learner strategy by training two independent regression models: one for the treated group ($\mu_1(x)$) and another for the untreated group ($\mu_0(x)$). The Conditional Average Treatment Effect (CATE) is calculated as the difference between the predicted outcomes of these models, $\tau(x) = \mu_1(x) - \mu_0(x)$. This approach is straightforward and effective when

there is significant divergence between the treated and untreated outcome distributions. However, it does not share information between the models, which can lead to inefficiency when the treated and untreated groups exhibit common patterns.

- **TARNet**: TARNet improves the S-Learner approach by learning a shared representation $\Phi(X)$ for covariates, which is used by two separate heads to predict the outcomes for the treated ($\mu_1(x)$) and untreated ($\mu_0(x)$) groups. Using a shared feature space, TARNet captures common patterns across treatment groups, improving efficiency and performance in scenarios with significant overlap or small sample sizes. However, it may struggle when the outcome functions for treated and untreated groups are substantially different, as the shared representation might not fully capture their distinct characteristics.

- **DR-Learner**: The DR-Learner combines regression adjustment and inverse propensity weighting to ensure robustness, making it accurate if either the outcome or propensity model is correctly specified. It estimates CATE in two steps: first, by estimating propensity scores ($\pi(x)$) and potential outcomes ($\mu_1(x)$ and $\mu_0(x)$); and second, by computing a pseudo-outcome using a doubly robust formula and regressing it on covariates. This method is theoretically optimal in large samples and is resilient to certain model misspecifications, but can be sensitive to propensity score errors, particularly in smaller datasets.

7.3.5 Evaluation

The evaluation of the models was conducted in two parts, designed to provide a comprehensive understanding of their performance in predicting outcomes and estimating treatment effects.

Discriminative Performance in Observed Outcomes

To evaluate predictive capabilities, we employ a novel approach that separates potential outcomes based on factual observations and treatment assignments. Specifically, the predicted potential outcomes of the model for untreated and treated cases (po_0 and po_1, respectively) were compared with the factual outcomes observed in the dataset. For each observation, the factual prediction was calculated as a combination of po_0 or po_1, determined by assigning the observed treatment. This separation ensures that the evaluation captures the alignment between the model's predictions and the

observed realities in the given treatment context, providing a treatment-specific evaluation of the accuracy of the prediction. The AUROC was calculated to measure the ability of the model to distinguish between positive and negative cases within each treatment group. Additionally, the Area Under the Precision-Recall Curve (AUPRC) was calculated to evaluate performance on imbalanced datasets, offering a comprehensive perspective on the model's discrimination and calibration capabilities. This evaluation represents a significant advancement, being the first to compute AUROC and AUPRC in this treatment-specific manner, tailored for observational datasets with imbalanced outcomes.

Comparison with Randomized Controlled Trial (RCT) Findings

In addition to evaluating predictive performance, the models were assessed for their ability to estimate treatment effects. This evaluation focused on comparing the average treatment effects (ATE) derived from the models with findings from established Randomized Controlled Trials (RCTs). Using benchmarks against RCT results, the evaluation validates the causal inference capabilities of the models, emphasizing their potential to provide insights aligned with clinical evidence. This step ensures that the models not only accurately predict outcomes, but also generate reliable estimates of treatment effects critical to decision-making in personalized medicine.

In general, these dual evaluation strategies highlight the robustness of the proposed methodology, which bridges predictive analytics and causal inference to address complex challenges in healthcare research. This approach not only evaluates the performance of the model in observed outcomes, but also validates its potential to generate actionable insights aligned with clinical evidence.

Estimating Risk Ratios for Individual Treatment Effects

In the potential outcomes framework, the treatment effect for an individual i is defined as

$$\beta(i) = y_1(i) - y_0(i),$$

where $y_1(i)$ and $y_0(i)$ represent the potential outcomes under treatment and control, respectively. Because treatment effects are generally heterogeneous across individuals, the *average treatment effect* (ATE) is defined as

$$\text{ATE} = E[y_1 - y_0]. \tag{7.11}$$

Under the law of large numbers, the ATE can be estimated by averaging the individual treatment effects on all N individuals:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_1(i) - y_0(i)\right).$$

For binary outcomes, where potential outcomes correspond to probabilities, we denote *baseline risk*, p_c, as the risk (or probability of events) in the control group. Then the average outcome under treatment is given by

$$p_t = p_c + \text{ATE}.$$

The risk ratio (RR), a relative measure of the treatment effect, is defined as

$$\text{RR} = \frac{p_t}{p_c}.$$

Substituting $p_t = p_c + \text{ATE}$ into the equation, we obtain

$$\text{RR} = \frac{p_c + \text{ATE}}{p_c} = 1 + \frac{\text{ATE}}{p_c}.$$

This conversion allows us to compare the performance of interventions on a relative scale, facilitating comparisons across randomized controlled trials and meta-analyses. This approach of linking individual-level potential outcomes with population-level effect measures is well-established [192, 107].

7.4 Results

7.4.1 Comparison with ITE estimators

To assess the performance of MOITE in treatment effect estimation, we compared it with established ITE models (TARNet, TNet, and DR-Net) across multiple adverse outcomes. The results show that MOITE outperforms other estimators in AUROC and AUPRC scores, indicating superior discrimination across a range of outcomes as shown in Table 7.1. MOITE exhibited the highest AUROC scores across all outcomes, particularly for Falls (0.794, 95% CI: 0.793–0.795), AKI (0.748, 95% CI: 0.746–0.752), and Gout (0.803, 95% CI: 0.803–0.805). Compared to DR-Net, TARNet, and TNet, MOITE demonstrated more consistent predictive performance across heterogeneous clinical conditions, supporting the advantage of a multi-task learning framework for modeling interrelated treatment effects. The AUPRC scores further

reinforce MOITE's superior performance, particularly in outcomes with class imbalance. MOITE achieved an AUPRC of 0.031 for AKI, surpassing 0.020 for DR-Net and 0.018 for TARNet, and an AUPRC of 0.057 for Hypotension, outperforming DR-Net with an AUPRC of 0.040 and TARNet with 0.038. These results suggest that MOITE effectively captures meaningful predictive signals even in conditions with a low prevalence of positive cases—a critical requirement for treatment effect modeling in real-world clinical settings.

Table 7.1: Performance Metrics Across Models and Outcomes with Confidence Intervals

Model	Falls	AKI	Fracture	Hypotension	Syncope	Gout	Elec Sensitivity
AUROC (95% CI)							
TNet	0.782 (0.780, 0.784)	0.671 (0.667, 0.675)	0.650 (0.649, 0.652)	0.740 (0.737, 0.743)	0.616 (0.613, 0.620)	0.767 (0.765, 0.769)	0.698 (0.696, 0.701)
TARNet	0.782 (0.781, 0.784)	0.660 (0.656, 0.664)	0.649 (0.647, 0.650)	0.737 (0.734, 0.740)	0.612 (0.608, 0.615)	0.767 (0.765, 0.769)	0.694 (0.691, 0.696)
DR-Net	0.783 (0.782, 0.785)	0.692 (0.688, 0.696)	0.648 (0.647, 0.650)	0.749 (0.746, 0.752)	0.617 (0.614, 0.621)	0.770 (0.767, 0.771)	0.713 (0.710, 0.715)
MOITE	**0.794 (0.793, 0.795)**	**0.748 (0.746, 0.752)**	**0.675 (0.673, 0.675)**	**0.789 (0.786, 0.790)**	**0.677 (0.676, 0.678)**	**0.803 (0.803, 0.805)**	**0.762 (0.761, 0.763)**
AUPRC (95% CI)							
TNet	0.200 (0.198, 0.202)	0.019 (0.019, 0.020)	0.174 (0.172, 0.176)	0.039 (0.039, 0.041)	0.029 (0.028, 0.030)	0.225 (0.221, 0.229)	0.183 (0.180, 0.185)
TARNet	0.201 (0.199, 0.203)	0.018 (0.018, 0.019)	0.173 (0.171, 0.175)	0.038 (0.038, 0.039)	0.029 (0.028, 0.030)	0.226 (0.222, 0.229)	0.184 (0.181, 0.186)
DR-Net	0.202 (0.200, 0.204)	0.020 (0.019, 0.020)	0.172 (0.171, 0.174)	0.040 (0.040, 0.041)	0.030 (0.029, 0.031)	0.227 (0.223, 0.230)	**0.185 (0.182, 0.187)**
MOITE	**0.218 (0.216, 0.219)**	**0.031 (0.030, 0.032)**	**0.189 (0.188, 0.191)**	**0.057 (0.056, 0.058)**	**0.039 (0.038, 0.040)**	**0.248 (0.245, 0.251)**	0.093 (0.092, 0.093)

In contrast, single-task models such as TARNet and TNet exhibited higher variability across outcomes. TARNet demonstrated competitive performance for Hypotension AUROC of 0.749 but performed suboptimally for AKI AUROC of 0.692, suggesting limitations in capturing complex dependencies between outcomes. Similarly, TNet performed poorly on the AKI prediction AUROC of 0.671 and Syncope AUROC of 0.616, highlighting difficulties in modeling the risk of adverse events during treatment. DR-Net showed improved generalization in some tasks, but exhibited inconsistencies for Syncope AUROC of 0.617 and Fracture AUROC of 0.648, reflecting sensitivity to dataset variations. The comparative evaluation highlights the robustness and stability of MOITE's multi-task learning approach, which effectively integrates shared representations across clinically related outcomes. This enables more reliable treatment effect estimation across multiple adverse events while reducing variance and improving generalizability. These findings support the adoption of multi-task architectures for real-world evidence generation, particularly in observational healthcare datasets where treatment effects are inherently heterogeneous.

Although statistical significance tests were not explicitly conducted, the improvements shown in Table 7.1 are consistent across multiple evaluation settings. The stability of performance across baselines suggests that the observed differences are unlikely to be due to random chance.

7.4.2 Ablation Study

To assess the impact of different components in our MOITE model, we conducted an ablation study that evaluated multiple variants across adverse outcomes. The full model (MOITE Combined, Attention, MMD, Contrastive) achieved the highest AUROC across all tasks, demonstrating the stability and robustness of a multi-task learning approach with integrated contrastive learning and domain alignment techniques.

Table 7.2 presents the AUROC scores for each variant of the model. The MOITE (Combined, Attention, MMD, Contrastive) consistently outperformed other versions, with improvements observed in AKI of 0.748 compared to 0.708 without attention, Hypotension of 0.789 compared to 0.758, and Syncope 0.675 compared to 0.633, highlighting the contribution of attention mechanisms and contrastive learning. The MOITE (Single-task) variant showed the lowest performance in most outcomes, reinforcing the importance of multi-task representation learning. Although some models performed similarly in select outcomes, their high variance across tasks suggests that MOITE's multi-task learning approach provides superior generalization and stability. The results confirm that the integration of attention, MMD, and contrastive learning enhances the shared representation space, leading to more reliable estimations of the effects of treatment across multiple adverse events.

Table 7.2: Performance when ablating various components of the proposed models, reported in terms of AUROC

Ablations	Falls	AKI	Fracture	Hypotension	Syncope	Gout	Electrolyte Sensitivity
MOITE (Combined, no Attention)	0.785	0.708	0.652	0.758	0.633	0.770	0.720
MOITE (Combined, Attention)	0.795	0.747	0.677	0.788	0.675	0.804	0.762
MOITE (Combined, Attention, MMD, Contrastive)	**0.795**	**0.748**	**0.677**	**0.789**	**0.675**	**0.804**	**0.762**
MOITE (Single)	0.792	0.726	0.671	0.783	0.660	0.801	0.758

7.4.3 Comparison With RCT Findings

To understand how our model's predictions compare with real-world clinical evidence, we examined findings from a large meta-analysis of 58 randomized controlled trials (RCTs) involving 280,638 participants [8]. This analysis assessed the risks associated with antihypertensive treatment across a range of adverse events, including acute kidney injury (AKI), electrolyte imbalances, hypotension, syncope, falls, fractures, and gout. The RCT findings reported increased risks for AKI (RR: 1.18, 95% CI: 1.01–1.39; 15 studies), Hypokalaemia (RR: 1.54, 95% CI: 0.63–3.75; 12 studies), Hypotension (RR: 1.97, 95% CI: 1.67–2.32; 35 studies), and Syncope (RR: 1.28, 95% CI:

1.03–1.59; 16 studies). Meanwhile, Falls (RR: 1.05, 95% CI: 0.89–1.24; 7 studies) and Fractures (RR: 0.93, 95% CI: 0.58–1.48; 5 studies) were not significantly associated with treatment, while evidence for Gout (RR: 3.84, 95% CI: 0.95–15.57; 5 studies) was inconclusive due to high variability.

Importantly, the degree of variation across studies—referred to as heterogeneity—was assessed using two standard measures: I^2 and τ^2, [8]. I^2 quantifies the percentage of total variation in effect estimates that is due to heterogeneity rather than chance, while τ^2 estimates the actual between-study variance in effect sizes. For example, hypotension and gout showed high heterogeneity (I^2 = 85.1% and 84.3%, respectively), indicating substantial inconsistency in effect sizes across studies—likely reflecting differences in study populations, drug classes, or definitions of adverse events. These heterogeneity estimates are critical for interpreting the reliability and generalizability of meta-analytic results. The complete results are shown in Table 7.3.

Table 7.3: Meta-analysis Summary Table. I^2 (%) refers to the percentage of variation across studies due to heterogeneity rather than chance. τ^2 represents the variance in effect sizes across studies [8].

Outcome	**No. of Studies**	**No of Patients**	**RR**	**I^2 (%)**	**τ^2**
Falls	7	29,481	1.05 (0.89 to 1.24)	31.5	0.009
AKI	15	95,600	1.18 (1.01 to 1.39)	48.1	0.037
Fractures	5	12,913	0.93 (0.58 to 1.48)	53.8	0.062
Hypotension	35	182,122	1.97 (1.67 to 2.32)	85.1	0.132
Syncope	16	102,261	1.28 (1.03 to 1.59)	42.9	0.050
Gout	5	32,661	3.84 (0.95 to 15.57)	84.3	1.374
Hypokalaemia	12	39,376	1.54 (0.63 to 3.75)	94.3	1.612

To evaluate how well our machine learning models align with these findings, we analyzed real-world observational data using TNet, TARNet, DR-Net, and MOITE to estimate the average treatment effect (ATE) for each adverse outcome. Since ATE values range from –1 to 1, where negative values indicate a reduced risk and positive values indicate increased risk, we converted them into risk ratios (RRs) for direct comparison with the RCT estimates as shown in Table 7.4.

In all models, MOITE exhibited a relatively close alignment with the RCT-reported risk ratios, with more stable and consistent performance across the outcomes. For example, MOITE estimated an RR of 1.27 (95% CI: 1.01-1.53) for AKI, which aligns closely with the RCT estimate and falls within its confidence interval. For Hypotension, although none of the models matched the RCT estimate of 1.97, MOITE provided one of the most conservative but consistent estimates (RR: 1.28, 95% CI: 0.97–1.60). Similarly, MOITE's RR for Syncope was 1.23 (95% CI: 1.07–1.39), within the RCT interval and comparable to other models. For Electrolyte Sensitivity, used

Table 7.4: Comparison of ATE-derived Risk Ratios (RR) across models and RCT values for various outcomes. * The clinical evidence reported the results for hypokalaemia, which is considered the closest to electrolyte sensitivity. The RCT confidence intervals are computed from 7, 15, 5, 35, 16, 5, and 12 studies for Falls, AKI, Fracture, Hypotension, Syncope, Gout, and Hypokalaemia, respectively.

Model	Falls	AKI	Fracture	Hypotension	Syncope	Gout	Electrolyte Sensitivity
TNet	1.21 (0.91, 1.51)	1.88 (1.65, 2.11)	1.02 (0.85, 1.19)	1.64 (0.95, 2.33)	1.28 (1.09, 1.47)	1.31 (0.63, 1.99)	1.78 (1.20, 2.36)
TARNet	1.12 (0.96, 1.28)	1.76 (1.68, 1.84)	0.99 (0.88, 1.11)	1.52 (1.23, 1.80)	1.38 (1.28, 1.48)	1.52 (0.83, 2.20)	2.12 (1.68, 2.56)
DR-Net	1.10 (0.78, 1.42)	1.34 (0.87, 1.81)	0.97 (0.85, 1.10)	1.41 (0.63, 2.20)	1.49 (1.21, 1.76)	1.44 (0.75, 2.14)	1.33 (0.97, 1.69)
MOITE	1.13 (0.96, 1.29)	1.27 (1.01, 1.53)	1.14 (1.03, 1.25)	1.28 (0.97, 1.60)	1.23 (1.07, 1.39)	1.21 (0.97, 1.46)	1.31 (1.03, 1.59)
RCT	1.05 (0.89, 1.24)	1.18 (1.01, 1.39)	0.93 (0.58, 1.48)	1.97 (1.67, 2.32)	1.10 (1.03, 1.59)	3.84 (0.95, 15.57)	1.54 (0.63, 3.75)*

here as a proxy for Hypokalaemia, MOITE's estimate (RR: 1.31, 95% CI: 1.03–1.59) was slightly lower than the effect size of the pooled RCT but within a clinically reasonable range. However, a notable observation was that MOITE's behavior diverged somewhat for Falls and Fractures. Although its estimates for these outcomes (Falls: RR 1.13, 95% CI: 0.96–1.29; Fractures: RR 1.14, 95% CI: 1.01–1.53) remained within plausible bounds, they were slightly elevated compared to the pooled estimates of the RCT. This may be attributed, in part, to differences in heterogeneity across outcomes.

In the meta-analysis, Falls and Fractures exhibited relatively low between-study heterogeneity (I^2: 31.5% and τ^2: 0.009, I^2: 53.8% and τ^2: 0.062, respectively), indicating more consistent findings across trials [8]. In contrast, outcomes such as Hypotension (I^2: 85.1%, τ^2: 0.132) showed much higher heterogeneity, suggesting substantial variation across studies, likely influenced by study design, patient characteristics, and drug classes. In this context, MOITE's relatively elevated estimates for Falls and Fractures could reflect its sensitivity to broader real-world heterogeneity that is not fully captured in the more homogeneous trial datasets. As a multi-task learning model, MOITE also learns shared representations across outcomes. Although this allows it to leverage cross-outcome dependencies effectively, it can also lead to signal spillover, whereby associations from outcomes with stronger treatment effects (e.g., AKI or Hypotension) subtly influence the estimation of outcomes with weaker or multifactorial etiologies, such as Falls and Fractures. Furthermore, these events are highly influenced by non-treatment-related factors such as frailty, environmental risk, and mobility, all of which are underrepresented in structured data and may introduce unmeasured confounding.

Although other models showed stronger performance in select outcomes, they often exhibited greater variability or deviation from RCT findings across the board. TNet and TARNet tended to overestimate the risk of AKI and Hypotension, while DR-Net, though more conservative, showed a greater spread for outcomes such as Syncope and

Electrolyte Sensitivity. MOITE, on the contrary, provided a balanced profile, maintaining consistency without extreme over- or under-estimation across tasks. These findings underscore the importance of considering the study heterogeneity when interpreting differences between model-predicted and RCT-reported treatment effects. They also highlight the practical value of multi-task models like MOITE in real-world decision-making. Despite some differences from pooled RCT results, possibly due to averaging effects or broader population variability, MOITE's individualized predictions offer clinically useful insight, especially in contexts where trial evidence is limited or difficult to generalize. Ultimately, integrating models like MOITE into practice can help bridge the gap between controlled trial settings and the complexity of real-world patient care.

7.5 Discussion

The results of this study demonstrate the effectiveness of the MOITE model in estimating treatment effects on multiple adverse outcomes, outperforming traditional ITE estimators of a single task. The MOITE multi-task architecture consistently achieved higher AUROC and AUPRC scores, particularly in outcomes with class imbalance, such as AKI, Hypotension, and Syncope, highlighting its ability to capture complex treatment-outcome relationships while maintaining model stability and generalizability.

A key part of our analysis was to compare the risk ratios predicted by our models with those reported in clinical trials. One of the most important findings was that MOITE closely mirrored real-world evidence, strengthening its case as a reliable tool to estimate treatment effects. MOITE's risk estimates for AKI (RR: 1.27), Electrolyte Sensitivity (RR: 1.31), Hypotension (RR: 1.28), and Syncope (RR: 1.23) were remarkably close to the values reported by the RCT for AKI (1.18), Hypokalaemia (1.54), Hypotension (1.97), and Syncope (1.10). This strong alignment suggests that MOITE successfully captures how antihypertensive treatments impact patients in real-world settings, translating clinical trial insights into practical data-driven risk assessments. However, not all models showed the same level of stability. TNet and TARNet consistently overestimated the risks for AKI and Hypotension, probably because they analyze each outcome in isolation without accounting for how different health risks are connected. DR-Net, on the other hand, tended to be more conservative but showed higher variations when estimating risks for Syncope and Electrolyte Sensitivity, meaning it might be more vulnerable to variations in the dataset. The

observed differences highlight a key advantage of the multi-task learning framework of MOITE, which integrates outcome dependencies and improves the stability of the model across diverse clinical conditions. Instead of treating each outcome separately, MOITE considers how different health risks interact, allowing it to make more stable and well-rounded predictions across a range of adverse events. This leads to more consistent estimates, which is exactly what is needed for real-world treatment planning, where patients often face multiple health risks at the same time. Despite some differences from the RCT results, this discrepancy may be due to the average effect in the observational data, which can smooth out extreme values when computing overall risk estimates. Unlike RCTs, which often study more controlled populations, real-world data encompass a broader spectrum of patient variability, potentially leading to more conservative estimates. Nevertheless, MOITE's ability to calculate individualized treatment effects at the patient level ensures that risk assessments remain personalized, allowing clinicians to tailor treatment decisions based on each patient's unique health profile rather than relying solely on population-wide averages.

From a methodological perspective, MOITE's superior performance in AUPRC across most outcomes highlights its ability to identify true positives in imbalanced datasets, a critical challenge in modeling treatment effects in the real world. The improved predictive capability observed in Hypotension (AUPRC: 0.057) and AKI (AUPRC: 0.031) compared to other models suggests that MOITE effectively learns subtle risk patterns that influence treatment response, even when the occurrence of adverse events is rare. This is particularly important for clinical decision-making, as the ability to accurately identify high-risk patients allows more targeted intervention strategies. Furthermore, the integration of shared representations across multiple outcomes improves generalizability and stability. Traditional single-task models, such as TARNet and TNet, exhibited greater performance variability across outcomes, suggesting that independently estimating treatment effects for each condition may not capture common risk factors. The ability of MOITE to maintain stable performance across conditions with different prevalence rates supports its suitability to model complex treatment-response relationships in observational healthcare data.

Despite these promising findings, several limitations warrant discussion. First, while multi-task learning enhances generalization, it may introduce task interference in highly heterogeneous outcomes. Additional studies are needed to optimize task weighting strategies and assess the impact of multi-task learning across broader clinical settings. Future research should focus on external validation across independent

datasets to assess MOITE's generalizability across different healthcare systems. Moreover, extending MOITE to handle longitudinal patient trajectories and integrating multimodal data sources could further improve its predictive capabilities. Additionally, prospective studies assessing the impact of MOITE-derived treatment effect estimates on clinical decision-making would provide valuable insights into its practical utility.

Overall, our findings demonstrate that MOITE provides a robust, stable, and generalizable framework for individualized treatment effect estimation, surpassing conventional ITE models in predictive accuracy and reliability. By leveraging multi-task learning to model shared risk factors across clinical outcomes, MOITE presents a promising approach for improving precision medicine and treatment optimization in observational healthcare data. Furthermore, its alignment with RCT-derived risk ratios supports its potential use in real-world evidence generation, offering an AI-driven tool for personalized treatment decision support in clinical practice.

7.6 Relevant Publications

- **Ghosheh, G.**, Gogl. M., Zhu, T. (2025). A Perspective on Individualized Treatment Effects Estimation from Time-Series Data. Journal of the American Medical Informatics Association, 2025; https://doi.org/10.1093/jamia/ocae323
- **Ghosheh, G.**, Sheppard, J, Zhu, T. (2025) Multi-Objective Individualized Treatment Effects Estimation (ITE) for antihypertensive medications in five million patients longitudinal data. [In Submission]. [Chapter 7]

Chapter 8

Conclusion

This chapter presents a summary of the results, significance, and limitations of each of the works presented in the thesis chapters. The chapter concludes with the overall contributions of the thesis and future work in the area of personalized medicine.

8.1 Summary of Results

8.1.1 Generative Models for Predictive Modeling in Tabular EHRs

Significance: This thesis addressed the research question of improving the quality of EHRs for use in CDSS, particularly in low- and middle-income countries (LMICs). Using a deep-generative model to generate synthetic EHR data, this work tackled the critical challenges associated with data scarcity, incomplete records, and limited feature sets in resource-constrained settings. The generated data not only improved predictive modeling capabilities but also preserved the underlying data distributions and predictive importance of the features, demonstrating that synthetic data can improve EHR quality for downstream applications.

The ability to predict hospital-acquired infections (HAI) using CDSS trained on synthetic data exemplifies this improvement. HAIs pose significant risks to patient outcomes and operational costs, especially in LMICs, where data collection is often limited. By allowing the creation of realistic and high-quality synthetic datasets, this approach allows clinicians to derive actionable insights without the need for extensive additional data collection. Interpretability analysis also demonstrated that models trained on synthetic datasets maintained consistency in predictive ranking of features, aligned with established medical knowledge, and provided transparency in risk assessment.

Furthermore, this work explored the impact of synthetic dataset size on predictive model performance, addressing an underexplored dimension of EHR quality improvement. The findings show that synthetic data can effectively augment small and incomplete datasets, thereby improving the feasibility of building robust models for resource-constrained healthcare settings. This aligns directly with the overall research goal of improving the quality and usability of EHRs for critical clinical applications.

Limitations: Despite its contributions, this study acknowledges several limitations. First, the validation of this study was retrospective, which limited its applicability in real-world clinical workflows. Prospective studies, in which synthetic datasets are used to train predictive models in live settings, would provide more comprehensive information on the practical value of improving the quality of the EHR through generative models. Comparative analysis that include original, oversampled, and synthetic datasets under these conditions would offer regulators and stakeholders a clearer understanding of the potential and limitations of these methods.

Finally, while interpretability analysis were included to assess the impact of synthetic data on feature rankings, future work could focus on developing parsimonious models with reduced feature sets. This approach could improve usability and reduce the burden of data collection, further contributing to the goal of improving the quality and utility of EHRs in LMICs.

8.1.2 Individualized Generation of Imputations in Time-series EHRs

Significance: This thesis introduced IGNITE, a novel generative model that improves the quality of electronic health records (EHRs) by synthesizing personalized imputations in highly sparse and irregularly sampled time-series data. IGNITE addresses a significant gap in the field by leveraging individualized missingness patterns to generate patient data that reflect the underlying health dynamics. Unlike traditional imputation models, which often rely on rigid assumptions about missingness mechanisms, IGNITE adapts to diverse real-world missing data scenarios, making it highly versatile for clinical applications.

A key innovation of IGNITE is the introduction of the Individualized Missingness Mask (IMM), a novel representation of measurement patterns and frequencies tailored to individual patients. This IMM not only improved imputation tasks but also demonstrated potential utility in broader time-series applications, such as prediction, detection, and clustering. The model's ability to robustly generate features that were completely unobserved in the original dataset prevents the exclusion of patients

with high missingness and ensures that valuable data is not wasted. This flexibility has transformative implications for improving the usability of EHRs, particularly for personalized medicine applications where high-quality imputations can enhance diagnostic accuracy and treatment recommendations.

The robust performance of IGNITE on three large-scale ICU datasets underscores its generalizability and reliability. It consistently outperformed state-of-the-art models in downstream tasks, such as mortality prediction, achieving superior AUROC and AUPRC scores. In addition, IGNITE excelled in reconstruction tasks, demonstrating its ability to recover original patient values even under extreme missingness conditions. These results highlight IGNITE's potential to enhance data quality and modeling accuracy, particularly in high-stakes clinical applications.

Limitations: While IGNITE demonstrates significant advances, this work acknowledges several limitations that present opportunities for future research. First, the model was validated primarily on retrospective ICU datasets, which, while rich in clinical detail, may not fully represent the broader healthcare landscape. Future studies should evaluate IGNITE on datasets with more diverse sources, such as primary care records and wearable data, to assess its performance in different healthcare environments and settings with varying levels of sparsity and missingness.

Second, IGNITE was tested only on downstream tasks that involved the prediction of mortality. Although this choice aligns with the available datasets, it limits the exploration of the impact of the model on other critical healthcare tasks, such as recommendation of treatment, disease phenotyping and early diagnosis. Extending the evaluation of IGNITE to these domains would provide deeper insights into its utility and further validate its contributions to improving the quality of the EHR.

In addition, the computational complexity of IGNITE remains a consideration. Although deep learning models like IGNITE require significant resources during training, they offer fast inference times in practice. Future research could explore methods to optimize training efficiency, ensuring that IGNITE remains accessible and scalable, particularly for resource-constrained settings.

Moreover, this study relied on publicly available datasets to ensure transparency and reproducibility. Although these datasets are commonly used in healthcare analytics research, the focus on ICU data may limit perceptions of IGNITE's applicability. Expanding its evaluation to other clinical datasets will demonstrate its broader applicability and uncover potential challenges specific to different healthcare settings.

Finally, the selection of benchmark models focused on traditional statistical methods, such as MICE and LOCF, alongside advanced deep learning approaches. Al-

though this provides a comprehensive perspective on imputation effectiveness, the incorporation of additional benchmarks from medical statistics and causal inference could further enrich the comparisons and validate the advantages of IGNITE over alternative methods.

8.1.3 Individualized ITE Estimation Frameworks for Accounting for Missingness-related Confounding

Significance The SI-Mask framework represents a novel and impactful contribution to the field of causal inference by addressing missingness-related confounding in time-series data. Traditional causal inference methods often assume Sequential Ignorability, which requires that treatment assignments be independent of potential outcomes conditional on observed covariates. However, this assumption frequently fails in real-world datasets characterized by high missingness and complex confounding structures. The introduction of the SI-Mask framework offers a robust solution by dynamically accounting for missingness-related confounders through the use of adversarially trained individualized masks. This innovation improves the accuracy and reliability of the ITE estimation in scenarios where missing data can obscure causal relationships.

Using adversarial training to create individualized masks for each data point, the SI-Mask framework adapts to the unique missingness patterns and the temporal dynamics of the data. This flexibility allows the framework to mitigate the biases introduced by the observed confounders, providing more reliable causal estimates. Furthermore, the ability of SI-Mask to handle varying degrees of missingness and high-dimensional data highlights its adaptability to real-world applications across diverse domains, including healthcare, economics, and social sciences. This advancement positions SI-Mask as a valuable tool for improving the quality of causal insights derived from incomplete and observational datasets, paving the way for more informed decision-making.

Limitations Although the SI-Mask framework addresses key challenges in causal inference, certain limitations must be acknowledged. The framework relies on the assumption of Sequential Ignorability, which does not account for scenarios involving MNAR mechanisms where missingness depends on unobserved variables. This limitation could hinder the performance of SI-Mask in datasets with severe missingness or high-dimensional covariates that lack sufficient observed data. Future iterations of the framework may need to integrate techniques that can explicitly address MNAR scenarios or incorporate information about unobserved confounders.

Another limitation lies in the use of LOCF for initial imputation, which, while computationally efficient, may not fully capture the complex temporal dependencies present in many real-world datasets. Exploring more advanced imputation techniques, such as deep generative models or model-based approaches, could enhance the framework's ability to accurately represent missing values in highly non-linear and high-dimensional contexts.

The computational complexity of the adversarial training process is another challenge, particularly in large-scale applications. Balancing multiple objectives simultaneously can be resource-intensive, potentially limiting the scalability of the framework. Future work should focus on optimizing the training process, using methods such as multi-task optimization, distributed training, or architectural simplifications to improve computational efficiency.

Finally, while SI-Mask has shown promise in addressing time-varying confounding in static datasets, it has yet to be applied to dynamic treatment regimes or reinforcement learning scenarios. These contexts, which involve sequential decision-making and real-time data streams, represent important areas for extending the applicability of the framework. Future research should explore the potential of SI-Mask to handle interactive systems and dynamic datasets to further broaden its impact.

8.1.4 Multi-objective ITE Estimation from Tabular EHRs

Significance: The MOITE model is a unique contribution to the field as one of the few works to investigate the estimation of ITE in observational data across multiple outcomes. By addressing the complexities of real-world healthcare datasets, MOITE advances the understanding and application of causal inference methods in settings where randomized controlled trials are not feasible or available.

Leveraging a multi-task architecture, MOITE captures shared representations of underlying risk factors while simultaneously adapting to the unique complexities of individual outcomes. This dual capability makes it particularly effective in scenarios where outcomes are interrelated or share common predictors, such as hypotension, falls, and gout, which frequently arise from overlapping patient characteristics.

The robust performance of the model underscores its ability to generalize across complex datasets while maintaining high predictive accuracy. Its effectiveness in handling unbalanced outcomes further demonstrates its ability to address sparse positive cases, a common challenge in real-world healthcare datasets. Unlike traditional single-task models, MOITE reduces performance variability across tasks by exploiting shared learning across outcomes. This capability mitigates overfitting and improves

generalization, enabling the model to perform consistently well even in observational datasets with heterogeneous complexities.

Another significant aspect of MOITE is its alignment with the risk ratios of randomized controlled trials, further validating its robustness and clinical relevance. This combination of predictive power, consistency with established metrics, and adaptability to real-world observational data makes MO-ITE a valuable tool for causal inference studies and clinical applications.

Limitations: Despite its strengths, the MOITE model has several limitations that require further investigation. Its performance on rare or highly variable outcomes, such as syncope, remains limited. These results highlight the inherent challenges of predicting outcomes with sparse data or substantial variability, suggesting the need for additional mechanisms to improve prediction in such scenarios, such as adaptive weighting or specialized architectures.

The computational complexity of MOITE can be a barrier, particularly when applied to large datasets or high-dimensional feature spaces. While the multi-task architecture provides substantial benefits in capturing shared learning, it also increases training time and resource requirements. In contrast, simpler single-task models may be more practical for resource-constrained settings or when focusing on a single outcome. Future work could explore strategies to optimize the computational efficiency of MOITE without compromising its predictive power.

Lastly, while MOITE provides robust estimates of risk risks, its interpretability remains limited compared to traditional statistical methods such as propensity scores or hazard ratios. These traditional methods offer clear, theory-driven insights into relative risks and treatment effects, which are essential for clinical decision-making. Future efforts should aim to integrate the interpretability of these statistical approaches with the scalability and flexibility of deep learning models, potentially creating hybrid frameworks that combine the strengths of both methodologies.

8.2 Key Contributions

This thesis makes several significant contributions to advancing personalized medicine and causal inference methodologies. It introduces innovative generative models designed to improve the quality of EHRs for CDSS. These models address critical challenges, such as data scarcity, incomplete records, and limited feature sets, particularly in resource-constrained settings. By generating synthetic data that improve predic-

tive accuracy while preserving underlying data distributions, this work demonstrates how synthetic EHR data can improve healthcare decision making.

The IGNITE framework represents a novel approach to imputing highly sparse and irregularly sampled time-series data. Using individualized missingness representations, IGNITE advances data imputation methodologies and improves downstream predictive tasks, such as mortality prediction, demonstrating its transformative potential for personalized medicine applications. Similarly, the MOITE model addresses the estimation of ITE across multiple outcomes. Its multi-task architecture captures shared representations of underlying risk factors while adapting to the specific complexities of individual outcomes, offering a robust solution for causal inference in observational data.

Another significant contribution is the introduction of the SI-Mask framework, which tackles missingness-related confounding in time-series data. By dynamically accounting for confounders using adversarially trained individualized masks, the SI-Mask framework improves the accuracy and reliability of treatment effect estimations under real-world data constraints. Collectively, these methodologies not only advance computational health informatics, but also contribute to a broader understanding of improving EHR quality, optimizing treatment effect estimation, and addressing confounding in causal inference. Together, these contributions form a comprehensive methodological toolkit tailored for the complexities of modern healthcare data. The proposed models address core bottlenecks in data quality and causal estimation, offering scalable, interpretable, and personalized solutions. Moreover, they highlight the potential for harmonizing generative modeling and causal inference to create robust, clinically aligned AI systems.

This body of work lays a strong foundation for future efforts in ethical, effective, and individualized care—especially as healthcare systems increasingly adopt AI-driven tools. By rigorously tackling issues at the intersection of data availability, bias, and interpretability, the thesis contributes not just technical advances but also conceptual clarity to the field.

Code Availability

To promote transparency, reproducibility, and broader impact, the code associated with each chapter of this thesis will be released publicly upon the publication of the corresponding manuscripts. This includes implementations for generative modeling in tabular EHRs, individualized imputation of time-series data, frameworks for handling missingness-related confounding in treatment effect estimation, and multi-objective

individualized treatment effect models. The code will be made available under an open-source license and hosted on GitHub at `https://github.com/ghadeerghosheh`, with clear documentation and instructions to facilitate use by the research community.

8.3 Limitations and Future Work

Limitations: While this thesis presents significant advances, certain limitations must be acknowledged. The generalizability of the proposed methods remains a challenge due to dataset-specific assumptions and configurations. High computational demands also limit the scalability and feasibility of deploying these models in real-time clinical settings. Furthermore, reliance on retrospective validation methods limits the ability to assess the practical impact of these methods on clinical workflows. Finally, the interpretability of some proposed models is limited, posing barriers to their adoption by clinicians.

Integration of Modalities: A key challenge not yet fully addressed is the integration of multiple EHR modalities. While this thesis separately explores tabular and time-series data, combining both within a unified modeling framework could yield richer representations and stronger generalization, especially for individualized prediction tasks.

Challenges with Generative Models in Healthcare : Generative models, while powerful, come with inherent risks in clinical contexts. Issues such as mode collapse, or reproducing biases present in training data pose challenges to their safe deployment. Understanding their uncertainty and embedding domain-aware constraints are important directions for future work.

Causality vs. Interpretability : Another area deserving further exploration is the distinction between causal validity and model interpretability. While causal estimates (e.g., ITEs) are a central focus of this thesis, some deep models lack transparency in how predictions are derived. Bridging the gap between explainable modeling and causal inference remains an important open challenge.

Benchmarking and Reproducibility : A broader concern relates to the absence of standard benchmarks in individualized prediction with EHRs. While this thesis used STRATIFY and MIMIC datasets, variations in preprocessing, splitting, and

evaluation hinder comparability. This Thesis contributed a series of works that will be all opensource upon publication, creating standardized benchmarks with shared tasks, codebases, and metrics—would greatly enhance reproducibility and field-wide progress.

Incorporating Domain Knowledge : Lastly, future models should more explicitly integrate domain knowledge—such as clinical guidelines, pathophysiology, or expert heuristics. This could be achieved via model priors, soft constraints, or hybrid symbolic-neural architectures, helping ensure that learned representations are both clinically valid and trustworthy.

Future Work: Future research should focus on validating these methodologies in prospective studies and real-world clinical environments, including randomized controlled trials, to evaluate their impact on patient outcomes. Interdisciplinary collaboration with clinicians and healthcare stakeholders is essential to refine the models for better interpretability and usability. Simplifying model architectures and optimizing computational efficiency will also be critical to ensure scalability and broader applicability. Expanding the methodologies to incorporate multi-modal data, such as genomics, imaging, and wearable sensor data, can further enhance individualized predictions. In addition, addressing ethical and regulatory considerations, such as patient privacy and compliance, will be paramount to safely deploying AI-driven solutions in healthcare.

8.4 Closing Remarks

This thesis represents a significant step forward in the use of advanced machine learning techniques to address key challenges in personalized medicine. By addressing data quality issues and advancing treatment effect estimation methodologies, it provides a strong foundation for future innovations. These contributions not only advance the field of computational health informatics, but also pave the way for more precise, patient-centric healthcare delivery.

Appendix A

Leveraging Generative Models for Predictive Modeling in Tabular EHR Data

A.1 Hyperparameter Search

The searched hyperparameters for each of the models to predict HAIs are shown in Table A.1. The final parameters were chosen using GridSearch.

Table A.1: The ranges considered for the hyperparameter search for the downstream predictive modelling section.

Model	Hyperparameters	Values
Random Forest	N estimators	[100, 150,200]
K-Nearest Neighbors	N neighbors	[5,10,15,20]
	Power parameter	[1,2]
	Leaf size	[20,25,30,35,40,45,50]
Support Vector Machine	Kernel	[*poly*, *rbf*]
	Gamma	[1.e-02, 1.e+03]
	Regularization parameter C	[0.1, 1, 10, 100]

Appendix B

Individualized Generation of Imputations in Time-series EHRs

B.1 Hyper-parameters for IGNITE

The hyperparameters used for IGNITE optimization. The final parameters were chosen based on a hyper-parameter Bayesian search tuned for each dataset.

Table B.1: Hyper-parameter Search Ranges for IGNITE

Parameter	**Search Range**
γ Reconstruction	[2.e-02, 2.e+02]
δ KL Divergence	[2.e-04, 2.e+04]
ϵ Matching	[1.e-04, 1.e+04]
ζ Semantic	[1.e-04, 1.e+04]
η Contrastive	[1.e-04, 1.e+04]
θ MIT	[1.e-01, 3.e+01]

B.2 Hyper-parameters for Downstream Tasks

The hyper-parameters range that were used for training the downstream LSTM network.

B.3 Reconstruction Results for Female & Male Populations

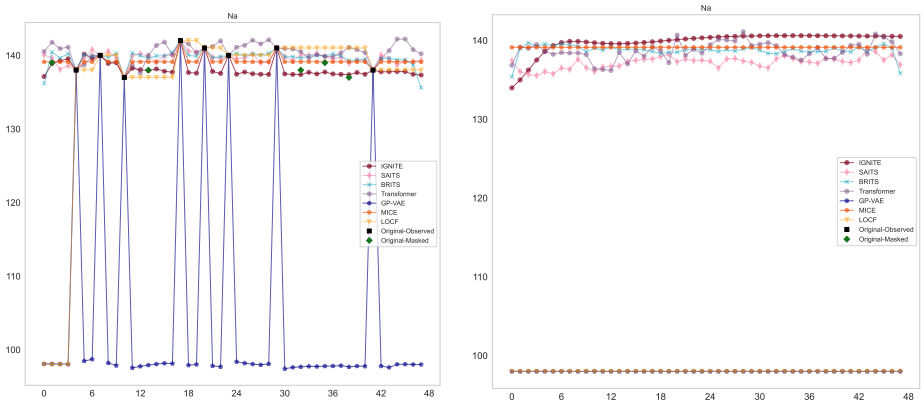
The performance of the baseline methods compared to IGNITE in terms of the reconstruction task. We present the results for Females and Male Populations B.2, respectively. In both experiments and in all percentages of masking, IGNITE consistently outperformed all other baselines.

Table B.2: Performance of the various baselines of the reconstruction task on the PhysioNet 2012 dataset. The results are calculated across all the patients in the test set and reported in terms of RMSE and MAE with mean and standard deviation. * denotes statistical significance based on the Wilcoxon Test

	RMSE			**MAE**		
Introduced Missingness	**10%**	**20%**	**50%**	**10%**	**20%**	**50%**
Females						
LOCF	0.184 (0.06)*	0.190 (0.052)*	0.202 (0.037)*	0.160 (0.041)*	0.107 (0.037)*	0.114 (0.031)*
MICE	0.108 (0.051)*	0.114 (0.044)*	0.140 (0.067)*	0.065 (0.025)*	0.066 (0.021)*	0.067 (0.019)*
CATSI	0.284 (0.067)*	0.353 (0.037)*	0.353 (0.038)*	0.181 (0.0517)*	0.196 (0.041)*	0.247 (0.031)*
GP-VAE	0.402 (0.060)*	0.405 (0.0492)*	0.407 (0.041)*	0.324 (0.051)*	0.325 (0.036)*	0.325 (0.036)*
Transformer	0.113 (0.058)*	0.123 (0.055)*	0.120 (0.040)*	0.065 (0.028)*	0.068 (0.026)*	0.066 (0.023)*
BRITS	0.109 (0.054)*	0.114 (0.045)*	0.114 (0.039)*	0.064 (0.027)	0.064 (0.023)	0.063 (0.021)*
TimesNet	0.118 (0.045)*	0.118 (0.035)*	0.122 (0.028)*	0.071 (0.026)*	0.070 (0.022)	0.071 (0.021)*
SAITS	0.116 (0.057)*	0.122 (0.052)	0.120 (0.039)*	0.067 (0.030)*	0.067 (0.026)	0.065 (0.022)*
IGNITE	**0.097 (0.038)**	**0.126 (0.028)**	**0.101 (0.027)**	**0.061 (0.02)**	**0.061 (0.020)**	**0.061 (0.017)**
Males						
LOCF	0.180 (0.066)*	0.188 (0.053)*	0.200 (0.039)*	0.104 (0.041)*	0.107 (0.036)*	0.114 (0.031)*
MICE	0.103 (0.048)*	0.109 (0.041)*	0.136 (0.069)*	0.063 (0.023)*	0.063 (0.020)*	0.066 (0.019)*
CATSI	0.284 (0.067)*	0.305 (0.052)*	0.352 (0.039)*	0.181 (0.051)*	0.197 (0.043)*	0.247 (0.033)*
GP-VAE	0.406 (0.063)*	0.406 (0.051)*	0.406 (0.041)*	0.327 (0.053)*	0.324 (0.044)*	0.324 (0.037)*
Transformer	0.112 (0.057)*	0.122 (0.049)*	0.118 (0.039)*	0.064 (0.026)*	0.066 (0.024)*	0.065 (0.021)*
BRITS	0.108 (0.051)*	0.111 (0.044)*	0.111 (0.036)*	0.063 (0.025)*	0.063 (0.022)*	0.062 (0.019)*
TimesNet	0.118 (0.046)*	0.117 (0.037)*	0.121 (0.029)*	0.071 (0.026)*	0.069 (0.022)*	0.071 (0.020)*
SAITS	0.115 (0.055)*	0.120 (0.051)*	0.119 (0.037)*	0.067 (0.027)*	0.065 (0.024)*	0.064 (0.021)*
IGNITE	**0.093 (0.034)**	**0.126 (0.027)**	**0.098 (0.024)**	**0.059 (0.019)**	**0.061 (0.019)**	**0.060 (0.015)**

B.4 Visualizations for Reconstructions

Below are the visualized reconstructions for two patients from the physioNet dataset. In this experiment, trained models were used to impute randomly masked samples for each patient. The approach used here was to introduce 50% missingness in the observed values for each patient record. We can see that GP-VAE seems to overfit the training set, while LOCF is imputed by zero when there is no observed value at the start of the patient observation recording. The visualizations for all evaluated models are shown in B.1.



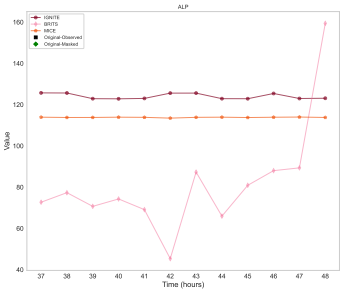
(a) Imputation for a patient with sample-wise missingness.

(b) Imputation for a patient with feature-wise missingness.

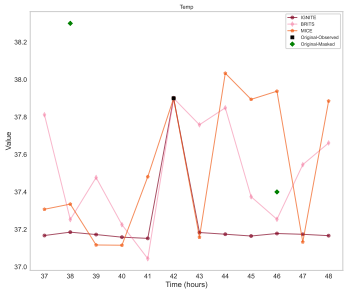
Figure B.1: Visualized imputations for two patients from the PhysioNet 2012 dataset. In a and b, we show examples of patients with different types of missingness. The examples shown are from the test set used for reconstruction experiments where 50% of the observed values in the overall patient record are masked, and various imputations are compared with the ground truth. The original masked values are shown in green.

B.5 Visualizations of Imputations for a Dead Patient

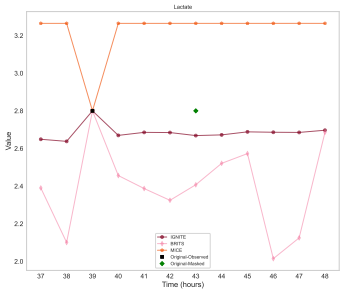
This section provides a detailed overview of imputed values, for example, variables from dead patients, using IGNITE, BRITS and MICE models. In Figure 3.1, we present a focused comparison of imputed values for urine output, HCO3, and FiO2, showcasing the performance of each model. These three variables serve as key examples, with additional imputed results for the full set of patient variables included in this section. These results underscore the superior ability of IGNITE to generate clinically meaningful imputations for all patient variables.



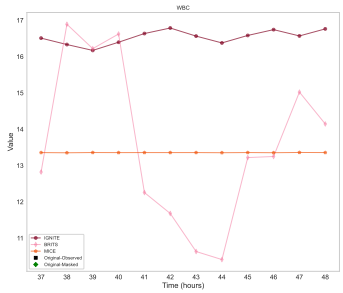
(a) ALP



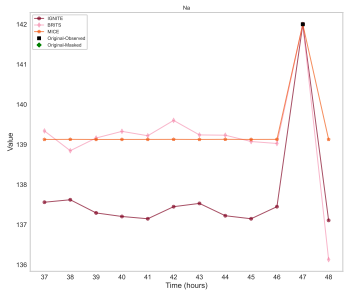
(b) Temperature



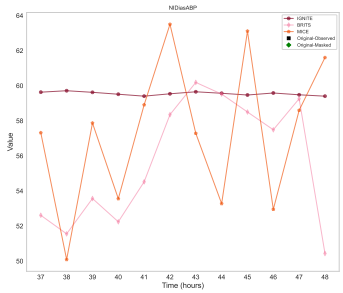
(c) Lactate



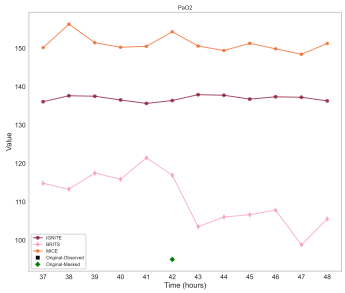
(d) WBC



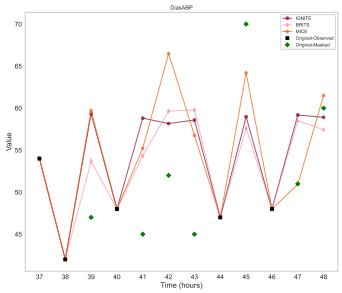
(e) Na



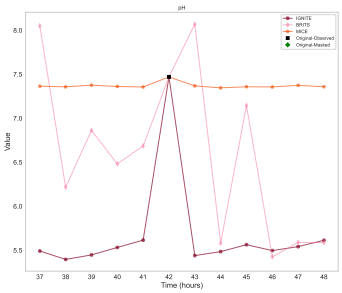
(f) Non Invasive Diastolic Blood pressure



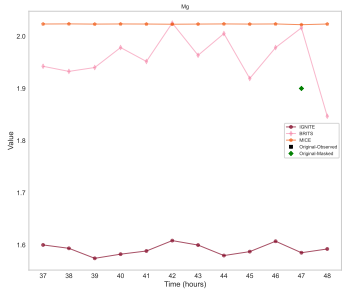
(a) Heart Rate



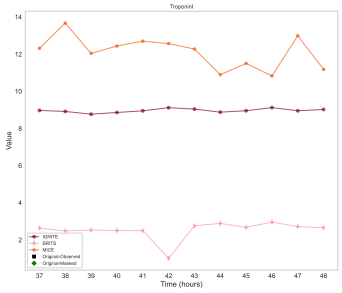
(b) Diastolic Blood Pressure



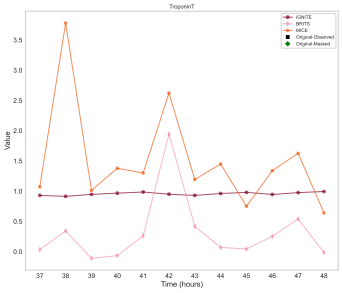
(c) pH



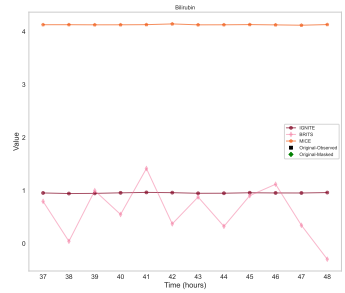
(d) Mg



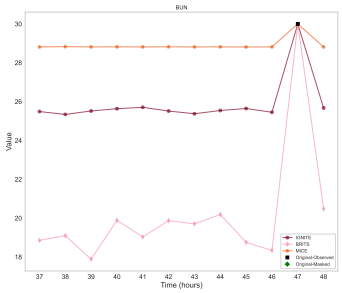
(e) TroponinI



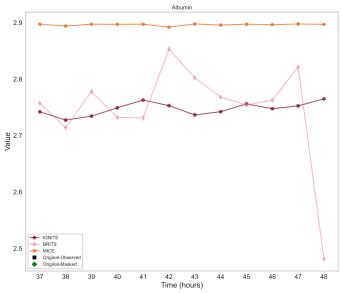
(f) TroponinT



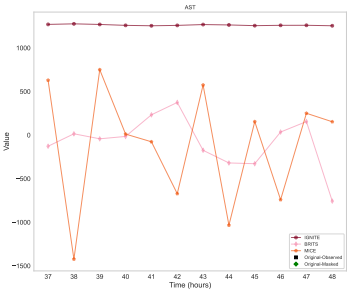
(g) Bilirubin



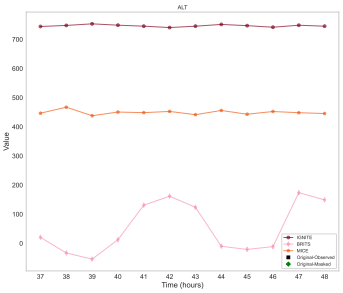
(h) BUN



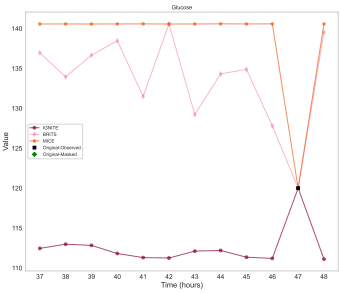
(i) Albumin



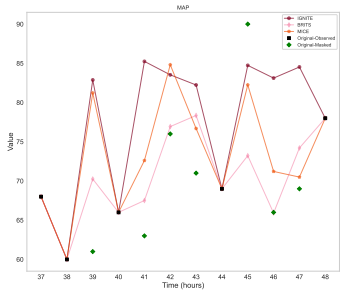
(j) AST



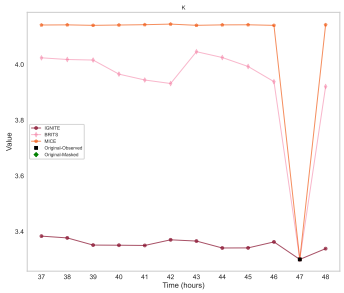
(k) ALT



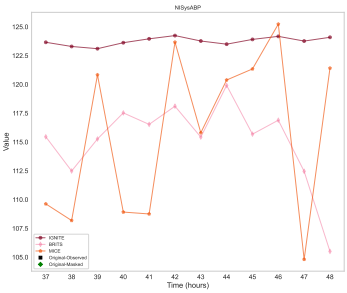
(l) Glucose



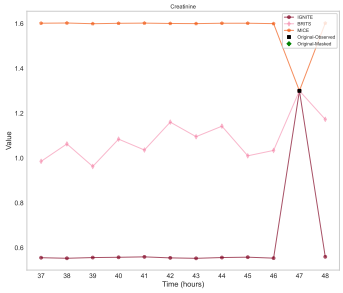
(a) MAP



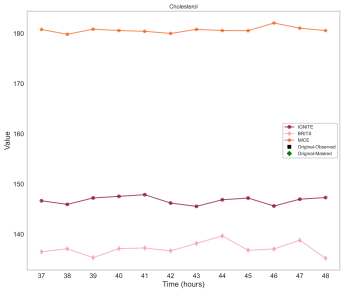
(b) K



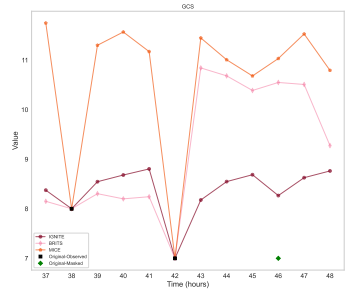
(c) NISysABP



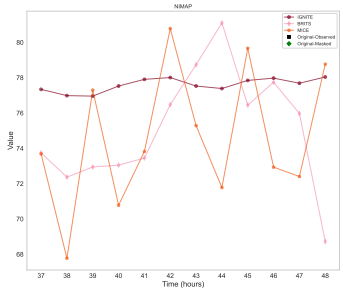
(d) Creatinine



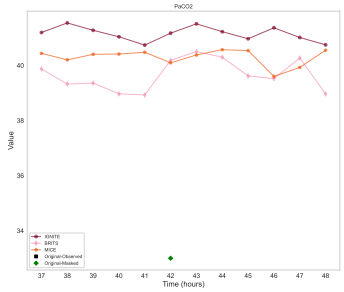
(e) Cholesterol



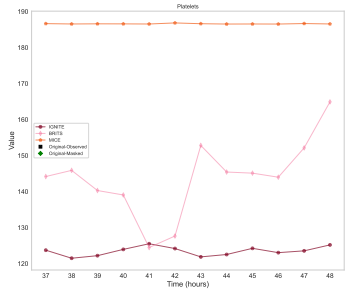
(f) GCS



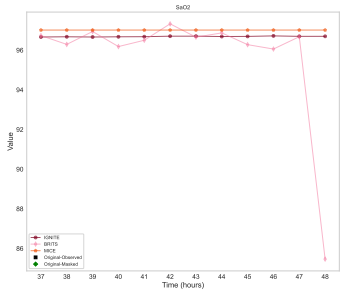
(g) NIMAP



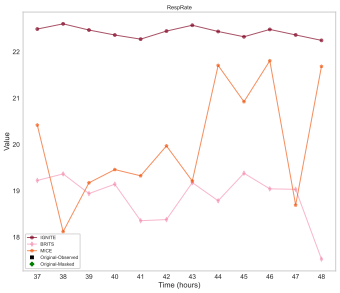
(h) PaCo2



(i) Platlets



(j) SaO2



(k) Respiratory Rate

Appendix C

Relaxing Sequential Ignorability in ITE Estimation from Time-series EHRs

C.1 Hyper-parameters for SI-Mask

The hyper-parameters range that were used for training the SI-Mask.

Table C.1: Hyper-parameter Search Ranges for the SI Mask.

Parameter	**Search Range**
Dropout	[0.1, 0.9]
Learning Rate	[1e-4, 1e-2]
Batch Size	[32, 64, 128, 256]
Hidden Dimensions	[64, 128, 256]
α (MMD weight)	[0.1, 10.0]
β (Contrastive weight)	[0.1, 5.0]

Appendix D

Multi-objective ITE Estimation from Tabular EHRs

D.1 Hyper-parameters for MOITE

The hyper-parameters range that were used for training the for Multi-Outcome ITE Estimator.

Table D.1: Hyper-parameter Search Ranges for Multi-Outcome ITE Estimator

Parameter	**Search Range**
Dropout	[0.1, 0.5]
Learning Rate	[1e-4, 1e-2]
Batch Size	[64, 256]
Hidden Dimensions	[128, 256, 512]
α (MMD weight)	[0.1, 10.0]
β (Contrastive loss weight)	[0.1, 10.0]

Bibliography

[1] Center for disease control and prevention surveillance definitions for specific types of infections, 2015.

[2] Samira Abbasgholizadeh Rahimi, Michelle Cwintal, Yuhui Huang, Pooria Ghadiri, Roland Grad, Dan Poenaru, Genevieve Gore, Hervé Tchala Vignon Zomahoun, France Légaré, and Pierre Pluye. Application of artificial intelligence in shared decision making: scoping review. *JMIR Medical Informatics*, 10(8):e36199, 2022.

[3] Noura S Abul-Husn and Eimear E Kenny. Personalized medicine and the power of electronic health records. *Cell*, 177(1):58–69, 2019.

[4] Edgar Acuna and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pages 639–647. Springer, 2004.

[5] Davies Adeloye, Peige Song, Yajie Zhu, Harry Campbell, Aziz Sheikh, and Igor Rudan. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (copd) in 2019: a systematic review and modelling analysis. *The Lancet Respiratory Medicine*, 10(5):447–458, 2022.

[6] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information systems*, 53:16–38, 2015.

[7] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[8] Ali Albasri, Miriam Hattle, Constantinos Koshiaris, Anna Dunnigan, Ben Paxton, Sarah Emma Fox, Margaret Smith, Lucinda Archer, Brooke Levis, Rupert A Payne, et al. Association between antihypertensive treatment and adverse events: systematic review and meta-analysis. *bmj*, 372, 2021.

[9] Sarah C Anoke, Sharon-Lise Normand, and Corwin M Zigler. Approaches to treatment effect heterogeneity in the presence of confounding. *Statistics in medicine*, 38(15):2797–2815, 2019.

[10] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[11] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[12] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[13] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 2023.

[14] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

[15] Pedro Baqui, Valerio Marra, Ahmed M Alaa, Ioana Bica, Ari Ercole, and Mihaela van der Schaar. Comparing covid-19 risk factors in brazil using machine learning: the importance of socioeconomic, demographic and structural factors. *Scientific reports*, 11(1):1–10, 2021.

[16] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

[17] Dmitry I Belov and Ronald D Armstrong. Distributions of the kullback–leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*, 64(2):291–309, 2011.

[18] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[19] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

[20] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

[21] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

[22] Guthrie S Birkhead, Michael Klompas, and Nirav R Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, 36:345–359, 2015.

[23] Nick Black. Why we need observational studies to evaluate the effectiveness of health care. *Bmj*, 312(7040):1215–1218, 1996.

[24] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33, 2021.

[25] Iuliia Branco and Altino Choupina. Bioinformatics: new tools and applications in life science and personalized medicine. *Applied microbiology and biotechnology*, 105(3):937–951, 2021.

[26] M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114, 2010.

[27] Dorothy Bulas and Alexia Egloff. Benefits and risks of mri in pregnancy. In *Seminars in perinatology*, pages 301–304. Elsevier, 2013.

[28] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[29] Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating multi-categorical samples with generative adversarial networks. *arXiv preprint arXiv:1807.01202*, 2018.

[30] Ramiro D Camino, Christian A Hammerschmidt, and Radu State. Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666*, 2019.

[31] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *arXiv preprint arXiv:1805.10572*, 2018.

[32] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[33] Taha Ceritli, Ghadeer O Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P Creagh, and David A Clifton. Synthesizing mixed-type electronic health records using diffusion models. *arXiv preprint arXiv:2302.14679*, 2023.

[34] Ying-Jui Chang, Min-Li Yeh, Yu-Chuan Li, Chien-Yeh Hsu, Chao-Cheng Lin, Meng-Shiuan Hsu, and Wen-Ta Chiu. Predicting hospital-acquired infections by scoring system with simple parameters. *PloS one*, 6(8):e23137, 2011.

[35] 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. IEEE, 2017.

[36] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

[37] Peipei Chen, Wei Dong, Xudong Lu, Uzay Kaymak, Kunlun He, and Zhengxing Huang. Deep representation learning for individualized treatment effect estimation using electronic health records. *Journal of biomedical informatics*, 100:103303, 2019.

[38] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, pages 1–5, 2021.

[39] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2180–2188, 2016.

[40] Zhaoyi Chen, Xiong Liu, William Hogan, Elizabeth Shenkman, and Jiang Bian. Applications of artificial intelligence in drug development using real-world data. *Drug Discovery Today*, 26(5):1256–1264, 2021.

[41] Nicholas C Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15(1):14–20, 2022.

[42] Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of heterogeneous synthetic electronic health records using gans. In *workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019.

[43] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[44] 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 *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

[45] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016.

[46] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[47] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[48] Jiebin Chu, Wei Dong, Jinliang Wang, Kunlun He, and Zhengxing Huang. Treatment effect prediction with adversarial deep learning using electronic health records. *BMC Medical Informatics and Decision Making*, 20(4):1–14, 2020.

[49] Paul Collinson. Troponin measurement in routine clinical practice: the reality behind the guidelines, 2022.

[50] Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

[51] Mike Conway, Richard L Berg, David Carrell, Joshua C Denny, Abel N Kho, Iftikhar J Kullo, James G Linneman, Jennifer A Pacheco, Peggy Peissig, Luke Rasmussen, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In *AMIA annual symposium proceedings*, volume 2011, page 274. American Medical Informatics Association, 2011.

[52] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.

[53] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[54] D Curran, M Bacchi, SF Schmitz, G Molenberghs, and RJ Sylvester. Identifying the types of missingness in quality of life data from clinical trials. *Statistics in medicine*, 17(5-7):739–756, 1998.

[55] Ulrich Curth, Vladimir Prokhorenko, and Jan Schneider. Nonparametric estimation of heterogeneous treatment effects with random forests. *arXiv preprint arXiv:1906.02028*, 2019.

[56] Arianna Dagliati, Alberto Malovini, Valentina Tibollo, and Riccardo Bellazzi. Health informatics and ehr to support clinical research in the covid-19 pandemic: an overview. *Briefings in bioinformatics*, 22(2):812–822, 2021.

[57] Ralph B D'Agostino Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281, 1998.

[58] Zongyu Dai, Zhiqi Bu, and Qi Long. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 791–798. IEEE, 2021.

[59] Melissa A Daubert and Allen Jeremias. The utility of troponin measurement to detect myocardial infarction: review of the current findings. *Vascular health and risk management*, 6:691, 2010.

[60] Chloe de Grood, Jeanna Parsons Leigh, Sean M Bagshaw, Peter M Dodek, Robert A Fowler, Alan J Forster, Jamie M Boyd, and Henry T Stelfox. Patient, family and provider experiences with transfers from intensive care unit to hospital ward: a multicentre qualitative study. *CMAJ*, 190(22):E669–E676, 2018.

[61] Spiros Denaxas, J Lee, AH Shah, P Bankhead, A Wong, J van Vlymen, I Wong, B Molaee-Ardekani, TP van Staa, J Sheppard, et al. Stratifying risk of emergency hospital admission and death in adults in england: national derivation and validation data linkage study. *JMIR medical informatics*, 7(4):e14337, 2019.

[62] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[63] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[64] Yuhan Du, Anthony R Rafferty, Fionnuala M McAuliffe, Lan Wei, and Catherine Mooney. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Scientific Reports*, 12(1):1170, 2022.

[65] Christopher J Duff, Ivonne Solis-Trapala, Owen J Driskell, David Holland, Helen Wright, Jenna L Waldron, Clare Ford, Jonathan J Scargill, Martin Tran, Fahmy WF Hanna, et al. The frequency of testing for glycated haemoglobin, hba1c, is linked to the probability of achieving target levels in patients with suboptimally controlled diabetes mellitus. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 57(2):296–304, 2019.

[66] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[67] Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174:114582, 2021.

[68] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[69] M Faltys, M Zimmermann, X Lyu, M Hüser, S Hyland, G Rätsch, and TM Merz. Hirid, a high time-resolution icu dataset (version 1.1. 1), 2021.

[70] Jerome Fan, Suneel Upadhye, and Andrew Worster. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20, 2006.

[71] Bassam Farran, Arshad Mohamed Channanath, Kazem Behbehani, and Thangavel Alphonse Thanaraj. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait—a cohort study. *BMJ open*, 3(5):e002457, 2013.

[72] Renee Fekieta, Alana Rosenberg, Beth Hodshon, Shelli Feder, Sarwat I Chaudhry, and Beth L Emerson. Organisational factors underpinning intrahospital transfers: a guide for evaluating context in quality improvement. *Health Systems*, 10(4):239–248, 2021.

[73] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[74] Centers for Disease Control, Prevention, et al. Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report*, 52(Suppl 1):1–17, 2003.

[75] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.

[76] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[77] Arturo Galindo-Fraga, Marco Villanueva-Reza, and Eric Ochoa-Hein. Current challenges in antibiotic stewardship in low-and middle-income countries. *Current Treatment Options in Infectious Diseases*, 10(3):421–429, 2018.

[78] Yaroslav Ganin and V Lempitsky. Unsupervised domain adaptation by backpropagation. arxiv. *arXiv preprint arXiv:1409.7495*, 2014.

[79] Qiyang Ge, Xuelin Huang, Shenying Fang, Shicheng Guo, Yuanyuan Liu, Wei Lin, and Momiao Xiong. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in genetics*, 11, 2020.

[80] Changran Geng, Harald Paganetti, and Clemens Grassberger. Prediction of treatment response for combined chemo- and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. *Scientific Reports*, 7(1), October 2017.

[81] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Chen, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

[82] Afshin Gholamy, Vladik Kreinovich, and Olga Kosheleva. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. 2018.

[83] Shantanu Ghosh, Christina Boucher, Jiang Bian, and Mattia Prosperi. Propensity score synthetic augmentation matching using generative adversarial networks (pssam-gan). *Computer methods and programs in biomedicine update*, 1:100020, 2021.

[84] Ghadeer Ghosheh, Jin Li, and Tingting Zhu. A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources. *arXiv preprint arXiv:2203.07018*, 2022.

[85] Ghadeer O Ghosheh, Jin Li, and Tingting Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys*, 2023.

[86] Ghadeer O Ghosheh, Jin Li, and Tingting Zhu. Ignite: Individualized generation of imputations in time-series electronic health records. *arXiv preprint arXiv:2401.04402*, 2024.

[87] Ghadeer O Ghosheh, Jin Li, and Tingting Zhu. Understanding missingness in time-series electronic health records for individualized representation. *arXiv preprint arXiv:2402.15730*, 2024.

[88] Michael Steven Gold and Peter M Bentler. Treatments of missing data: A monte carlo comparison of rbhdi, iterative stochastic regression imputation, and expectation-maximization. *Structural equation modeling*, 7(3):319–355, 2000.

[89] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[90] Cory E Goldstein, Charles Weijer, Jamie C Brehaut, Dean A Fergusson, Jeremy M Grimshaw, Austin R Horn, and Monica Taljaard. Ethical issues in pragmatic randomized controlled trials: a review of the recent literature identifies gaps in ethical argumentation. *BMC medical ethics*, 19(1):1–10, 2018.

[91] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

[92] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[93] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[94] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[95] Mehak Gupta, Thao-Ly T Phan, H Timothy Bunnell, and Rahmatollah Beheshti. Concurrent imputation and prediction on ehr data using bi-directional gans: Bi-gans for ehr imputation and prediction. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

[96] Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[97] Zhongyang Han, Jun Zhao, Henry Leung, King Fai Ma, and Wei Wang. A review of deep learning models for time series prediction. *IEEE Sensors Journal*, 21(6):7833–7848, 2019.

[98] John Hancock, Taghi M Khoshgoftaar, and Justin M Johnson. Informative evaluation metrics for highly imbalanced big data classification. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1419–1426. IEEE, 2022.

[99] Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. *arXiv preprint arXiv:2104.09323*, 2021.

[100] Tobias Hatt and Stefan Feuerriegel. Sequential deconfounding for causal inference with unobserved confounders. In *Causal Learning and Reasoning*, pages 934–956. PMLR, 2024.

[101] RJ Hayes and S Bennett. Simple sample size calculation for cluster-randomized trials. *International journal of epidemiology*, 28(2):319–326, 1999.

[102] Thomas Heldt, Ramakrishna Mukkamala, George B Moody, and Roger G Mark. CVSim: An open-source cardiovascular simulator for teaching and research. *Open Pacing Electrophysiol. Ther. J.*, 3:45–54, 2010.

[103] Catherine Helmer, Karine Pérès, Antoine Pariente, Florence Pasquier, Sophie Auriacombe, Michel Poncet, Florence Portet, Olivier Rouaud, Karen Ritchie, Christophe Tzourio, et al. Primary and secondary care consultations in elderly demented individuals in france. *Dementia and geriatric cognitive disorders*, 26(5):407–415, 2008.

[104] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35:1–9, 2016.

[105] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 2022.

[106] Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd Van Staa, and Liam Smeeth. Data resource profile: clinical practice research datalink (cprd). *International journal of epidemiology*, 44(3):827–836, 2015.

[107] Julian PT Higgins, Sally Green, et al. Cochrane handbook for systematic reviews of interventions. 2008.

[108] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[109] R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.

[110] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[111] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[112] Susan D Horn, Gerben DeJong, David K Ryser, Peter J Veazie, and Jeffrey Teraoka. Another look at observational studies in rehabilitation research: going beyond the holy grail of the randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, 86(12):8–15, 2005.

[113] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. *Cambridge university press*, 28, 2015.

[114] Michael Imhoff. Acquisition of icu data: concepts and demands. *International journal of clinical monitoring and computing*, 9(4):229–237, 1992.

[115] Giovanni Improta, Valeria Mazzella, Donatella Vecchione, Stefania Santini, and Maria Triassi. Fuzzy logic–based clinical decision support system for the evaluation of renal function in post-transplant patients. *Journal of evaluation in clinical practice*, 26(4):1224–1234, 2020.

[116] R Indhumathi and S Sathiya Devi. Healthcare cramér generative adversarial network (hcgan). *Distributed and Parallel Databases*, pages 1–17, 2021.

[117] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17(1):1–10, 2017.

[118] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[119] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE, 2013.

[120] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[121] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[122] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), May 2016.

[123] 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.

[124] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

[125] Rita R Kalyani, Sherita H Golden, and William T Cefalu. Diabetes and aging: unique considerations and goals of care. *Diabetes Care*, 40(4):440–443, 2017.

[126] Maged N Kamel Boulos and Peng Zhang. Digital twins: from personalised medicine to precision public health. *Journal of Personalized Medicine*, 11(8):745, 2021.

[127] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[128] Edward H Kennedy et al. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 5, 2020.

[129] Ismail Keshta and Ammar Odeh. Security and privacy of electronic health records: Concerns and challenges. *Egyptian Informatics Journal*, 22(2):177–183, 2021.

[130] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[131] Alison M Kim, Candace M Tingen, and Teresa K Woodruff. Sex bias in trials and treatment must end. *Nature*, 465(7299):688–689, 2010.

[132] Bo-Guen Kim, Minwoong Kang, Jihyun Lim, Jin Lee, Danbee Kang, Minjung Kim, Jinhee Kim, Hyejeong Park, Kyung Hoon Min, Juhee Cho, et al. Comprehensive risk assessment for hospital-acquired pneumonia: sociodemographic, clinical, and hospital environmental factors associated with the incidence of hospital-acquired pneumonia. *BMC pulmonary medicine*, 22:1–11, 2022.

[133] Ellen Kim, Samuel M Rubinstein, Kevin T Nead, Andrzej P Wojcieszynski, Peter E Gabriel, and Jeremy L Warner. The evolving use of electronic health records (ehr) for research. In *Seminars in radiation oncology*, pages 354–361. Elsevier, 2019.

[134] Jaeyoon Kim, Donghyun Tae, and Junhee Seok. A survey of missing data imputation using generative adversarial networks. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 454–456. IEEE, 2020.

[135] Joanne Kim, Gilad Horowitz, Michael Hong, Mario Orsini, Sylvia L Asa, and Kevin Higgins. The dangers of parathyroid biopsy. *Journal of Otolaryngology-Head & Neck Surgery*, 46(1):1–4, 2017.

[136] Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. List-wise deletion is evil: what to do about missing data in political science. In *Annual Meeting of the American Political Science Association, Boston*, volume 52, 1998.

[137] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[138] Helena Chmura Kraemer. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, 26(3):533–541, 2000.

[139] David M Kreindler and Charles J Lumsden. The effects of the irregular sample and missing data in time series analysis. *Nonlinear dynamics, psychology, and life sciences*, 2006.

[140] Margaret E Kruk, Anna D Gage, Catherine Arsenault, Keely Jordan, Hannah H Leslie, Sanam Roder-DeWan, Olusoji Adeyi, Pierre Barker, Bernadette Daelmans, Svetlana V Doubova, et al. High-quality health systems in the sustainable development goals era: time for a revolution. *The Lancet global health*, 6(11):e1196–e1252, 2018.

[141] S R Kunzel, J S Sekhon, A Bennett, and B Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[142] Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. Deconfounding temporal autoencoder: estimating treatment effects over time using noisy proxies. In *Machine Learning for Health*, pages 143–155. PMLR, 2021.

[143] Young Joon Kwon, Danielle Toussie, Lea Azour, Jose Concepcion, Corey Eber, G Anthony Reina, Ping Tak Peter Tang, Amish H Doshi, Eric K Oermann, and Anthony B Costa. Appropriate evaluation of diagnostic utility of machine learning algorithm generated images. In *Machine Learning for Health*, pages 179–193. PMLR, 2020.

[144] SK Lakshmanaprabu, Sachi Nandan Mohanty, Sujatha Krishnamoorthy, J Uthayakumar, K Shankar, et al. Online clinical decision support system using optimal deep neural networks. *Applied Soft Computing*, 81:105487, 2019.

[145] The Lancet. Personalised medicine in the uk. *Lancet (London, England)*, 391(10115):e1, 2018.

[146] Daniel T Larose and Chantal D Larose. k-nearest neighbor algorithm. 2014.

[147] Dongha Lee, Hwanjo Yu, Xiaoqian Jiang, Deevakar Rogith, Meghana Gudala, Mubeen Tejani, Qiuchen Zhang, and Li Xiong. Generating sequential electronic health records using dual adversarial autoencoder. *Journal of the American Medical Informatics Association*, 27(9):1411–1419, 2020.

[148] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[149] Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digital Medicine*, 6(1):98, 2023.

[150] Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1):98, 2023.

[151] Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, et al. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pages 282–299. PMLR, 2021.

[152] Wenyuan Li, Yunlong Wang, Yong Cai, Corey Arnold, Emily Zhao, and Yilian Yuan. Semi-supervised rare disease detection using generative adversarial network. *arXiv preprint arXiv:1812.00547*, 2018.

[153] Yunzhe Li, Kun Kuang, Bo Li, Peng Cui, Jianrong Tao, Hongxia Yang, and Fei Wu. Continuous treatment effect estimation via generative adversarial deconfounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 4–22. PMLR, 2020.

[154] Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31, 2018.

[155] WA Lindsay, MM Murphy, DS Almghairbi, and IK Moppett. Age, sex, race and ethnicity representativeness of randomised controlled trials in peri-operative medicine. *Anaesthesia*, 75(6):809–815, 2020.

[156] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a better measure than accuracy in comparing learning algorithms. In *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings 16*, pages 329–341. Springer, 2003.

[157] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[158] Ruoqi Liu, Changchang Yin, and Ping Zhang. Estimating individual treatment effects with time-varying confounders. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 382–391. IEEE, 2020.

[159] Katerina G Lourida and George E Louridas. Constraints in clinical cardiology and personalized medicine: Interrelated concepts in clinical cardiology. *Cardiogenetics*, 11(2):50–67, 2021.

[160] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[161] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[162] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[163] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[164] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th international joint conference on artificial intelligence*, pages 3094–3100. AAAI Press, 2019.

[165] C Mack, Z Su, and D Westreich. Types of missing data. agency for healthcare research and quality (us). 2018, 2022.

[166] Jeanne M Madden, Matthew D Lakoma, Donna Rusinak, Christine Y Lu, and Stephen B Soumerai. Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association*, 23(6):1143–1149, 2016.

[167] AP Majtey, PW Lamberti, and DP Prato. Jensen-shannon divergence as a measure of distinguishability between mixed quantum states. *Physical Review A*, 72(5):052310, 2005.

[168] R Malarvizhi and Antony Selvadoss Thanamani. K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1):5–7, 2012.

[169] Tatiana Malygina, Elena Ericheva, and Ivan Drokin. Data augmentation with gan: Improving chest x-ray pathologies prediction on class-imbalanced cases. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 321–334. Springer, 2019.

[170] Shie Mannor, Dori Peleg, and Reuven Rubinstein. The cross entropy method for classification. In *Proceedings of the 22nd international conference on Machine learning*, pages 561–568, 2005.

[171] Benjamin Marlin. *Missing data problems in machine learning*. PhD thesis, University of Massachusetts Amherst, 2008.

[172] Stan Matwin, Jordi Nin, Morvarid Sehatkar, and Tomasz Szapiro. A review of attribute disclosure control. *Advanced Research in Data Privacy*, pages 41–61, 2015.

[173] Argyro Mavrogiorgou, Athanasios Kiourtis, Spyridon Kleftakis, Konstantinos Mavrogiorgos, Nikolaos Zafeiropoulos, and Dimosthenis Kyriazis. A catalogue of machine learning algorithms for healthcare risk predictions. *Sensors*, 22(22):8615, 2022.

[174] Matthew McDermott, Tom Yan, Tristan Naumann, Nathan Hunt, Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. Semi-supervised biomedical translation with cycle wasserstein regression gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[175] Roseanne McNamee. Confounding and confounders. *Occupational and environmental medicine*, 60(3):227–234, 2003.

[176] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, pages 15293–15329. PMLR, 2022.

[177] Nathan K Mensah, Richard O Boadu, Godwin Adzakpah, Obed U Lasim, Ruth D Amuakwa, Hannah B Taylor-Abdulai, and Samuel T Chatio. Electronic health records post-implementation challenges in selected hospitals: A qualitative study in the central region of southern ghana. *Health Information Management Journal*, page 18333583221096899, 2022.

[178] Bradley D Menz, Sophie L Stocker, Nick Verougstraete, Danijela Kocic, Peter Galettis, Christophe P Stove, and Stephanie E Reuter. Barriers and opportunities for the clinical implementation of therapeutic drug monitoring in oncology. *British Journal of Clinical Pharmacology*, 87(2):227–236, 2021.

[179] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[180] Lu Mi, Macheng Shen, and Jingzhao Zhang. A probe towards understanding gan and vae models. *arXiv preprint arXiv:1812.05676*, 2018.

[181] Anne Mills. Health care systems in low-and middle-income countries. *New England Journal of Medicine*, 370(6):552–557, 2014.

[182] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[183] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022.

[184] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An introduction to g methods. *International journal of epidemiology*, 46(2):756–762, 2017.

[185] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

[186] Fulufhelo V Nelwamondo, Shakir Mohamed, and Tshilidzi Marwala. Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*, pages 1514–1521, 2007.

[187] Daniel A Newman and Jonathan M Cottrell. Missing data bias: Exactly how bad is pairwise deletion? In *More statistical and methodological myths and urban legends*, pages 143–171. Routledge, 2014.

[188] Simon J Newsome, Ruth H Keogh, and Rhian M Daniel. Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty. *Statistics in medicine*, 37(15):2367–2390, 2018.

[189] Do Thi Thuy Nga, Nguyen Thi Kim Chuc, Nguyen Phuong Hoa, Nguyen Quynh Hoa, Nguyen Thi Thuy Nguyen, Hoang Thi Loan, Tran Khanh Toan, Ho Dang Phuc, Peter Horby, Nguyen Van Yen, et al. Antibiotic sales in rural and urban pharmacies in northern vietnam: an observational study. *BMC Pharmacology and Toxicology*, 15(1):1–10, 2014.

[190] Kinh Van Nguyen, Nga Thuy Thi Do, Arjun Chandna, Trung Vu Nguyen, Ca Van Pham, Phuong Mai Doan, An Quoc Nguyen, Chuc Kim Thi Nguyen, Mattias Larsson, Socorro Escalante, et al. Antibiotic use and resistance in emerging economies: a situation analysis for viet nam. *BMC public health*, 13(1):1–10, 2013.

[191] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[192] Marlies Noordzij, Merel van Diepen, Fergus C Caskey, and Kitty J Jager. Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrology Dialysis Transplantation*, 32(suppl_2):ii13–ii18, 2017.

[193] Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859, 2015.

[194] Michal Ozery-Flato, Pierre Thodoroff, Matan Ninio, Michal Rosen-Zvi, and Tal El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

[195] S Palmer, A Jansen, K Leitmeyer, H Murdoch, and F Forland. Evidence-based medicine applied to the control of communicable disease incidents when evidence is scarce and the time is limited. *Eurosurveillance*, 18(25), 2013.

[196] Saeed Piri. Missing care: A framework to address the issue of frequent missing values; the case of a clinical decision support system for parkinson's disease. *Decision Support Systems*, 136:113339, 2020.

[197] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[198] Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning: Methods and applications*, pages 307–323. Springer, 2012.

[199] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[200] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, pages 29–48. Citeseer, 2003.

[201] Manizheh Ranjbar, Parham Moradi, Mostafa Azami, and Mahdi Jalili. An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. *Engineering Applications of Artificial Intelligence*, 46:58–66, 2015.

[202] Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. Smooth-gan: towards sharp and smooth synthetic ehr data generation. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 37–48. Springer, 2020.

[203] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[204] James Robins and Miguel Hernan. Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599, 2008.

[205] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

[206] Laurence D Robinson and Nicholas P Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 227–240, 1991.

[207] Paul R Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

[208] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[209] Donald B Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health services & outcomes research methodology*, 2(3-4):169–188, 2001.

[210] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[211] Kristina E Rudd, Christopher W Seymour, Adam R Aluisio, Marc E Augustin, Danstan S Bagenda, Abi Beane, Jean Claude Byiringiro, Chung-Chou H Chang, L Nathalie Colas, Nicholas PJ Day, et al. Association of the quick sequential

(sepsis-related) organ failure assessment (qsofa) score with excess hospital mortality in adults with suspected infection in low-and middle-income countries. *Jama*, 319(21):2202–2211, 2018.

[212] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

[213] Sergio Sanchez-Martinez, Oscar Camara, Gemma Piella, Maja Cikes, Miguel Angel Gonzalez Ballester, Marius Miron, Alfredo Vellido, Emilia Gomez, Alan Fraser, and Bart Bijnens. Machine learning for clinical decision-making: challenges and opportunities. 2019.

[214] Robert William Sanson-Fisher, Billie Bonevski, Lawrence W Green, and Cate D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2):155–161, 2007.

[215] Sara Santiso, Arantza Casillas, and Alicia Pérez. The class imbalance problem detecting adverse drug reactions in electronic health records. *Health informatics journal*, 25(4):1768–1778, 2019.

[216] Ashish Sarraju, Andrew Ward, Sukyung Chung, Jiang Li, David Scheinker, and Fàtima Rodriguez. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open heart*, 8(2):e001802, 2021.

[217] Judi Scheffer. Dealing with missing data. 2002.

[218] Roosmarijn MC Schelvis, Karen M Oude Hengel, Alex Burdorf, Birgitte M Blatter, Jorien E Strijk, and Allard J van der Beek. Evaluation of occupational health interventions using a randomized controlled trial: challenges and alternative research designs. *Scandinavian journal of work, environment & health*, pages 491–503, 2015.

[219] Kenneth F Schulz, Iain Chalmers, Richard J Hayes, and Douglas G Altman. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama*, 273(5):408–412, 1995.

[220] Jason Schwend, Susan Athey, and Simone Wager. Recursive partitioning for heterogeneous causal effects. In *Advances in Neural Information Processing Systems*, pages 10124–10134, 2018.

[221] Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect using counterfactual regression. In *Advances in neural information processing systems*, pages 3461–3469, 2016.

[222] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[223] Farah Shamout, Tingting Zhu, and David A Clifton. Machine learning for clinical outcome prediction. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2020.

[224] Jun Shao and Bob Zhong. Last observation carry-forward and last observation analysis. *Statistics in medicine*, 22(15):2429–2441, 2003.

[225] Anis Sharafoddini, Joel A Dubin, David M Maslove, Joon Lee, et al. A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR medical informatics*, 7(1):e11605, 2019.

[226] RW Shaw. Adverse long-term effects of oral contraceptives: a review. *British journal of obstetrics and gynaecology*, 94(8):724–730, 1987.

[227] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[228] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[229] Yajuan Si, Mari Palta, and Maureen Smith. Bayesian profiling multiple imputation for missing electronic health records. *arXiv preprint arXiv:1906.00042*, 2019.

[230] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in

cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[231] Helen R Sofaer, Jennifer A Hoeting, and Catherine S Jarnevich. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577, 2019.

[232] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[233] Leif Sörnmo and Pablo Laguna. Electrocardiogram (ecg) signal processing. *Wiley encyclopedia of biomedical engineering*, 2006.

[234] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[235] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[236] Charles A Stiller. Centralised treatment, entry to trials and survival. *British journal of cancer*, 70(2):352–362, 1994.

[237] Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data. *arXiv preprint arXiv:2010.12493*, 2020.

[238] Ping Sun. Mice-da: a mice method with data augmentation for missing data imputation in ieee ichi 2019 dacmi challenge. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–3. IEEE, 2019.

[239] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):1–10, 2020.

[240] Assel Syzdykova, André Malta, Maria Zolfo, Ermias Diro, José Luis Oliveira, et al. Open-source electronic health record systems for low-resource settings: systematic review. *JMIR medical informatics*, 5(4):e8131, 2017.

[241] Jonathan M Teich. Information systems support for emergency medicine. *Annals of emergency medicine*, 31(3):304–307, 1998.

[242] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[243] Duong Bich Thuy, James Campbell, Le Thanh Hoang Nhat, Nguyen Van Minh Hoang, Nguyen Van Hao, Stephen Baker, Ronald B Geskus, Guy E Thwaites, Nguyen Van Vinh Chau, and C Louise Thwaites. Hospital-acquired colonization and infections in a vietnamese intensive care unit. *PLoS One*, 13(9):e0203600, 2018.

[244] Sheldon W Tobe, Diane Hua, and Patrick Twohig. Clinical practice guidelines. Future Medicine, 2013.

[245] Amirsina Torfi and Edward A Fox. Cor-gan: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. *arXiv preprint*, 2020.

[246] Bheki ETH Twala, MC Jones, and David J Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

[247] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[248] Tjeerd Van der Ploeg, Peter C Austin, and Ewout W Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, 14(1):1–13, 2014.

[249] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[250] Kaushik P Venkatesh, Marium M Raza, and Joseph C Kvedar. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation, 2022.

[251] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:3152676, 2017.

[252] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[253] Jun Wang, Wenjie Du, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: A survey. *arXiv preprint arXiv:2402.04059*, 2024.

[254] Lu Wang, Wei Zhang, and Xiaofeng He. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In Guoliang Li, Jun Yang, Joao Gama, Juggapong Natwichai, and Yongxin Tong, editors, *Database Systems for Advanced Applications*, pages 36–52, Cham, 2019. Springer International Publishing.

[255] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. MIMIC-extract. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, April 2020.

[256] Shuo Wang, Carsten Rudolph, Surya Nepal, Marthie Grobler, and Shangyu Chen. Part-gan: Privacy-preserving time-series sharing. In *International Conference on Artificial Neural Networks*, pages 578–593. Springer, 2020.

[257] Yuanjia Wang, Peng Wu, Ying Liu, Chunhua Weng, and Donglin Zeng. Learning optimal individualized treatment rules from electronic health record data. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 65–71. IEEE, 2016.

[258] Ann M Weber, Ribhav Gupta, Safa Abdalla, Beniamino Cislaghi, Valerie Meausoone, and Gary L Darmstadt. Gender-related data missingness, imbalance and bias in global health surveys. *BMJ global health*, 6(11):e007405, 2021.

[259] John Weldon, Tomas Ward, and Eoin Brophy. Generation of synthetic electronic health records using a federated gan. *arXiv preprint arXiv:2109.02543*, 2021.

[260] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3), 2013.

[261] Tyler Williamson and Pietro Ravani. Marginal structural models in clinical research: when and how to use them? *Nephrology Dialysis Transplantation*, 32(suppl_2):ii84–ii90, 2017.

[262] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[263] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

[264] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[265] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.

[266] Alexandre Yahi, Rami Vanguri, Noémie Elhadad, and Nicholas P Tatonetti. Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. *arXiv preprint arXiv:1712.00164*, 2017.

[267] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

[268] Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A Malin. Generating electronic health records with multiple data types and constraints. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1335. American Medical Informatics Association, 2020.

[269] Fan Yang, Zhongping Yu, Yunfan Liang, Xiaolu Gan, Kaibiao Lin, Quan Zou, and Yifeng Zeng. Grouped correlational generative adversarial networks for discrete electronic health records. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 906–913. IEEE, 2019.

[270] Yinchong Yang, Zhiliang Wu, Volker Tresp, and Peter A Fasching. Categorical ehr imputation with generative adversarial nets. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–10. IEEE, 2019.

[271] Yun Yang, Fengtao Nan, Po Yang, Qiang Meng, Yingfu Xie, Dehai Zhang, and Khan Muhammad. Gan-based semi-supervised learning approach for clinical decision support in health-iot platform. *IEEE Access*, 7:8048–8057, 2019.

[272] Xin Yi and Paul Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669, 2018.

[273] Kejing Yin, Liaoliao Feng, and William K Cheung. Context-aware time series imputation for multi-analyte clinical data. *Journal of Healthcare Informatics Research*, 4:411–426, 2020.

[274] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics*, 24(8):2378–2388, 2020.

[275] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

[276] Jinsung Yoon, James Jordon, and Mihaela Schaar. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In *International Conference on Machine Learning*, pages 5699–5707. PMLR, 2018.

[277] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[278] Michael P Young, Valerie J Gooder, Karen McBride, Brent James, and Elliott S Fisher. Inpatient transfers to the intensive care unit. *Journal of general internal medicine*, 18(2):77–83, 2003.

[279] Kezi Yu, Yunlong Wang, Yong Cai, Cao Xiao, Emily Zhao, Lucas Glass, and Jimeng Sun. Rare disease detection by sequence modeling with generative adversarial networks. *arXiv preprint arXiv:1907.01022*, 2019.

[280] Nikolaos Zafeiropoulos, Argyro Mavrogiorgou, Spyridon Kleftakis, Konstantinos Mavrogiorgos, Athanasios Kiourtis, and Dimosthenis Kyriazis. Interpretable stroke risk prediction using machine learning algorithms. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2*, pages 647–656. Springer, 2023.

[281] Hongyang Zhang and David P Woodruff. Medical missing data imputation by stackelberg gan. *Carnegie Mellon University*, 2018.

[282] Xinmeng Zhang, Chao Yan, Cheng Gao, Bradley A Malin, and You Chen. Predicting missing values in medical data via xgboost regression. *Journal of healthcare informatics research*, 4:383–394, 2020.

[283] Y Zhang, Q Yang, Z-H Zhou, and L-L Chen. A survey on multi-task learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):2400–2421, 2021.

[284] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Unsupervised x-ray image segmentation with task driven generative adversarial networks. *Medical Image Analysis*, 62:101664, 2020.

[285] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.

[286] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. Synteg: a framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association*, 28(3):596–604, 2021.

[287] Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association*, 27(1):99–108, 2020.

[288] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[289] Yinghao Zhu, Zixiang Wang, Long He, Shiyun Xie, Zixi Chen, Jingkun An, Liantao Ma, and Chengwei Pan. Leveraging prototype patient representations with feature-missing-aware calibration to mitigate ehr data sparsity. *arXiv preprint arXiv:2309.04160*, 2023.

[290] Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657, 2007.