

To the Graduate Council:

I am submitting herewith a dissertation written by Senne Van Steelandt entitled "Generative Machine Learning Models in Healthcare Applications." I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Michel Ballings, Major Professor

We have read this dissertation
and recommend its acceptance:

Dr. Michel Ballings

Dr. Anahita Khojandi

Dr. Ephy Love

Dr. Sean Willems

Accepted for the Council:

Dixie I. Thompson

Vice Provost and Dean of the Graduate School

To the Graduate Council:

I am submitting herewith a dissertation written by Senne Van Steelandt entitled "Generative Machine Learning Models in Healthcare Applications." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Michel Ballings, Major Professor

We have read this dissertation
and recommend its acceptance:

Dr. Michel Ballings

Dr. Anahita Khojandi

Dr. Ephy Love

Dr. Sean Willems

Accepted for the Council:

Dixie I. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Generative Machine Learning Models in Healthcare Applications

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Senne Van Steelandt
May 2023**

© by Senne Van Steelandt, 2023
All Rights Reserved.

*né, sì chinato, lì fece dimora,
e come albero in nave si levò.*

DANTE ALIGHIERI, DIVINA COMMEDIA

Acknowledgements

I am immensely grateful to Dr. Michel Ballings for his guidance throughout my Ph.D. We first met at Ghent University when he was the visiting lecturer for the Social Media and Web Analytics I took as part of my Master's degree. Through this encounter, I was introduced to the Management Science Ph.D. program at the Haslam College of Business, University of Tennessee. The collaborative work we did on digital marketing and e-commerce was foundational for my understanding of academic writing and reinforcement learning. I am sincerely thankful for the invaluable opportunity he has provided me.

I am deeply grateful to Dr. Anahita Khojandi for her unwavering support throughout the course of my Ph.D. She generously agreed to advise me on my research on the use of reinforcement learning in intensive care units, and her patience and support enabled me to overcome the writer's block I had been struggling with for a significant part of my Ph.D. track. I cannot overstate her contribution to my work. I'd also like to extend my gratitude to Ben Smith and Dr. Vasudevan, with whom I had the pleasure to work together on my initial research on healthcare applications.

I would like to express my thanks to Dr. Russell Zaretski for introducing me to Dr. Ephy Love, Head of Data Science at Thalamus. This introduction has been invaluable to my Ph.D., as it enabled me to pursue my Curricular Practical Training at Thalamus and undertake research into topic modeling in Graduate Medical Education. Dr. Love has provided invaluable guidance on my second chapter and has been a great advisor, colleague, and friend. I am incredibly grateful for the opportunity that he has given me, and he has been an outstanding advisor on my research at Thalamus.

I also want to express my thanks to Dr. Jason Reminick, CEO, and co-founder of Thalamus, for providing a unique environment that supports and fosters conducting fundamental research in Graduate Medical Education. I'm convinced that the resulting open-source peer-reviewed research will have a positive and long-overdue impact on Graduate Medical Education in the United States. I am also thankful to everyone at Thalamus for their hard work, positive outlook, and dedication to improving the future of Graduate Medical Education.

I am immensely grateful to Dr. Arat for his generous hospitality in accommodating me during my first month in Knoxville, ever-willing to take me around the city to help me settle in. Not only was he a formidable host, colleague, and co-student, but he has also become a dear friend. Murat went above and beyond to support me during the first three years of my Ph.D., and it was with a heavy heart that we had to say our farewells when he graduated amidst the pandemic. I look forward to the day when I can travel again and visit him.

I'd like to extend my gratitude to Dr. Willems, the director of the Management Science Ph.D. program, for extending to me the opportunity to participate in this program and to attain my Ph.D. in Management Science. Additionally, I am thankful to Dr. Galbreth, the department head of Business Analytics and Statistics, and Christine Vossler for their assistance throughout the process of submitting my thesis. I am also appreciative of my colleagues within the department for creating a supportive research environment.

Most importantly, I would like to thank my beloved family. My wife Idil, whom I met while I was completing my Ph.D., and who has commiserated with me and supported me throughout this journey. I wouldn't have made it without you. I'd like to thank my parents, Rudi and Annemie, and my sister, Rilke, for supporting me as I made my journey to the United States, and I am grateful for their love and encouragement. Hopefully, I can travel again soon so I can visit them.

Abstract

In this doctoral dissertation, I develop two applications of generative machine learning models for healthcare settings. These applications are organized in two chapters. In the first chapter, I analyze the impact of imputation methods on the ability of deep generative models (DGMs) to estimate patient health dynamics in the Intensive Care Unit (ICU). Despite data-rich environments, ICU data often suffer from a considerable amount of missing information. I demonstrate how imputed values can lead to discrepancies between the DGM's loss (difference in the observed and imputed values) and the DGM's failure to accurately reconstruct patient health trajectories. I propose a novel imputation technique, Generative Iterative Multiple Imputation (GIMI), as an on-training imputation strategy with a masked loss function and encoded missingness patterns. I evaluate GIMI's ability to reconstruct patient health trajectories on a dataset of ICU patients receiving heparin treatment. I find that it outperforms benchmarks such as k-nearest neighbors imputation, MICE, and GAIN. I also find that GIMI considerably reduces the reconstruction error of the patient health trajectory.

In my second chapter, I explore the impact of historical selection bias on a corpus of 35,000 personal statements collected between 2019 and 2023 from applicants to Graduate Medical Education programs. I examine one outcome of this selection bias, which is a substantial distortion of the ethnic distribution of the US population. Word frequency and topic analyses uncover both differences and similarities among applicants from various ethnic backgrounds, which are manifested in the form of ethnicity-specific topics as well as topics that are shared by all applicants. Furthermore, I find that ethnic

diversity is suppressed when all ethnicities share the same topic distribution, particularly affecting Black and Latino applicants. In light of these findings, I propose a topic model that combines both shared and ethnicity-specific topics. I demonstrate that the topics generated by this model preserve ethnic diversity and are more coherent compared to those derived from models using exclusively shared or ethnicity-specific topic distributions.

Table of Contents

1	GIMI: Generative Iterative Multiple Imputation for Improving Health Trajectory Estimation	1
1.1	Abstract	2
1.2	Introduction	2
1.3	Background and Literature Review	5
1.3.1	DGMs for estimating patient health dynamics	6
1.3.2	Missing data techniques	7
1.3.3	Patient Health Transition Models Using Incomplete Datasets	8
1.4	Methodology	10
1.4.1	Addressing the imputation error	10
1.4.2	Imputation error in the output vector	11
1.4.3	Distribution shift at inference	12
1.4.4	Generative Iterative Multiple Imputation	13
1.5	Experimental Design	15
1.5.1	Data	15
1.5.2	Model setup	17
1.5.3	Benchmark methods	19
1.5.4	Evaluation methods	20
1.6	Results	22
1.6.1	Observed and Imputed Sample Statistics	22
1.6.2	Discrepancies in Accuracy between Observed and Imputed Values	25

1.6.3	Generative Iterative Multiple Imputation	27
1.6.4	Robustness Analysis: Data Collection Method and Model Performance	30
1.7	Conclusion	32
1.7.1	Future work and limitations	34
	Appendix	35
2	Preserving Ethnic Diversity in Topic Models for Residency Applications	37
2.1	Abstract	38
2.2	Introduction	38
2.3	Methods	41
2.3.1	Ethics statement	41
2.3.2	Data Collection	41
2.3.3	Data Cleaning and Text Pre-Processing	42
2.3.4	Descriptive study	42
2.3.5	Predictive study	43
2.4	Results	47
2.4.1	The Effect of Historical Selection Bias on Data Quality	47
2.4.2	Corpora Differ between Applicants with Different Ethnic Backgrounds	49
2.4.3	Ethnicity-Agnostic Topic Models Suppress Ethnic Variety	52
2.4.4	Ethnic Variety Preserves Semantic Relationships	57
2.5	Conclusion	63
	Appendix	66
	Bibliography	67
	Vita	81

List of Tables

1.1	Overview of missing data techniques in health trajectory estimation. . . .	9
1.2	Overview of variables by variable group.	18
1.3	Descriptive statistics for all variables in the dataset.	23
1.4	Comparison of Key Statistics for Heartrate, CO2, and Creatinine for each Imputed Dataset	24
1.5	MRMSE ₁ and MRMSE _{traj} for the VAE and Transformer models.	28
1.6	MRMSE ₁ for each hyperparameter setup with 16 latent dimensions tested during tuning of the VAE model.	35
1.7	MRMSE ₁ for each hyperparameter setup with 32 latent dimensions tested during tuning of the VAE model.	36
1.8	MRMSE ₁ for each hyperparameter setup tested during tuning of the Transformer model.	36
2.1	Hyperparameters for the ELDA models	66

List of Figures

1.1	Autoregressive input	14
1.2	Comparison of a) the $RMSE_1$ and $MRMSE_1$, and b) the $RMSE_{traj}$ and $MRMSE_{traj}$	26
1.3	$MRMSE_t$ by step for each imputation technique.	29
1.4	$MRMSE_t$ by step for each variable type.	31
2.1	a) Ethnic distribution at different stages of the physician pipeline and its effect on data quality. b) Proportion of personal statements, words, and unique words by ethnicity; note that, for unique words, the normalizing constant is less than the sum of unique words across groups (i.e., the ethnicity's unique words overlap), and therefore these proportions sum to greater than 100% while the proportions of words and statements are mutually exclusive and sum to 100%.	48
2.2	Most over- and under-represented terms for each ethnicity.	50
2.3	Topic distributions for the general and ethnicity-specific models	53
2.4	Comparison of the density estimated by the general model and the ethnicity-specific models for the four most important and shared topics.	55
2.5	Distances between the General Model Topics and the Ethnicity-Specific Model Topics.	56
2.6	Distributions of the Distances between the General Model Topics and the Ethnicity-Specific Model Topics.	58

2.7	Mean coherence scores for the topics of the ethnicity-specific, hybrid, and shared topic models, calculated for the top 30 terms.	60
2.8	Relative mean coherence scores for the topics of the ethnicity-specific, hybrid, and shared topic models, calculated for the top 5 to 30 terms. . .	61
2.9	Hybrid Model Topic Distributions for All Ethnicities.	62

Chapter 1

GIMI: Generative Iterative Multiple Imputation for Improving Health Trajectory Estimation

1.1 Abstract

Predicting the health trajectories of patients undergoing treatment in the Intensive Care Unit (ICU) is essential for developing data-driven treatment models. Deep generative models can be used to capture the underlying dynamics of these health trajectories, paving the way for the development of such treatment models. However, despite the data-rich environment of ICUs, often the collected data suffer from significant amounts of missing values. In this study, we analyze the impact of imputation methods on the ability of deep generative models to estimate these patient health dynamics. I describe and empirically validate how imputed values can lead to a discrepancy between the loss of the imputed values and the observed value, resulting in the model's failure to reconstruct the patient health trajectory accurately. To address this issue, I develop a novel imputation technique, namely Generative Iterative Multiple Imputation or GIMI. GIMI is developed as an on-training imputation strategy and uses a masked loss function and encoded missingness patterns to capture patient health dynamics. I evaluate the ability of GIMI to accurately reconstruct the patient health trajectory on a dataset of patients receiving heparin treatment in the ICU. I benchmark GIMI against state-of-the-art imputation methods, including sample-and-hold, KNN, MICE, and GAIN. Our results show that GIMI outperforms benchmarks and considerably reduces the reconstruction error of the predicted patient health trajectory.

1.2 Introduction

The intensive care unit (ICU) is a data-rich environment that provides a wealth of well-structured information on patient health, including vital signs, laboratory results, and medication dosing [Celi et al., 2013]. These data can be used as an alternative to —often costly and time-intensive— randomized clinical trials to study the safety, effectiveness, and outcomes of treatments and interventions [Black, 1996]. In particular, the data on patients' past treatment history, current health status, and response to

different treatments collected in electronic health records (EHRs) can be used to train machine learning (ML) models that accurately characterize the transitions of the patient health trajectory [Rajkomar et al., 2018]. Various mathematical methodologies, such as the Markov decision process (MDP) — the foundation of reinforcement learning (RL) algorithms, a powerful class of ML models — can effectively leverage EHRs to characterize such transitions. Specifically, MDP models capture the dynamics of the patient’s health over time by mapping the patient state (a vector with observed health indicators) and a treatment action to the patient’s next state in the trajectory, with the goal of ultimately identifying the optimal treatment strategy for patients. That is, these models are typically used to train personalized treatment models, such as dynamic treatment regimes (DTRs), to determine the best sequence of treatments for individual patients over time [Bennett and Hauser, 2013].

The state and action spaces that must be accounted for by these MDP models are extensive, necessitating an estimator with high expressivity and capacity to accurately model the relationship between the patient’s health dynamics and the administered treatment. Deep neural networks (DNNs) are generally considered to be the most appropriate choice [Sejnowski, 2020]. However, rather than using the more common discriminative DNNs to predict the new state given past state and actions using the conditional probability $P(X|Y)$, generative DNNs are preferable since they are able to capture the joint probability distribution $P(X,Y)$, allowing for the generation of new samples [Bennett and Hauser, 2013] and thereby decreasing the possibility of a distributional shift occurring between model training and inference [Levine et al., 2020].

The utility of these data-driven generative DNNs is subject to the practical limitations of the quality of corresponding data on which the models are developed [Beaulieu-Jones et al., 2018]. The ICU is a fast-paced, busy environment, and even though data are usually collected continuously from patients in this environment, it is not unusual to have missing data as a result of loose sensors or patients being moved for scans [Kramer et al., 2021]. Additionally, note that data are often collected at different frequencies depending on the corresponding collection procedure

[Fu et al., 2021], resulting in missing values for one feature or biomarker compared with another in the dataset. For example, heart rate can be monitored continuously, but the white blood cell count may only be measured every few hours or less frequently — this is mainly because heart rate can be easily monitored using non-invasive electrocardiographic or photoplethysmographic sensors while measuring white blood cell count requires a healthcare professional to draw a blood sample for further lab analysis. Note that regardless of the cause of data missingness, such omission values can cause difficulties when training models [Lizotte et al., 2008].

Before we consider the impact of these missing values on patient health transition models estimated by generative DNNs, we first examine the most common application of DNNs. That is, a discriminative model estimating point predictions for a target vector, Y , given an input vector, X . Both vectors are sampled from different data-generating processes, and the length of the target vector is typically considerably smaller — often equal to one — than the length of the input vector. The literature on data missingness techniques for these discriminative models typically focuses on the relationship that exists between the missingness patterns in the input vector and the value of the target vector (e.g., MCAR, MAR, and MNAR) rather than the missingness patterns in the target vector. Furthermore, missingness in the target vector is usually addressed through deletion as a result of the limited vector length.

In contrast, health transition models are used to find the mapping between past states and actions, and future states for a target vector, X_t , given an input vector, X_{t+1} [Srinivasan and Doshi-Velez, 2020]. Therefore, both the input and target vector are sampled from the same time series/data-generating process, and a strong correlation exists between the missingness patterns in both vectors. Furthermore, the transition model predicts a multi-dimensional health state for a patient resulting in a vector that is considerably lengthier and which can have both missing and observed values. As this length increases, the probability of a missing value in the target vector increases exponentially. Alternatively, discarding every observation where at least one value is missing in the target vector would typically result in a considerable loss of information.

In this study, we investigate imputation techniques for health transition models based on generative DNNs, focusing on improving the handling of missing data in such applications. Specifically, we propose Generative Iterative Multiple Imputation (GIMI), an on-training imputation technique that combines encoded missingness patterns and a masked loss function to accurately capture the dynamics of the patient health trajectory. This approach is unique in that, to the best of our knowledge, it is the first imputation technique specifically designed for patient health transition models based on DGMs. We showcase the performance of this approach using a longitudinal dataset of patients receiving heparin treatment in the ICU [Johnson et al., 2016]. I compare the results of our GIMI approach to those obtained using state-of-the-art imputation techniques, including k-nearest neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and Generative Adversarial Networks (GAIN). I find that GIMI outperforms all benchmarks, demonstrating the potential of GIMI to improve the handling of missing data in health transition models that use DGMs as estimators, providing more accurate predictions and enabling better decision-making in clinical settings.

The remainder of this manuscript is organized as follows. First, in Section 2, I review the relevant literature on DGMs and missing data techniques and their utilization in patient transition model applications. Next, in Section 3, motivated by the shortcomings of current imputation techniques, I propose GIMI, a novel imputation technique for DGM-based transition models. Section 4 then discusses the experimental design, including the dataset, the used DGMs, the benchmark techniques, and the evaluation methods. Next, the results of this experiment are discussed in Section 5. Finally, Section 6 concludes with practical implications, study limitations, and guidance for future work.

1.3 Background and Literature Review

In this section, our objective is to offer the reader a deeper understanding of DGMs and some imputation techniques, particularly in the context of patient health transition models. I begin by presenting an overview of the most prevalent DGM paradigms and

illustrating their application in predicting patient health trajectories. Next, I describe the most popular techniques for handling missing data. Finally, I review studies focused on estimating patient health transitions using incomplete datasets.

1.3.1 DGMs for estimating patient health dynamics

Although all DGMs share the common feature of modeling the joint probability distribution, $P(X, Y)$, various DGM families exist. Here, I provide an overview of the three most popular approaches developed for deep generative modeling, namely Variational Autoencoders (VAEs), Deep Autoregressive Models (DAMs), and Generative Adversarial Networks (GANs). I will briefly describe each approach and provide examples of their applications in healthcare.

VAEs [Kingma and Welling, 2013] utilize an encoder-decoder architecture. The input state vector is projected onto a latent dimension by one or more neural network layers (i.e., the encoder). Subsequently, this latent dimension is decoded through another neural network (i.e., the decoder) to the target state vector. VAEs have been employed to estimate patient health dynamics in several studies, such as Baicum et al. [Baicum et al., 2020], where a transitional VAE (tVAE) is proposed to map patient state-treatment pairs to the next patient state, and Allen et al. [Allen et al., 2021] who use a VAE to estimate patient health trajectories of patients experiencing an ischemic stroke.

DAMs are neural architectures designed to model sequential data by conditioning the prediction of each output element on previously generated elements. Key examples of DAMs include recurrent neural networks (RNNs) [Hochreiter and Schmidhuber, 1997], autoregressive convolutional networks [Hochreiter and Schmidhuber, 1997] [Oord et al., 2016], and autoregressive transformer models [Vaswani et al., 2017]. The use of an autoregressive model to generate patient health dynamics is exemplified by Phetrattikun et al. [Phetrattikun et al., 2021], who employ a Temporal Fusion Transformer to predict future vital signs' quantiles based on time-varying measurements of past vital signs.

GANs [Goodfellow et al., 2014] are comprised of two competing neural networks: a generator and a discriminator. The generator's objective is to create new data points, given a vector of random noise, while the discriminator's role is to classify these points as real or fake. In a continuous process of mutual improvement, the generator learns to deceive the discriminator, and the discriminator refines its ability to identify the authenticity of the generated data. A notable example of GANs in generating longitudinal patient health records is found in the work of Li et al. [Li et al., 2018], who proposed a GAN-based model named EHR-M-GAN to create synthetic mixed-type longitudinal EHRs.

1.3.2 Missing data techniques

I will now exhibit four prevalent methods for addressing missing data: deletion, imputation, interpolation, and modeling. Firstly, deletion, also known as complete case analysis, entails eliminating observations with missing values from the dataset [Little and Rubin, 2019]. This approach is mostly of interest because of its ease of use and computational simplicity. However, deletion can lead to the loss of valuable information and a considerable reduction in available data, especially in sparse datasets. Consequently, this technique is only feasible when missing values are sufficiently uncommon.

The second approach is imputation, a statistical method that addresses missing data in datasets by estimating and substituting missing values with suitable replacements. Numerous imputation techniques are available, spanning from elementary strategies, such as mean or forward imputation, to more sophisticated methods, like KNN imputation and MICE. The selection of an imputation technique hinges on factors including the data's structure, the missingness type (e.g., MCAR, MAR, and MNAR), and the intended application of the imputed dataset [Luengo et al., 2012].

Imputation methods can be broadly categorized into two groups: Simple Imputation (SI) and Multiple Imputation (MI) [Rubin and Schenker, 1986]. SI techniques utilize a single estimate, such as the sample mean, to replace missing values. Although these methods are straightforward and computationally inexpensive to implement, they do not

account for the uncertainty surrounding the missing value estimate. Consequently, the estimate's accuracy can be adversely impacted. In contrast, MI techniques replace missing values with a value sampled from a set of available estimates, considering the estimate's uncertainty. This can be achieved by sampling from a learned density function (e.g., GAIN) or through iterative sampling from a set of chained equations (e.g., MICE). While these approaches are more intricate and computationally costly, they offer enhanced flexibility in modeling missing values and are more resilient to data sparsity and outliers.

Interpolation is the third category and is, in fact, a special case of imputation where missing values are estimated by considering the neighboring data points in a structured manner. Typically, interpolation assumes a specific pattern or trend in the underlying data [Shahar, 1999]. In the context of time series data, this entails using both prior and subsequent values of the variable of interest to accurately estimate the missing data point. As a result, time series interpolation has limited practical value in real-world scenarios, where the model is typically forecasting into the future and, therefore, subsequent values are usually unknown.

Finally, missing data can also be explicitly modeled as part of the transition model. This approach is particularly interesting when the absence of data is not random and can provide valuable information about the underlying process, a phenomenon known as informative missingness [Groenwold, 2020]. A prominent example of missingness-modeling can be found in the literature on RNNs, where indicator variables are often used to signal to the model whether a value was observed or imputed.

1.3.3 Patient Health Transition Models Using Incomplete Datasets

In consideration of the aforementioned approaches for handling missing data, Table 1.1 offers an overview of dynamic treatment studies that estimate patient health dynamics using incomplete datasets. Each study is characterized by the method utilized to address missing values (SI, MI, interpolation, exclusion, or modeling) and whether it was applied before or during the model training phase. Further, if imputation was employed, the

Table 1.1: Overview of missing data techniques in health trajectory estimation.

Paper	Missing Method	Data	Pre/On-Training	Imputation Method
Nanayakkara et al., 2021 [Kim et al., 2021]	SI		Pre	Sample-and-hold
Roggeveen et al., 2021 [Roggeveen et al., 2021]	SI		Pre	KNN
Zheng et al., 2021 [Zheng et al., 2021]	SI		Pre	KNN
Tomkins et al., 2021 [Tomkins et al., 2008]	Exclusion		Pre	-
	SI		Pre	Zero imputation
Sun & Wang, 2021 [Sun and Wang, 2021]	MI		Pre	Gibbs sampling
Aguilera et al., 2020 [Aguilera et al., 2020]	SI		Pre	Sample-and-hold
Kahkoska et al., 2020 [Kahkoska et al., 2020]	MI		Pre	Gibbs sampling
Yu et al., 2020 [Yu et al., 2020]	SI		Pre	SVM
Yu et al., 2019 [Yu et al., 2019]	SI		Pre	SVM
Chang et al., 2018 [Chang et al., 2018]	SI		Pre	Mean imputation
Futoma et al., 2018 [Futoma et al., 2018]	Model		On	MGP
Jagannatha et al., 2018 [Jagannatha et al., 2018]	SI		Pre	Sample-and-hold
Komorowski et al., 2018 [Komorowski et al., 2018]	SI		Pre	KNN
	SI		Pre	Sample-and-hold
Li et al., 2018 [Li et al., 2018]	Interpolation		Pre	-
	SI		Pre	Sample-and-hold
Lin et al., 2018 [Lin et al., 2018]	SI		Pre	Mean imputation
	SI		Pre	Sample-and-hold
Peng, Xuefeng, et al., 2018 [Peng et al., 2018]	Exclusion		Pre	-
Raghu et al., 2018 [Raghu et al., 2018]	Exclusion		Pre	-
Tao et al., 2018 [Tao et al., 2018]	MI		Pre	Gibbs sampling
Krakov et al., 2017 [Krakov et al., 2017]	MI		Pre	Gibbs sampling
Prasad et al., 2017 [Prasad et al., 2017]	Model		On	-
Raghu et al., 2017 [Raghu et al., 2017b]	SI		Pre	KNN
Raghu et al., 2017 [Raghu et al., 2017a]	Exclusion		Pre	-
Weng et al., 2017 [Weng et al., 2017]	Interpolation		Pre	-
Ertefaie et al., 2016 [Ertefaie et al., 2016]	MI		Pre	Gibbs sampling
Murphy et al., 2016 [Murphy et al., 2016]	Exclusion		Pre	-
	Interpolation		Pre	-
Nemati et al., 2016 [Nemati et al., 2016]	Simple		Pre	Sample-and-hold
Marivate et al., 2014 [Marivate et al., 2014]	Interpolation		Pre	-
Parbhoo, 2014 [Parbhoo, 2014]	Interpolation		Pre	-
Chakraborty et al., 2013 [Chakraborty et al., 2013]	Exclusion		Pre	-
	SI		Pre	Sample-and-hold
	SI		Pre	KNN
Shortreed et al., 2011 [Shortreed et al., 2011]	Interpolation		Pre	-
Shortreed et al., 2011 [Shortreed et al., 2011]	MI		Pre	Gibbs sampling
Martín-Guerrero et al., 2009 [Martín-Guerrero et al., 2009]	Exclusion	2009	Pre	-
Gaweda et al., 2005 [Gaweda et al., 2005]	Interpolation		Pre	-

specific type is categorized into one of the following approaches: zero, mean, sample-and-hold, KNN, Gibbs sampling, Multivariate Gaussian Process (MGP), or Support Vector Machine (SVM) imputation.

Simple imputation is the most frequently employed technique, accounting for 48.48%, followed by exclusion and interpolation, each contributing 21.21%, and finally, multiple imputation, comprising 18.18%. Among the studies that utilized either simple or multiple imputation, the most popular techniques include sample-and-hold, with 34.78%, Gibbs sampling, with 26.09%, and KNN, with 21.74%. It is noteworthy that all imputation techniques presented in this table are performed prior to model training. Futoma et al. [Futoma et al., 2018] present a unique case in which no imputation is conducted, but missingness dynamics are learned concurrently with patient health dynamics using a Multi-output Gaussian process.

1.4 Methodology

In this methodology section, I first investigate how pre-training imputation affects a dataset and leads to imputation errors in both input and target vectors of DGMs. Subsequently, I introduce encoded missingness patterns and a masked loss function to address these errors. Next, I discuss the necessity for on-training imputation to mitigate a distributional shift that occurs between the training and inference stages of the transition model. Finally, I conclude this section by incorporating these methods into the GIMI algorithm.

1.4.1 Addressing the imputation error

First, I will examine how imputation errors, denoted as ϵ_t , can impact both the input and target vector in transition models. I define the imputation error vector as the difference between the oracle input vector, s_t^* , that we would observe if no values were missing and the imputation input vector, s_t^+ , where each value is estimated by an imputation method:

$$\epsilon_t = s_t^* - s_t^+. \quad (1.1)$$

Furthermore, let $\odot$ denote the Hadamard product and m_t the binary mask vector that is zero if the value was observed and one otherwise. Then I can define the imputed input state vector s_t^i as:

$$s_t^i = s_t^+ \odot m_t + s_t^* \odot (1 - m_t). \quad (1.2)$$

Finally, let $D()$ denote the transition function of the patient health trajectory. The target vector, s_{t+1}^i is then defined as:

$$s_{t+1}^i = D(s_t^i).$$

By substituting the imputed state vector, s_t^+ , with equation (1.2), and the imputation state vector with equation (1.1), I derive the following expression for the relationship between the imputed values and the health trajectory dynamics model:

$$(s_{t+1}^* + \epsilon_{t+1}) \odot m_{t+1} + s_{t+1}^* \odot (1 - m_{t+1}) = D((s_t^* + \epsilon_t) \odot m_t + s_t^* \odot (1 - m_t)). \quad (1.3)$$

This equation can be further simplified to

$$s_{t+1}^* + \epsilon_{t+1} \odot m_{t+1} = D(s_t^* + \epsilon_t \odot m_t), \quad (1.4)$$

which illustrates the presence of an imputation error in both the input and target vectors on which the DGM is trained.

1.4.2 Imputation error in the output vector

Having demonstrated that imputation errors exist in both the input and target vectors of the transition model, I propose two techniques to mitigate their impact. First, I address

the imputation error in the target vector, drawing inspiration from the use of indicator variables in RNNs [Rodenburg et al., 2019]. More specifically, I suggest imputing missing values with zeros and introducing an indicator variable for each feature, signaling to the model whether the value was observed or not. This method enables the model to learn from both the observed values and missingness patterns, as the absence of a variable can offer valuable insights into the patient’s subsequent state — a concept known as informative missingness [Groenwold, 2020]. Additionally, I expand the input vector across multiple time periods to maintain the Markovian property of our state. Due to the zero imputation employed, this expansion is necessary as each time period now contains information only from the values observed during that specific period.

Second, to address the imputation error in the target vector, I employ a masked loss function. In this approach, unobserved values in the target vector are masked using an indicator vector mask, ensuring that they do not contribute to the model loss. Let the root mean squared error (RMSE) be the loss function and let N denote the number of observations for which we calculate this loss. Then the following equation represents the masked loss function:

$$\sum_{n=1}^N (\hat{s}_{t+1,n} - s_{t+1,n})^2 \otimes (1 - m_{t+1,n}). \quad (1.5)$$

1.4.3 Distribution shift at inference

Building upon our earlier discussion of predicting the next state of a patient, I now consider the challenge of reconstructing the complete trajectory during the inference stage, where we only have the true state at the beginning of the trajectory. During the training phase, the model is exposed to observed patient states from the dataset, with the missingness patterns preserved. Consequently, the transition dynamics learned by the model are based on these incomplete observations, where missing values are represented as zeros.

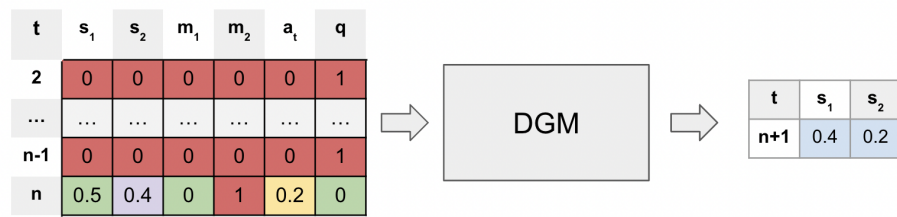
However, at the inference stage, the model relies on its own predicted states to generate subsequent estimates. This difference in input data —observed, incomplete states during training and predictions, complete states during inference— results in a distributional shift. The model may not have learned the underlying patterns and relationships that exist when using its own predicted states, leading to potential inaccuracies and degraded performance when generating patient health trajectories at the inference stage.

Figure 1.1 illustrates this phenomenon. The first subfigure represents the input vector for the DGM in both the training and inference phase, where we start the trajectory at time $t = 3$. In this case, the input vector for both phases is the same, and no issue occurs. However, when we progress to the next step in the trajectory, $t = 4$, we observe a discrepancy between the input vectors in the training and inference phases. For the training phase, the new input vector is deterministically drawn from the zero-imputed dataset, where only the observed values have non-zero entries. In contrast, for the inference phase, the new input vector is drawn from the DGM's output, which generates estimates for each variable, rather than taking an input vector interspersed with missing values.

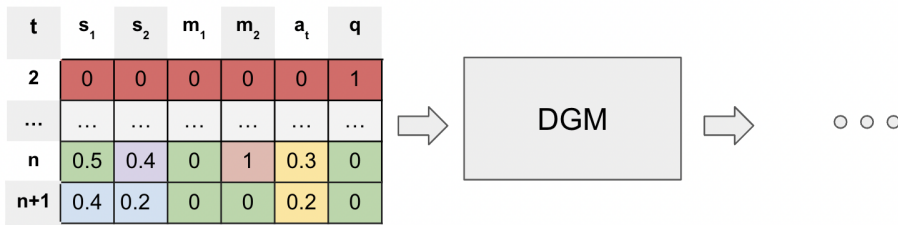
To address this distributional shift, I implement a strategy that involves imputing values during the estimation stage rather than imputing missing values prior to estimation. Initially, we set all missing values to zero, and then, throughout each training epoch, we update these missing values with the predictions generated by the DGMs. This approach ensures that the model learns to make predictions based on its own generated estimates, thus minimizing the distributional discrepancy between the training and inference stages.

1.4.4 Generative Iterative Multiple Imputation

Drawing from the methods presented in the previous three subsections, we now present the Generative Iterative Multiple Imputation (GIMI) Algorithm 1 as a superior means of imputing missing values in a DGM-based transition model. GIMI uses encoded missingness patterns to minimize imputation errors in the input vector, a masked loss function to



(a) t=1



(b) t=2

Figure 1.1: Autoregressive input

mitigate imputation errors in the target vector, and on-training imputation to address the distribution shift between the training and inference phase.

More specifically, given a dataset T with N observations, a batch size b , and a deep generative model (DGM) D , GIMI proceeds through the following steps: First, an imputation dataset, denoted by $\hat{S}$, is initialized to zero, where the on-training estimates of the DGM are stored. Then, in each iteration, GIMI randomly selects a batch of b input vectors, input vector masks, target vectors, and target vector masks from T , as well as the corresponding values from $\hat{S}$. Next, the input vector for the next timestep, $\hat{s}_{t+1}$, is sampled from the DGM, conditioned on s_t and m_t . The imputation dataset $\hat{S}$ is then updated with the newly generated input vector $\hat{s}_{t+1}$. In the final step of each iteration, a gradient descent update is performed on a masked loss function that filters out missing values from the target vector. Finally, GIMI repeats steps 2-6 until convergence is achieved, which is defined as the point where the average per-sample error does not improve by more than a predefined threshold.

1.5 Experimental Design

I now outline an experiment that assesses the accuracy of GIMI in reconstructing patient health trajectories of ICU patients receiving heparin treatment. I present an overview of the dataset and the DGMs utilized in this experiment. Additionally, I detail the imputation methods that serve as benchmarks and the metrics used to evaluate each approach.

1.5.1 Data

Our dataset was derived from MIMIC-III v1.4, which is a publicly available database containing the data of over 38,000 patients admitted to the intensive care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [Johnson et al., 2016]. Our experiment focused on patients who received heparin treatment during their ICU

Algorithm 1: Generative Iterative Multiple Imputation (GIMI)

Input : Dataset T , batch size b , DGM D

- 1 Initialize imputation dataset $\hat{S}$ to zero **do**
 - 2 **sample** $s_t, m_t, s_{t+1}, m_{t+1}, \hat{s}_t$ from $T, \hat{S}$
 - 3 $s_t \leftarrow s_t \otimes (1 - m_t) + \hat{s}_t \otimes m_t$
 - 4 $\hat{s}_{t+1} \sim D(s_t, m_t)$
 - 5 **update** $\hat{S}$ with $\hat{s}_{t+1}$
 - 6 Perform a gradient descent step on $\sum_{i=1}^N (\hat{s}_{t+1,i} - s_{t+1,i})^2 \otimes (1 - m_{t+1,i})$
 - 7 **until convergence**
-

stay. Specifically, I selected the first 30 hours for each patient who had received heparin treatment during their stay. For each time-variant variable, the values were calculated as the mean of all the values observed during the hour of recording. A comprehensive list of all variables and their descriptions can be found in Table 1.2. After applying these criteria, I extracted data from a total of 2,503 patients, resulting in a total of 75,090 observations.

The variables in the dataset are categorized into three different groups: state variables are the time-variant variables that describe the patient’s health state, such as the patient’s heart rate and hematocrit values. Covariates are the time-invariant variables that describe the patient’s health state, such as the patient’s age, weight, and height at admission. Finally, the treatment variable, Heparin, is the hourly intravenous heparin infusion dose that the patient receives as treatment during their stay at the ICU.

The state variables can be further categorized based on their frequency of observation, with the most frequently observed signals being vital signals, such as heart rate, temperature, and SaO₂, which are continuously monitored and recorded at regular intervals, ranging from seconds to minutes depending on the monitoring device and patient’s condition. Next in frequency are the observational signals, such as the (qualitative) Glasgow Coma Score, which requires human intervention for collection. The frequency of observation for these signals depends on the frequency of assessment by the clinical team. Finally, the least frequently observed signals are lab signals, including the partial pressure of arterial carbon dioxide, serum creatinine level, hematocrit level, hemoglobin level, INR, blood platelet count, PT, and white blood cell count, which require (invasive) collection, laboratory analysis, and collation.

1.5.2 Model setup

In this experiment, we implement two different families of DGMs to reconstruct patient health trajectories: the transformer model, representing autoregressive models, and the VAE model. For both, we follow the configuration described in the original papers by Kingma et al. [Kingma and Welling, 2013] and Vaswani et al. [Vaswani et al., 2017],

Table 1.2: Overview of variables by variable group.

Type	Name	Description
Covariates	Age	Age of the patient at the time of admission.
	Ethnicity	Ethnicity of the patient at the time of admission.
	Gender	Gender of the patient at the time of admission.
	Weight	Weight of the patient measured at the time of admission.
	Height	Height of the patient measured at the time of admission.
State	PaCO ₂	Partial pressure of arterial carbon dioxide.
	Heart rate	Heart rate.
	Creatinine	Serum creatinine level.
	GCS	Glasgow Coma Scale.
	Hematocrit	Hematocrit level.
	Hemoglobin	Hemoglobin level.
	INR	International Normalized Ratio.
	Platelets	Blood platelet count.
	PT	Prothrombin time.
	SaO ₂	Arterial oxygen saturation.
	Temperature	Body temperature.
	WBC	White blood cell count.
Treatment	Heparin	Intravenous heparin dosage.

respectively, with the exception of network size. This hyperparameter is optimized for each dataset-model combination to accommodate for the differences in optimal capacity. The details and results for the hyperparameter tuning are provided in Tables 1.6 and 1.7 for the VAE model, and Table 1.8 for the transformer model (which are all located in the Appendix).

Before dataset imputation and model training, we reserve a hold-out validation sample of 376 patient trajectories. The remaining 2,127 patients, with a total of 63,810 observations, are used for hyperparameter tuning and training, utilizing 10-fold cross-validation. We use a fixed optimizer, Adam [Kingma and Welling, 2013], and a fixed batch size of 64. Additionally, we use an adaptive learning rate schedule that initializes at 10^{-3} and decays by a factor of 0.1 when the test loss has not declined for at least five subsequent epochs. We stop model training early when the model reaches a learning rate of 10^{-8} .

1.5.3 Benchmark methods

The following imputation techniques are used as benchmarks:

- **Mean imputation:** This SI method replaces missing values with the mean value of the non-missing values for the respective variable and patient. It is easy to implement, computationally efficient, and commonly used as a baseline method [Donders et al., 2006]. However, it fails to utilize the information contained within the dataset to generate more accurate estimates.
- **Forward imputation:** This SI method replaces missing values by the last known value in time for the respective variable and patient. If no value was observed prior to the data point that needs to be imputed, then mean imputation is used. This is a popular baseline method for time-series data, and it intuitively aligns with the human, heuristic process of handling missing temporal data. Furthermore, while leveraging the temporal structure of the variables, it does not utilize the correlations between variables in the dataset [Schafer and Graham, 2002].

- **KNN imputation:** This SI method replaces missing values by the mean of its k closest neighbors, as measured by the distance between their non-missing values. It is less sensitive to outliers and utilizes the correlations between features in the dataset. However, it requires us to make a choice of k and the used distance metric, and can be computationally expensive when dealing with large datasets [Schafer and Graham, 2002].
- **Multivariate Imputation By Chained Equations (MICE)** This MI method uses chained linear regressions to iteratively impute missing values. MICE can efficiently account for the relationships between variables but assumes that the imputation model is correctly specified, and it may produce biased estimates if model assumptions are violated [Van Buuren and Oudshoorn, 1999].
- **Generative Adversarial Imputation Nets (GAIN):** This MI method uses a GAN to iteratively impute missing values, where the generator network, given the observed values, generates the imputed values and the discriminator network evaluates the quality of the imputations. GAIN is relatively robust to outliers compared with other methods and is capable of leveraging both the temporal structure of the dataset, as well as the correlation between features. However, it is data-inefficient, computationally expensive, and requires careful tuning of hyperparameters [Yoon et al., 2018].
- **No-GIMI:** This is a variation on GIMI where we keep the missingness patterns and the masked loss function, but we omit the on-training imputation. I introduce No-GIMI as an additional point of comparison to illustrate how the on-training imputation alone stabilizes the reconstructed health trajectory.

1.5.4 Evaluation methods

I evaluate the model’s performance for each patient in the validation set by autoregressively reconstructing their health trajectory. More specifically, for each patient in the validation set, I sample a starting state, $s_{t=w}$, from the dataset, where w is the

timestep in the trajectory at which we begin reconstructing the patient health trajectory. This is done to ensure the model can observe sufficient non-missing values. Subsequently, for each starting state, I sample a treatment variable, $a_{t=w}$, from the policy used in the dataset. $s_{t=w}$ and $a_{t=w}$ are then utilized as input vectors for the DGM, sampling the subsequent state $\hat{s}_{w+1}$ from this model. I then employ this predicted subsequent state, $\hat{s}_{w+1}$, along with the corresponding action, $a_{t=w+1}$, sampled from the dataset, as input for predicting the following step in the patient trajectory. This procedure is repeated until the entire patient trajectory has been predicted.

Throughout this study, I use four metrics to evaluate the reconstructed patient health trajectory. I Let s^i_i denote the target vector where the missing values are imputed and $\hat{s}^i_i$ the target vector estimated by the DGM. Then, I define the RMSE, RMSE_t , at each timestep, t , of the reconstructed patient health trajectory:

$$\text{RMSE}_t = \frac{1}{N} \sum_{i=1}^N (s^i_i - \hat{s}^i_i)^2.$$

Next, I calculate the masked root mean squared error, MRMSE_t , a modification of the RMSE_t that considers only observed values. By masking missing values in the target vector using a target vector mask, m_i , the loss is calculated exclusively for the observed values. The MRMSE_t is defined as:

$$\text{MRMSE}_t = \frac{1}{N} \sum_{i=1}^N (s^i_i - \hat{s}^i_i)^2 \otimes (1 - m_i).$$

Furthermore, I define the $\text{RMSE}_{\text{traj}}$ and the $\text{MRMSE}_{\text{traj}}$ as the average RMSE_t and MRMSE_t , respectively, of all steps in the reconstructed patient health trajectory. Let T represent the number of steps in the trajectory. I can define $\text{RMSE}_{\text{traj}}$ as

$$\text{RMSE}_{\text{traj}} = \frac{1}{T} \sum_{t=1}^T \frac{1}{N} \sum_{i=1}^N (s^i_i - \hat{s}^i_i)^2,$$

and the $\text{MRMSE}_{\text{traj}}$ as:

$$\text{MRMSE}_{\text{traj}} = \frac{1}{T} \sum_{t=1}^T \frac{1}{N} \sum_{i=1}^N (s^i_i - \hat{s}^i_i)^2 \otimes (1 - m_i).$$

1.6 Results

In this section, I first provide a brief analysis of the dataset of ICU patients utilized in our study and explore the differences in key statistics between the imputed and observed datasets. I then discuss the observed discrepancies between the MRMSE and the RMSE measures. Subsequently, I present the outcomes of the proposed algorithm, GIMI, by examining its ability to reconstruct the initial patient health trajectories compared to benchmark imputation methods. Finally, I assess how this performance differs for variables acquired through distinct data collection techniques.

1.6.1 Observed and Imputed Sample Statistics

Table 1.3 presents the number of observations (Obs), mean, median, standard deviation (Std), maximum (Max), and minimum (Min) values for each variable. There are substantial differences in the number of observed values for each variable, ranging from as few as 3,560 (4.66%) for creatinine to 68,249 (89.32%) for heart rate. The patient population primarily comprises older patients, with an average age of 66.88 and a median age of 69.2. The average heparin dose was 765.11, while the median was 522.1, as patients did not receive heparin every day during their stay.

We illustrate the impact of each imputation technique on the key measures of vital, observational, and lab signals in Table 1.4. Specifically, we examine the heart rate, CO2 level, and creatinine level in the training set and compare their mean and standard deviation after imputation to the observed values for each variable. The results indicate that most imputation techniques perform well in preserving the mean of each variable, with the exception of GAIN for all three models. Additionally, KNN and MICE show limitations when applied to the arterial carbon dioxide signal. MICE is the only method that expands the range of each variable, while all imputation methods underestimate the standard deviation, particularly for variables with more missing values. Overall, forward imputation is the most effective method for retaining the original statistics.

Table 1.3: Descriptive statistics for all variables in the dataset.

	Variable	Obs	Mean	Median	Std	Max	Min
State variables	CO2	6251	25.68	25.0	6.112	98.0	9.0
	Heartrate	68249	86.50	85.0	18.295	200.0	0.0
	Creatinine	3560	1.87	1.2	4.426	180.0	0.1
	GSC	20172	12.20	15.0	3.678	15.0	3.0
	Hematocrit	6544	31.44	30.8	4.845	58.0	0.0
	Hemoglobin	5069	10.75	10.5	1.766	43.1	5.5
	INR	5603	1.55	1.4	1.409	66.1	0.8
	Platelet Count	4440	222.50	201.0	117.366	1047.0	11.0
	PT	5606	15.54	14.5	4.660	150.0	10.7
	SaO2	3012	96.42	97.0	4.298	108.0	24.0
	Temperature	16247	98.47	98.4	2.678	104.8	0.0
	WBC	4203	12.54	11.4	7.069	129.0	0.1
	aPTT	8296	68.77	58.9	34.712	158.4	0.0
Covariates	Age	76410	66.88	69.2	14.397	89.0	18.0
	Weight	72120	84.47	80.0	26.771	295.0	32.5
	Height	46020	170.44	170.2	19.515	464.8	13.0
Treatment Variable	Heparin	76410	765.11	522.1	58.900	4000.0	0.0

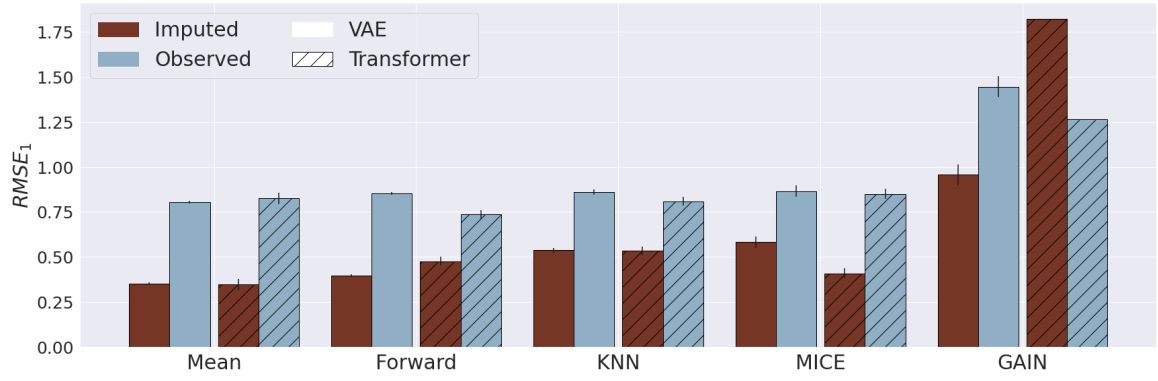
Table 1.4: Comparison of Key Statistics for Heartrate, CO2, and Creatinine for each Imputed Dataset

	Heartrate			Co2			Creatinine		
	Mean	Std	Range	Mean	Std	Range	Mean	Std	Range
Observed	0.000	1.000	9.738	0.000	1.000	7.231	0.000	1.000	6.806
Mean	-0.003	0.941	9.738	0.000	0.283	7.231	0.000	0.217	6.806
Forward	-0.013	0.998	9.738	0.042	0.670	7.231	-0.023	0.711	6.806
KNN	-0.001	0.942	9.738	0.093	0.577	7.231	-0.028	0.434	6.806
MICE	-0.003	0.413	10.567	0.150	0.944	9.945	0.023	0.325	9.960
GAIN	0.024	0.944	9.738	-0.293	0.404	7.231	0.222	0.648	6.806

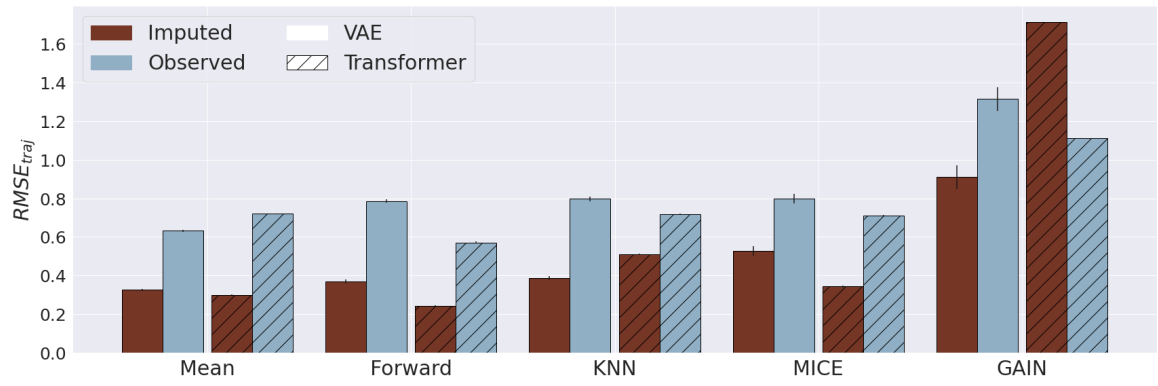
1.6.2 Discrepancies in Accuracy between Observed and Imputed Values

Figure 1.2.a presents a comparison of the $RMSE_1$ and the $MRMSE_1$ for each imputed dataset (mean, forward, KNN, MICE, or GAIN imputation) and model type (VAE or transformer). The confidence intervals are depicted by the black lines in the center of each bar. For all dataset and model type combinations, the confidence intervals between the $RMSE_1$ and the $MRMSE_1$ do not overlap, indicating significantly distinct errors. Excluding the GAIN imputation approach, the $MRMSE_1$ is underestimated by 86% for the VAE model and by 94% for the Transformer model in comparison to the $RMSE_1$. Furthermore, when assessing the relative performance across different models, disparities in rankings are observed between the $RMSE_1$ and the $MRMSE_1$. For example, the combination of mean imputation and the transformer model achieves the highest ranking based on the $RMSE_t$, whereas the optimal model selected according to the $MRMSE_1$ is the combination of forward imputation and the transformer model.

Figure 1.2.b presents a comparison of the $RMSE_{traj}$ and the $MRMSE_{traj}$ for each imputed dataset (mean, forward, KNN, MICE, or GAIN imputation) and model type (VAE or transformer). Similar patterns to those in Figure 1.2.a are observed, with statistically significant differences between the $RMSE_{traj}$ and the $MRMSE_{traj}$ of each dataset-model combination. Additionally, the $MRMSE_{traj}$ is underestimated compared to the $RMSE_{traj}$ by a factor of 72% and 71% for the VAE and transformer models, respectively (again excluding GAIN). Moreover, the relative ranking among different dataset-model combinations varies between the $RMSE_{traj}$ and the $MRMSE_{traj}$. For instance, the combination of mean imputation and the transformer model combination ranks first for the $RMSE_{traj}$, while it holds the third position for the $MRMSE_{traj}$. The analyses in Figures 1.2.a and 1.2.b reveal consistent underestimation of the $MRMSE$ in comparison to the $RMSE$, indicating that the traditional $RMSE$ might not adequately represent the performance of models on observed values in the presence of a considerable number of imputed missing values, as is the case in our study. Furthermore, by utilizing



(a) (M)RMSE₁



(b) (M)RMSE_{traj}

Figure 1.2: Comparison of a) the RMSE₁ and MRMSE₁, and b) the RMSE_{traj} and MRMSE_{traj}.

the MRMSE, we can ensure a more reliable assessment of model performance, accounting for the unique characteristics of datasets with substantial imputed values.

1.6.3 Generative Iterative Multiple Imputation

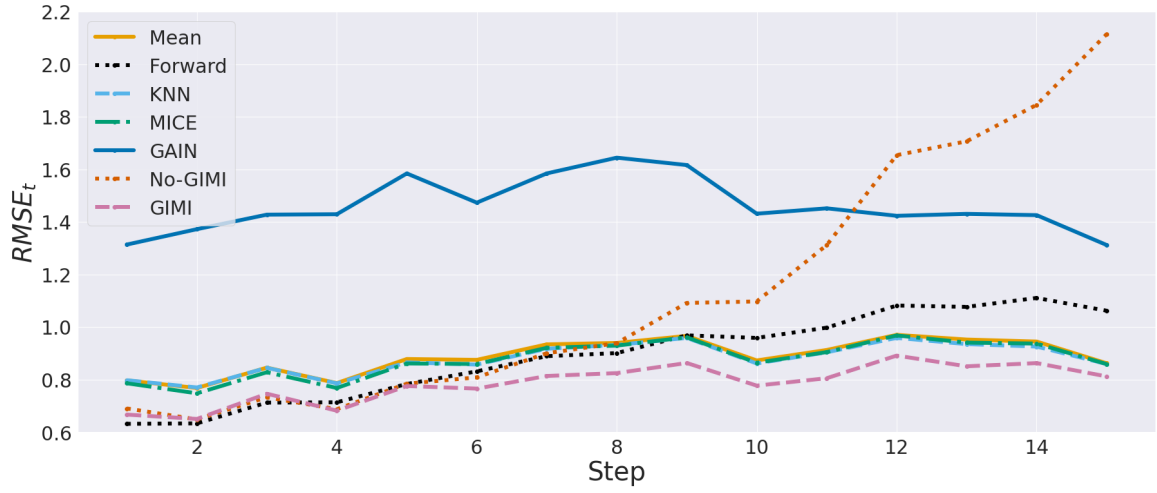
We can now compare the performance of GIMI to the benchmark methods described in the previous section using the MRMSE. Table 1.5 presents the MRMSE_1 and $\text{MRMSE}_{\text{traj}}$ for both the VAE and transformer models, with standard deviations provided in parentheses. The combination of GIMI and the Transformer model attains the lowest overall MRMSE_1 , with 0.5675. For the VAE model, the lowest MRMSE_1 of 0.6317 is achieved by forward imputation. Both models interchange positions for the second place. The mean, KNN, and MICE imputation techniques all exhibit similar performance with MRMSE_1 scores of 0.7970, 0.7977, and 0.7860 for the VAE model, and 0.7204, 0.7169, and 0.7121 for the transformer model, respectively. Finally, GAIN is the worst performing model with MRMSE_1 scores of 1.3138 and 1.1104 for the VAE and transformer models, respectively.

For the $\text{MRMSE}_{\text{traj}}$, GIMI achieves the highest performance for both the VAE and transformer models, with scores of 0.7858 and 0.6905, respectively. MICE is the best-performing benchmark technique for the VAE model, with a score of 0.8757, while forward imputation achieves the highest score for the transformer model, with 0.8124. The $\text{MRMSE}_{\text{traj}}$ of the remaining benchmark techniques demonstrate patterns consistent with the MRMSE_1 . For the VAE model, mean, forward, and KNN imputation achieve similar performance, with scores of 0.8870, 0.8901, and 0.8782, respectively. For the transformer model, mean, KNN, and MICE imputation also achieve similar performance, with scores of 0.8791, 0.8495, and 0.9371, respectively. Again, GAIN is the worst-performing technique, with scores of 1.4613 and 1.2988 for the VAE and Transformer models, respectively.

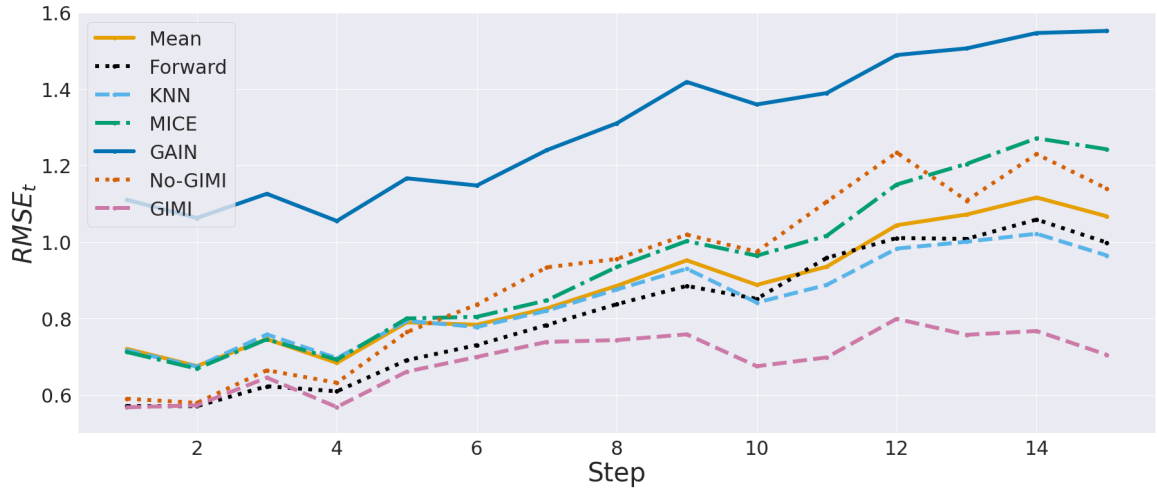
Figure 1.3.a displays the MRMSE_t score at each step of the reconstructed health trajectory for the VAE model, using various imputation techniques. Forward imputation, No-GIMI, and GIMI exhibit similar and the lowest MRMSE_t at the beginning of the

Table 1.5: MRMSE_1 and $\text{MRMSE}_{\text{traj}}$ for the VAE and Transformer models.

	VAE		Transformer	
	MRMSE_1	$\text{MRMSE}_{\text{traj}}$	MRMSE_1	$\text{MRMSE}_{\text{traj}}$
Mean	0.7970 (0.0185)	0.8870 (0.0201)	0.7204 (0.0042)	0.8791 (0.0590)
Forward	0.6317 (0.0162)	0.8901 (0.0321)	0.5710 (0.0020)	0.8124 (0.0157)
KNN	0.7977 (0.0159)	0.8782 (0.0173)	0.7169 (0.0060)	0.8495 (0.0420)
MICE	0.7860 (0.0151)	0.8757 (0.0094)	0.7121 (0.0044)	0.9371 (0.0479)
GAIN	1.3138 (0.0522)	1.4613 (0.0740)	1.1104 (0.4246)	1.2988 (0.3344)
No-GIMI	0.6902 (0.0219)	1.1341 (0.1000)	0.5901 (0.0113)	0.9178 (0.0196)
GIMI	0.6668 (0.0229)	0.7858 (0.0076)	0.5675 (0.0079)	0.6905 (0.0130)



(a) VAE



(b) Transformer

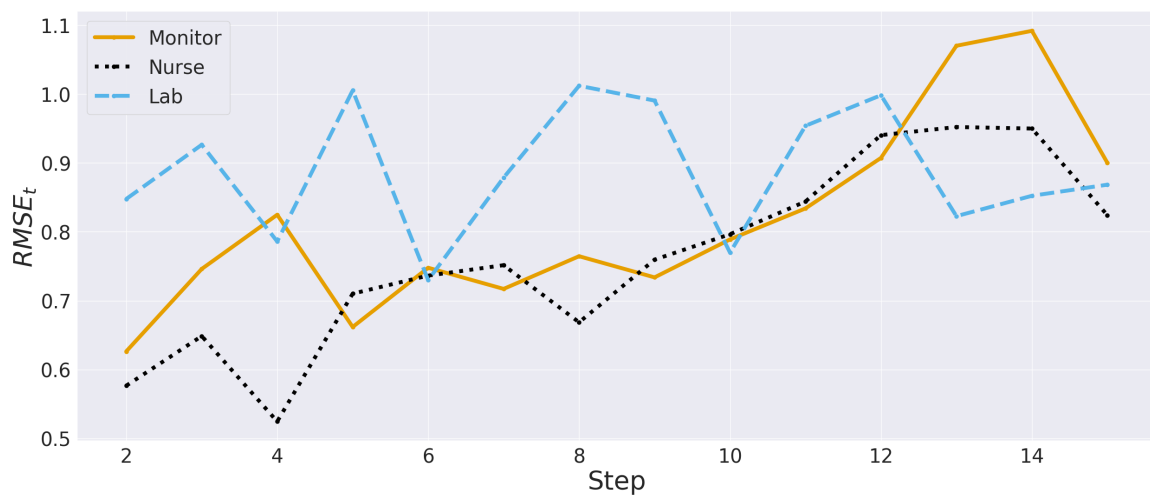
Figure 1.3: MRMSE_t by step for each imputation technique.

trajectory. Meanwhile, mean, KNN and MICE demonstrate comparable performance but underperform relative to the previous three methods. GAIN performs substantially worse compared to the other imputation techniques. The MRMSE_t increases over the trajectory for each imputation technique, with the exception of GAIN. This could suggest that GAIN's results are erroneous. Mean, KNN and MICE exhibit a similar increase in MRMSE_t . However, for forward imputation, No-GIMI, and GIMI, there are considerable differences between the three imputation methods. The error for both forward imputation and No-GIMI rises significantly over the trajectory, while GIMI's performance remains relatively stable and emerges as the best-performing technique, starting from the fourth step.

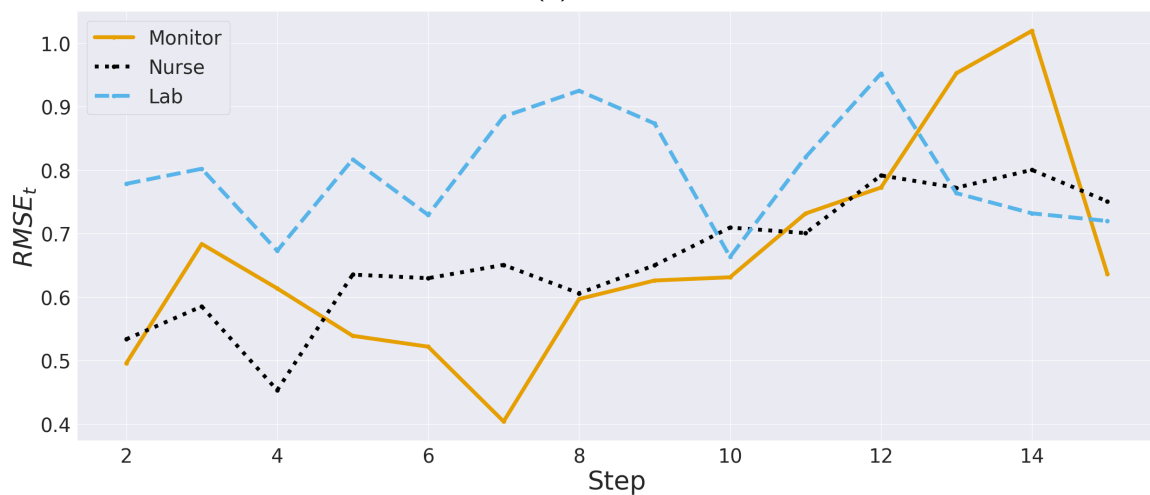
Figure 1.3.b presents the MRMSE_t score at every step of the reconstructed health trajectory for each imputation technique for the transformer model. Similar patterns are observed compared to the VAE model. Forward imputation, No-GIMI, and GIMI all achieve comparable scores and the lowest MRMSE_t at the beginning of the trajectory. Mean imputation, KNN imputation, and MICE exhibit similar performance but underperform compared to the previous three methods. GAIN performs notably worse than the other imputation techniques but shows improved performance compared to its performance in the VAE model. The degradation of forward and No-GIMI imputation is less pronounced but still significant. GIMI emerges as the best-performing technique at each step of the trajectory, with the exception of steps 2 and 3. The MRMSE_t for GIMI remains stable throughout. Relative to the VAE model, the MRMSE_t increases more. Mean imputation, KNN imputation, and MICE demonstrate comparable trajectories.

1.6.4 Robustness Analysis: Data Collection Method and Model Performance

Here, we calculate the MRMSE_t for each step in the trajectory for each signal category (vital, observational, and lab) to assess the models' performance under varying degrees of sparsity. The results for the VAE model are depicted in Figure 1.4.a. For all signal types,



(a) VAE



(b) Transformer

Figure 1.4: $MRMSE_t$ by step for each variable type.

we observe an increase in MRMSE_t over the trajectory. Furthermore, the difference in MRMSE_t between the vital and observational signals, as well as the lab signals, diminishes as the trajectory advances. Figure 1.4.b presents the results for the transformer model. We observe that the MRMSE_t increases over the trajectory for all signal categories, although this increase is less pronounced compared to the VAE model. Similarly, the gap between the MRMSE_t of the vital and observational signals and the lab signals narrows as the trajectory progresses.

1.7 Conclusion

In this study, my goal was to develop and assess imputation techniques for health transition models based on DGMs using ICU data. My motivation lies in addressing the challenges of missing data, which are common in ICUs due to factors such as loose sensors, patient movement, and varying data collection frequencies for costly measurements. Furthermore, these health transition models play a crucial role in MDP and RL models, ultimately contributing to enhanced medical decision-making by enabling more accurate and personalized treatment strategies for patients and therefore deserve further attention and development.

Using a dataset of patients receiving heparin in the ICU, we provided empirical evidence indicating that the RMSE_t for imputed datasets is not an accurate measure for the MRMSE_t . Both VAE and Transformer models exhibited substantially lower RMSE_1 and $\text{RMSE}_{\text{traj}}$ compared to MRMSE_1 and $\text{MRMSE}_{\text{traj}}$. Our investigation showed that, with the exception of GAIN, the RMSE_t and $\text{RMSE}_{\text{traj}}$ considerably underestimated MRMSE_t and $\text{MRMSE}_{\text{traj}}$ for all other imputation techniques. Furthermore, the relative rankings of dataset-model combinations differ between the RMSE and the MRMSE metrics. These observations indicate that using the RMSE with imputed target vectors is an inadequate performance measure for evaluating the ability of DGM-based transition models to accurately generate the dynamics of the observed values. Consequently, the use

of MRMSE as a performance measure allows for a more informed selection of the most suitable imputation techniques and model types for specific applications.

In light of the aforementioned shortcomings, I introduce GIMI, an on-training imputation technique that uses encoded missingness patterns with a masked loss function to facilitate the accurate representation of patient health dynamics, mitigating the imputation error introduced by pre-training imputation techniques. To alleviate a distributional shift between the training and inference, introduced by the autoregressive reconstruction of the patient health trajectory, I adopt a strategy that imputes values during the training phase. Specifically, missing values are initialized to zero and are updated with the predictions generated by the DGM during each epoch. This method allows the model to estimate the state transition dynamics using an input vector without missing values — which are instead replaced by its own predictions from the previous epoch. Consequently, this approach reduces distributional differences between the training and inference stages.

Using the same dataset, we demonstrate that GIMI outperforms all benchmark imputation techniques in reconstructing patient trajectories, emphasizing its suitability as an imputation method for DGM-based transition models. GIMI's accurate reconstruction of patient trajectories enables more precise estimation of health dynamics, leading to improved performance of MDP and RL models. As a result, this enhanced performance facilitates a more effective and informed decision-making process for patient care, underlining GIMI's potential contributions to both theory and practice in the management of healthcare services.

The performance of GIMI, especially in comparison to No-GIMI, strongly suggests that on-training imputation adeptly mitigates the distribution shift between the training and inference stages. Moreover, the model's performance exhibits resilience as the proportion of missing values escalates, signifying the appropriateness of the proposed approach to datasets exhibiting diverse degrees of missing data. The empirical stability demonstrates its robustness and suggests it is well-suited for real-world healthcare situations where data incompleteness is a prevalent challenge.

1.7.1 Future work and limitations

This study utilized Generative Iterative Multiple Imputation (GIMI) with on-training imputation in DGMs to estimate transition models for patients who received intravenous heparin treatment during their ICU stay. Further research could investigate the efficacy of GIMI in estimating transition models for other ICU treatments, such as pain management and sedation, to gain deeper insights into GIMI's performance in varying contexts [Yu et al., 2020]. Additionally, we expect GIMI to perform effectively in problems that rely on similar datasets with high sparsity, but this remains to be tested. Future research could also explore testing GIMI on other DGMs, including normalizing flows and other iterative estimation methods.

Finally, future research could focus on how the pattern of missingness influences the effectiveness of imputation techniques. To improve performance, prior knowledge contextual to specific health dynamics could be incorporated by utilizing varied initialization methods for GIMI and assessing how they affect the algorithm's convergence speed and final performance. For instance, forward imputation may prove more effective when a patient's health trajectory is relatively stable.

Appendix

Table 1.6: MRMSE_1 for each hyperparameter setup with 16 latent dimensions tested during tuning of the VAE model.

En-/Decoder Size	32	64	128	256
Mean	0.3344	0.3286	0.3318	0.3333
Forward	0.4569	0.4130	0.4011	0.3990
KNN	0.5313	0.5306	0.5276	0.5236
MICE	0.3750	0.3783	0.3747	0.3776
GAIN	0.9250	0.9125	0.9109	0.9151
NoGIMI	0.6790	0.6941	0.7103	0.6994
GIMI	0.6662	0.6888	0.6950	0.6575

Table 1.7: MRMSE₁ for each hyperparameter setup with 32 latent dimensions tested during tuning of the VAE model.

Latent Dimensions	32			
	32	64	128	256
Mean	0.3269	0.3212	0.3223	0.3232
Forward	0.4059	0.3801	0.3620	0.3448
KNN	0.5313	0.5213	0.5234	0.5247
MICE	0.3668	0.3646	0.3667	0.3666
GAIN	0.9030	0.9070	0.8884	0.8894
NoGIMI	0.6799	0.6604	0.6837	0.6748
GIMI	0.6812	0.6599	0.6369	0.6442

Table 1.8: MRMSE₁ for each hyperparameter setup tested during tuning of the Transformer model.

Heads	1			4		
	16	32	64	16	32	64
Mean	0.3163	0.3056	0.2987	0.3044	0.3000	0.2990
Forward	0.3662	0.2478	0.2439	0.2641	0.2441	0.2424
KNN	0.5208	0.5060	0.5067	0.5212	0.5098	0.5050
MICE	0.3495	0.3437	0.3401	0.3498	0.3483	0.3444
GAIN	0.9094	0.9139	0.8827	0.9062	0.8966	0.8785
NI	0.6689	0.6022	0.6148	0.6654	0.6052	0.5908
GIMI	0.6360	0.5917	0.5650	0.6441	0.6011	0.5644

Chapter 2

Preserving Ethnic Diversity in Topic Models for Residency Applications

2.1 Abstract

Constrained by time and resources, U.S. graduate medical education programs are only able to use holistic review for a subset of applications they receive. Computational language models can assist these programs with the efficient and precise evaluation of applicant-generated texts, such as personal statements, but caution must be exercised as these models are known to be susceptible to mirroring the biases reflected in the datasets that they are trained on. This paper aims to evaluate the scope and impact of (historical) ethnic selection bias on a corpus of 35,000 personal statements extracted from residency applications collected between 2019 and 2023. An initial word-frequency analysis reveals significant differences between the word distributions of applicants from different ethnic backgrounds. Subsequently, a topic analysis shows that these differences are also observed in the personal statements' semantic structure. Furthermore, I find that ethnic diversity is suppressed when all ethnicities share the same topic distribution, and this suppression is negatively related to the ethnicity's representation in the dataset, thus, affecting Black and Latino applicants in particular. Taking these results into account, I propose a topic model that combines shared and ethnicity-specific topics. I demonstrate through the UMass coherence that the topics estimated by this model are more meaningful.

2.2 Introduction

Graduate medical education (GME) is placing increasing pressure on residency programs to expand the scope of their application review process, a practice referred to as holistic review [Marbin et al., 2021, Pope et al., 2021, Aibana et al., 2019, Otugo et al., 2021, Nehemiah et al., 2021]. This growing emphasis is motivated by a desire to recruit more diverse candidates and promote equality and inclusivity (DEI). Unlike traditional review methods, which primarily focus on quantitative measures (e.g., USMLE score, Clerkship grades, GPA), the goal of Holistic reviews is to evaluate the candidate based on a variety of relevant personal and interpersonal characteristics and skills. These

frequently include the perceived ability to work effectively as part of a broader healthcare team, exhibiting leadership potential, and participation in extracurricular activities or voluntary or elected scholastic organizations. In addition to a more fair process, medical schools that shifted from traditional to holistic review methods have also reported that no negative effect on the academic performance of the admitted applicants was observed [Schulz et al., 2022]. This is also confirmed by the NRMP's yearly program director survey in 2022, where "identifying promising applicants" was cited by 78% of the respondents as a driver for engaging in holistic review, highlighting the perceived value of holistic review beyond promoting DEI [NRMP, 2022b].

Today, a holistic review of applications to GME is far from ubiquitous. While the majority of residency programs (intend to) use holistic review throughout their admissions process, the majority of students never receive said holistic review because they fail to pass heuristic screening methods in the earliest stage of the process [Moynahan, 2018, NRMP, 2021]. Constrained by the substantial time, knowledge, and resources needed to complete a uniform holistic review of all applicants, programs are forced to focus their resources for holistic review on a small subset of applicants, which are regrettably selected through a traditional review process. The cost to review applicants holistically is being multiplied by the astonishing growth in applications submitted per applicant, where residency application rates have nearly doubled in the last decade [Carmody et al., 2021] and in some competitive residencies or fellowships, only 4% of *interviews* (post-application review) result in a match [Lin et al., 2021].

Notably, information-rich documents such as personal statements or letters of recommendation require substantial attention from a reviewer to provide an accurate—the extent to which this is feasible using a single, individual interpretation—evaluation. Personal statements (one per application) and letters of recommendation (typically three per application) extracted from the residency applications that were used in this study contained, on average, 340 and 237 words. At a typical reading speed of 238 words per minute [Brysbaert, 2019], an evaluator has to spend almost 4.5 minutes per applicant to evaluate these written texts. This means that, for example, a typical internal medicine

program that receives around 3000 applicants would have to spend a total of 225 hours to complete a single review of each of these personal statements and letters of recommendation.

The cost of reviewer effort drives the demand for appropriate decision support tools. Language models, such as GPT-4 [OpenAI, 2023], have demonstrated utility on a large scale in assisting with text-related tasks, such as academic writing [Aydin and Karaarslan, 2022] [Gao et al., 2022], simplifying reports [Jeblick et al., 2022], creating summaries of medical dialogues [Chintagunta et al., 2021] and test-taking [Kung et al., 2023]. While review of personal statements is costly, concern for ethnic biases in black-box machine learning solutions is warranted, as, when crafted without careful considerations, these same language models can reflect counter-productive beliefs and patterns that are represented in the dataset they're trained on [Abid et al., 2021] [Johnson and Khoshgoftaar, 2019]. In the case of GME, there is concern that this bias occurs along ethnic lines as a result of (historical) selection bias at the expense of applicants with a Black and Latino background. In 2022, the National Resident Matching Program found that 8.1% of active U.S.-citizen MD residents reported identifying as African American and 6.1% reported as Hispanic [NRMP, 2022a]. The 2020 Decennial Census found the American population to be 13.6% African American and 18.9% Hispanic or Latino [Bureau, 2022], more than double the admittance rates to residency. This concern is further motivated by findings from linguistic studies, which suggest that ethnic identity, among other identities, plays a crucial role in shaping the semantic structure of academic written texts, as it influences the experiences and subject positions available to individuals within their socio-cultural context [Ivanič, 1998] [Bucholtz and Hall, 2004]. In addition, various studies have noted semantic differences between the personal statements of male and female GME applicants [Babal et al., 2019] [Osman et al., 2015] [Ostapenko et al., 2018] [Demzik et al., 2021], as well as applicants from varying socioeconomic backgrounds [Alvero et al., 2020].

In the following study, I aim to advance our understanding of ethnic diversity in the applicant-generated documents from GME recruitment. Using a dataset of 35,000

personal statements of U.S. medical residency applicants, I present word and topic analyses of personal statements written by Asian, Black, Latino, and White applicants. A word frequency analysis initially reveals disparities in the word distributions of applicants from varied ethnic backgrounds. Subsequently, a topic analysis shows that these variations are likewise discernible in the semantic structure of the personal statements. I then demonstrate that when a single topic distribution is employed for all ethnic subgroups, the diversity of topics for a group is inversely correlated to their presence in the dataset. Consequently, I conduct a comparative study of topic models with varying topic distributions and use the UMass coherence score to evaluate the meaningfulness of the topics estimated by these models. I find that, on average, the topics estimated by a hybrid model, which has both shared and ethnicity-specific topic distributions, are more meaningful as compared to the topics estimated by topic models that have solely shared or ethnicity-specific topic distributions.

2.3 Methods

2.3.1 Ethics statement

Thalamus (SJ MedConnect, Inc dba Thalamus, Santa Clara, CA) provided the de-identified demographic and personal statement data for this work. Thalamus is a cloud-based GME interview scheduling software and management platform. Authors E.L. and S.V.S. are shareholders of Thalamus. For the purpose of anonymity, the data in this study were de-identified upon extraction from the Thalamus database, and no identifiable information was shared with the investigators.

2.3.2 Data Collection

The demographic and personal statement data in this study are secondary use and were originally extracted from ERAS applications that were collected by Thalamus between 2019 and 2023. The following specialties have been included in this study: Obstetrics and

Gynecology, Internal Medicine, Family Medicine, Emergency Medicine, Surgery, Pediatrics, Anesthesiology, and Psychiatry. I removed all candidates for whom ethnicity was unknown, and I only selected one personal statement per applicant to avoid duplicate personal statements. After this, 34,976 personal statements remained in the dataset for an equal number of applicants.

2.3.3 Data Cleaning and Text Pre-Processing

I perform basic text pre-processing techniques that are typically performed prior to a bag of words analysis. I remove stopwords, punctuation, and other special characters, lemmatize the corpus, and remove words that are observed too infrequently (< 100 documents), as well as too frequently ($> 30\%$ of the documents). I also include bigrams and trigrams that occur at least ten times and achieve a minimum score of 100 as calculated by the n-gram Scoring Function proposed by Mikolov et al. [Mikolov et al., 2013]. After the text pre-processing, the cleaned corpus contains 24,212 unique words in 34,976 documents.

2.3.4 Descriptive study

Topic model To estimate the semantic structure of the personal statements for the descriptive study, I use the Ensemble Latent Dirichlet Allocation (ELDA) topic model [Brigl, 2018] estimated with the online Variational Bayes (VB) algorithm proposed by Hoffman et al. [Hoffman et al., 2010]. This topic model is an extension of the Latent Dirichlet Allocation (LDA) model [Blei et al., 2003], where a set of LDA models is estimated, and an ensemble vote is inferred from this set of LDA models using a variation of the DBSCAN algorithm [Ester et al., 1996], the CB-DBSCAN algorithm [Ghazizadeh et al., 2020]. I chose an ensemble size of ten models and fixed the number of topics at 100. I am able to do this because the ELDA model uses the clustering step to filter out any spurious topics, where the certainty of this step is driven by the minimum number of topics that need to be observed to be considered a stable topic, which I fix at

8. Asymmetrical α and β parameters are inferred from the data during model estimation [Řehřek et al., 2011].

Hyperparameter optimization & model selection I conduct an exhaustive search over the two hyperparameters of the ELDA model; the maximum distance a topic is allowed to be from a cluster center to be potentially considered a core topic, ϵ , and the minimum number of cores a cluster has to have, to be considered a stable topic. The minimum perplexity is the evaluation criterion for model selection. The results for this hyperparameter optimization are provided in the appendix. The results and corresponding optimal hyperparameter configuration can be found in Table 2.1 (which is located in the Appendix).

2.3.5 Predictive study

Estimation method In the descriptive study, I used a Variational Bayesian method to estimate the optimal topic distribution. I opted for this estimation method because of its computational efficiency. However, when I use a VB method, I do not find the true data distribution; instead, I estimate the lower bound of the evidence. Gibbs sampling offers a solution to this problem, as the true distribution can be inferred with this algorithm, but it has a drawback that it's more computationally complex. It's for this reason that I also do not infer the hyperparameters from the dataset, and I opt for hyperparameter selection based on the literature rather than optimizing it.

Topic models Here, I describe the three different LDA models used in this study: the shared, ethnicity-specific and hybrid LDA models. The shared LDA model consists of a single LDA model where the topic distribution is shared by applicants of all ethnicities. The ethnicity-specific LDA model consists of a set of LDA models, where each LDA model, and corresponding topic distribution, maps to an ethnicity. The hybrid model combines the two previous approaches, having a topic distribution that is shared by all applicants

and a set of topic distributions that map to a specific topic distribution. The generative process of each LDA model is now presented.

The Shared LDA Model is the classic LDA model proposed by Blei et. al [Blei et al., 2003]. For each document d , the weights for a topic distribution θ_d are drawn from a Dirichlet distribution, denoted as $\text{Dir}()$, with parameter α . Then, for each topic k , a word distribution ϕ_k is drawn from a Dirichlet distribution with parameter β . Next, for each word w in document d , a topic z_{dw} is drawn from a Multinomial distribution, denoted as $\text{Mult}()$, with parameter θ_d . Finally, for each word w in document d , the actual word w_{dw} is drawn from a Multinomial distribution with parameter ϕ_k .

$$\begin{aligned}
 \theta_d &\sim \text{Dir}(\alpha) \\
 \phi_k &\sim \text{Dir}(\beta) \\
 z_{dw} &\sim \text{Mult}(\theta_d) \\
 w_{dw} &\sim \text{Mult}(\phi_k)
 \end{aligned} \tag{2.1}$$

The Ethnicity-Specific LDA Model: adds an additional layer of granularity to the model, allowing it to capture the relationship between the writer's ethnic identity and the themes found in their academic writing. Here, the document-topic distribution, θ , and topic-word distribution, ϕ , are both specific to the ethnic identity of the writer. For each document d and ethnic identity i , a topic distribution θ_{di} is drawn from a Dirichlet distribution, denoted as $\text{Dir}()$, with parameter α_i . For each topic k and ethnic identity i , a word distribution ϕ_{ki} is drawn from a Dirichlet distribution with parameter β_i . For each word w in document d with ethnic identity i , a topic z_{dwi} is drawn from a Multinomial distribution, denoted as $\text{Mult}()$, with parameter θ_{di} . Finally, for each word w in document d with ethnic identity i , the actual word w_{dwi} is drawn from a Multinomial distribution with parameter ϕ_{ki} .

$$\begin{aligned}
\theta_{di} &\sim \text{Dir}(\alpha_i) \\
\phi_{ki} &\sim \text{Dir}(\beta_i) \\
z_{dwi} &\sim \text{Mult}(\theta_{di}) \\
w_{dwi} &\sim \text{Mult}(\phi_{ki})
\end{aligned} \tag{2.2}$$

The Hybrid LDA Model: incorporates an additional parameter, ι , which represents the ethnic identity of the writer. The term and topic distributions are now conditionalized based on ι , which is drawn from a multinomial distribution with parameter ϵ_{wi} . The model differs from the previous two models by incorporating this additional layer of granularity, allowing for a more nuanced understanding of the relationships between ethnic identity and the terms and topics found in the documents. In this model, the generative process is as follows:

For each document d , and word w , the writer's ethnic identity ι_{dw} is drawn from a Multinomial distribution, denoted as $\text{Mult}()$, with parameter ϵ_{wi} . For each document d , a topic distribution θ_d is drawn from a Dirichlet distribution, denoted as $\text{Dir}()$, with parameter α , conditional on ι_{dw} . For each topic k and ethnic identity i , a word distribution ϕ_k is drawn from a Dirichlet distribution with parameter β , conditional on ι_{dw} . For each word w in document d , a topic z_{dw} is drawn from a Multinomial distribution with parameter θ_d , conditional on ι_{dw} . Finally, for each word w in document d , the actual word w_{dw} is drawn from a Multinomial distribution with parameter ϕ_k , conditional on ι_{dw} .

$$\begin{aligned}
\theta_d &\sim \text{Dir}(\alpha|\iota_{dw}) \\
\phi_k &\sim \text{Dir}(\beta|\iota_{dw}) \\
\iota_{dw} &\sim \text{Mult}(\epsilon_{wi}), \\
z_{dw} &\sim \text{Mult}(\theta_d|\iota_{dw}) \\
w_{dw} &\sim \text{Mult}(\phi_k|\iota_{dw})
\end{aligned} \tag{2.3}$$

The probability of drawing the ethnicity (ι), $\epsilon_{w\iota}$, is defined as:

$$\epsilon_{w\iota} = \max \left(0, \left(\frac{1}{1 + e^{-\mu_{w\iota}}} - \frac{1}{2} \right) \right) \quad (2.4)$$

where $\mu_{w\iota}$ is given by:

$$\mu_{w\iota} = \frac{P(w|\iota)}{P(w) - P(w|\iota)} \quad (2.5)$$

This expression represents the conditional probability of observing the word w , given the writer's ethnic identity ι , divided by the difference between the overall probability of observing the word w and the probability of observing it given the writer's ethnic identity ι . The model is subject to the constraint:

$$\sum_{\substack{i \in I \\ i \neq \iota}} w_i < 1000 \quad (2.6)$$

This constraint ensures that the total weight of words in the term distribution for all other ethnic identities (excluding the one being considered) remains below a certain threshold (1000 in this case). I train topic models with 10, 30, and 50 topics. For the hybrid model, the ethnicity-specific topics are fixed at three topics, such that the shared numbers of topics are 7, 27, and 47, respectively.

Hyperparameter selection & Model Evaluation Given the considerable runtime of each model, I opted for a symmetrical prior for both hyperparameters α ($\frac{50}{K}$) and β (0.01), following the suggestions of the original LDA paper [Blei et al., 2003]. I feed the dataset to our model 200 times and calculate the per-word UMass coherence for both the train and validation dataset at the end of each pass over the corpus. The per-word normalization was performed to allow a comparison of perplexity over corpora of different sizes.

2.4 Results

I analyze the personal statements from residency applications between 2019 and 2023 for eight specialties: Obstetrics and Gynecology, Internal Medicine, Family Medicine, Emergency Medicine, Surgery, Pediatrics, Anesthesiology, and Psychiatry. I constrain the dataset to 1 personal statement per applicant, as applicants frequently re-use personal statements for multiple applications. In addition, I extracted the applicant's self-identified ethnicity as Asian, Black or African American, Latino, White, or any combination of these ethnicities, removing all applicants whose ethnic identities were either unknown or could not be classified in one or more of these ethnicities. From this sample, there were 2493 students with two self-identifications and 74 students with three self-identified ethnicities. A total of 34,976 personal statements are left for analysis. I use a topic model, Latent Dirichlet Allocation (LDA), to estimate the semantic structure of the corpus. It does this by assigning each document a mixture of topics (the document-to-topic distribution). Each topic, in its turn, is represented as a mixture of words (the topic-to-term distribution).

2.4.1 The Effect of Historical Selection Bias on Data Quality

I begin by investigating the effect of (historical) ethnic selection bias in GME in the corpus of personal statements for each ethnic group. Figure 2.1.a compares the ethnic distribution of the U.S. population (*stage 1*) to the ethnic distribution of aspiring physicians at different stages of the physician pipeline, respectively the number of medical school applicants (*stage 2*), the number of personal statements in this dataset (*stage 3*) and the number of active U.S. residents (*stage 4*). Asian applicants are the most over-represented ethnic group in the sample corpus, representing only 5.13% of the population while making up 26.49%, 27.59%, and 21.65% at the respective second, third, and fourth stages of the physician pipeline. We can see that the relative number of Asian applicants initially increases from the second to the third stage but then decreases significantly between the third and fourth stages. In contrast, the relative share of Black and Latino applicants decreases consistently throughout the pipeline, representing at each stage

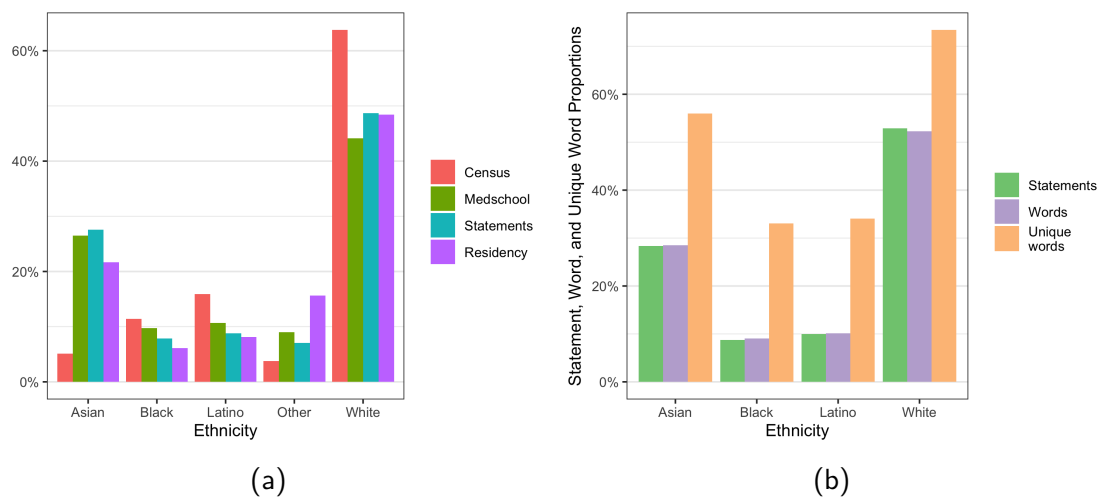


Figure 2.1: a) Ethnic distribution at different stages of the physician pipeline and its effect on data quality. b) Proportion of personal statements, words, and unique words by ethnicity; note that, for unique words, the normalizing constant is less than the sum of unique words across groups (i.e., the ethnicity's unique words overlap), and therefore these proportions sum to greater than 100% while the proportions of words and statements are mutually exclusive and sum to 100%.

11.44%, 9.71%, 7.83%, and 6.11% for Black applicants and 15.90%, 10.65%, 8.83%, and 8.13% for Latino applicants. Finally, White applicants are underrepresented at each stage of the physician pipeline, as compared to their relative share in the U.S. population of 63.75%. However, their relative share increases throughout the pipeline, from 44.16% at the second stage to 48.72% at the third stage and 48.43% at the fourth stage.

Figure 2.1.b shows how ethnic distributions are reflected in the corpus size and word diversity for each ethnicity. The relative number of personal statements, which maps directly to the relative number of applicants in the dataset, is 28.34%, 8.77%, 9.97%, and 52.92% for Asian, Black, Latino, and White applicants, respectively. This is also reflected in the relative numbers of words with respective shares of 28.47%, 9.09%, 10.13%, and 52.31% for Asian, Black, Latino, and White applicants. Furthermore, a dampened effect is also observed for the relative number of unique words, with Asian, Black, Latino, and White applicants sharing a respective 55.97%, 33.03%, 34.09%, and 73.40% of the vocabulary in the entire corpus.

2.4.2 Corpora Differ between Applicants with Different Ethnic Backgrounds

I first conduct a Chi-Square test to establish if there are significant differences between the frequencies of the words used in the corpora of Asian, Black, Latino, and White applicants. With a Chi-Square test statistic of 706,449, I reject the null hypothesis that the vocabulary of the personal statements is consistent across applicants of differing ethnic backgrounds. In addition, I find a positive relationship between the proportion of samples in the general dataset, with Chi-Square values of 178,696, 213,113, 175,816, and 138,822 for Asian, Black, Latino, and White applicants, respectively.

The first four plots of Figure 2.2 present, for each ethnicity, which terms occur the most frequently relative to the average frequency of the term in the general dataset. For Asian applicants, four terms refer to countries in Asia (“Bangladesh”, “India”, “Pakistan”, and “Taiwan”), while five terms refer to an identity relevant for Asian applicants to Asia

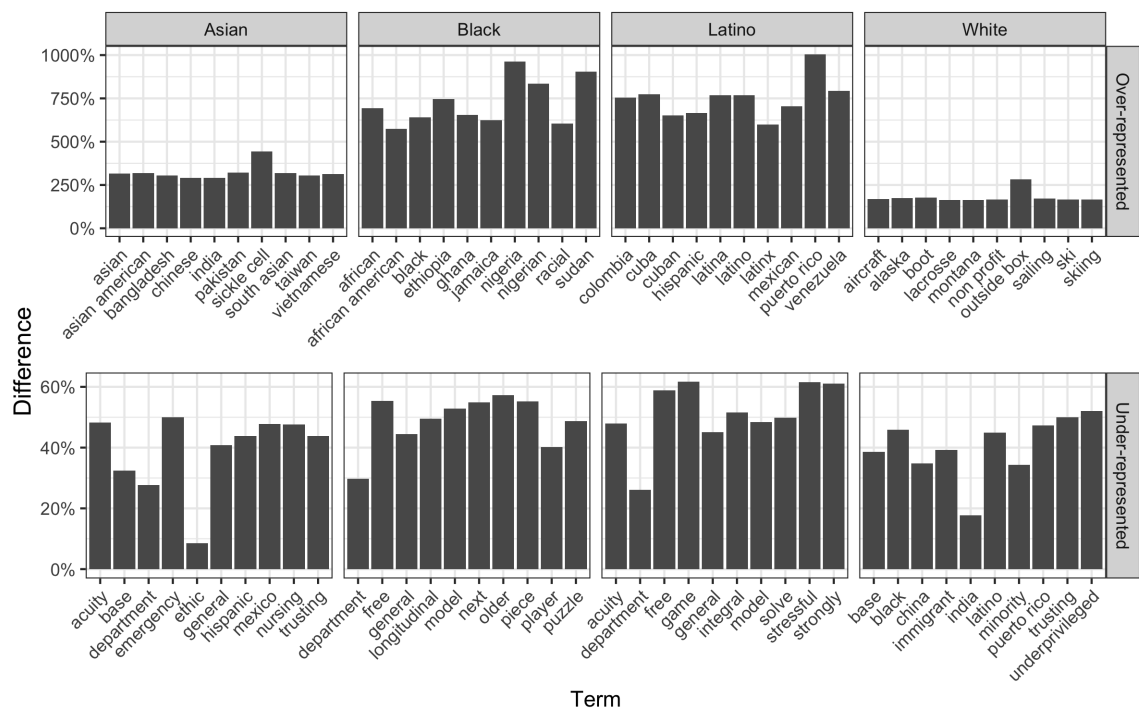


Figure 2.2: Most over- and under-represented terms for each ethnicity.

(“Asian”, “Asian American”, “Chinese”, “South Asian”, “Vietnamese”). One term, “sickle cell”, stands out as it fits in neither category and is significantly more over-represented compared to the other categories. This outlier may be explained by the high prevalence of different variants of sickle cell anemia in the Indian subcontinent [Piel et al., 2010].

For Black applicants, I observe a similar pattern, with five terms referring to African countries (“Ethiopia”, “Ghana”, “Jamaica”, “Nigeria”, “Sudan”), four terms referring or related to an identity relevant to Black applicants (“African”, “African American”, “Black”, “Nigerian”, “racial”). Similarly, for Latino applicants, I observe three terms referring to a Central or South American country (“Colombia”, “Cuba”, “Venezuela”) and one to a U.S. territory (“Puerto Rico”). The other six terms refer to identities relevant for Latino applicants (“Cuban”, “Hispanic”, “Latina”, “Latino”, “Latinx”, “Mexican”), with the gender-specific variants “Latina” and “Latino” more commonly used than their gender-neutral counterpart, “Latinx”. A different pattern is observed for white applicants, with seven frequently used terms referring or related to extracurricular activities (“aircraft”, “boat”, “non-profit”, “sailing”, “ski”, “skiing”), two U.S. states (“Alaska”, “Montana”), and one idiosyncratic term, “outside the box”. The extracurricular activities likely corroborate earlier findings that medical students with an affluent background are over-represented in GME [Shahriar et al., 2022]. Strikingly, the most commonly used terms for non-white applicants focus almost exclusively on location and identity, in contrast to the terms for white applicants, which focus mostly on extracurricular activities. A potential explanation for this observation is that, as a result of external pressure, applicants might feel the need to engage in self-tokenization, where they stress the more “exotic” aspects of their background in order to attain the image of diversity in order to appear more desirable to the admissions board and, ultimately, more likely to be accepted [Park and Bahia, 2022].

The bottom four plots of Figure 2.2 present, for each ethnicity, which terms occur the least frequently, relative to the frequency of the term in the general dataset. In contrast to the over-represented terms, I do not observe any strong discernible patterns for the under-represented terms of Asian, Black, and Latino applicants. For White applicants, however, I

observe an under-representation of terms specific to other Asian (“China”, “India”), Black (“Black”), and Latino (“Latino”, “Puerto Rico”) applicants or related to experiences shared by these groups (“immigrant”, “minority”, “underprivileged”). In addition, the term “ethic” is undoubtedly the most under-represented term for Asian applicants, while “department” is a recurring under-represented term among all non-White applicants alike. We can observe that terms related to problem-solving are under-represented for both Black (“model”, “piece”, “player”, “puzzle”) and Latino (“game”, “model”, “solve”) applicants.

2.4.3 Ethnicity-Agnostic Topic Models Suppress Ethnic Variety

I use an Ensemble Latent Dirichlet Allocation (ELDA) topic model to estimate the semantic structure of the corpus of personal statements. A total of five (Asian, Black, Latino, White, and General) models are estimated, for which the topic distributions are presented in Figure 2.3; For each topic, the theme, density, and five most important keywords are presented. I observe both topics that are limited to one or more ethnic groups (for example, “Black” for Black applicants or “Puerto Rico” for Latino applicants), as well as topics that are shared by all ethnicities (for example, the specialty-related topics “anesthesiology”, “emergency medicine”, “pediatrics” and “psychiatry”). Interestingly, only a subset of the ethnicity-specific topics are also observed in the topic distribution of the general model, and an association appears to exist between the number of ethnic groups for which the topic is observed. For example, the topics “grandmother”, “teaching” and “radiology” are recurrent for Asian, Black, and Latino applicants and, as such, are also observed in the general model, while topics such as “exam” (Asian), “Puerto Rico” (Latino,) “Black” (Black) and “rural” (White) are limited to a single ethnicity and, as such, are not captured in the general model.

Furthermore, I see that for each ethnicity, topics related to the specialty for which the applicant is applying are most important with respect to each ethnicity’s mean density. For Asian, Black, and Latino applicants, “psychiatry” is the most common topic, with a mean density of 15.14%, 24.29%, and 16.94%, while “internal medicine” is the most

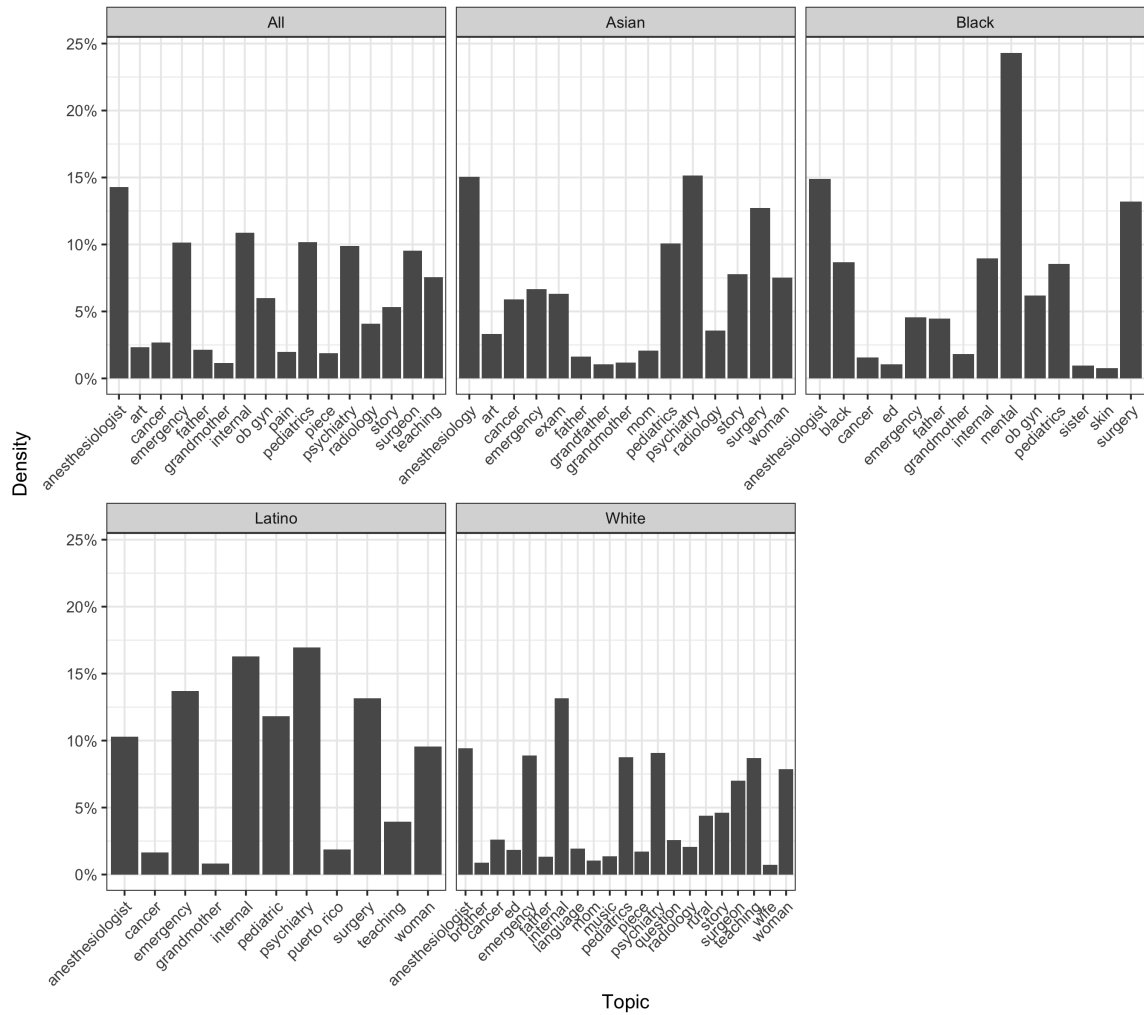


Figure 2.3: Topic distributions for the general and ethnicity-specific models

common topic for White applicants, with 13.17%. For each ethnicity, the first four topics all relate to a specialty. I find that the first three (if three exist) non-specialty-related topics for each ethnicity, ordered by mean density, are, “story”, “exam” and “art” for Asian, “Black”, “father” and “grandmother” for Black, “Puerto Rico” and “teaching” for Latino and “teaching”, “story” and “rural” for White applicants. Notably, the topic distribution of White applicants contains considerably more topics, as compared to the topic distribution of the other ethnicities, as well as the topic distribution of the general model. A potential explanation for this observation is that the larger corpus of White applicants allows the model to learn better as compared to the other ethnic groups. In contrast, the diversity in the general model inhibits the detection of an extended topic distribution.

Figure 2.4 visualizes the relationship between the density estimated by the general model and each ethnicity-specific model for the four topics with the highest mean density and which also were observed in the topic distribution of each ethnicity: “anesthesiology”, “emergency medicine”, “pediatrics” and “psychiatry”. The linear correlation for each ethnicity is visualized by the color-coded line, while the black line is the 45-degree line which we should expect all observations to fall on if the density estimated by the ethnicity-specific model is equal to the density estimated by the general model. The mean density for each topic is consistently underestimated by the general model for Asian, Black, and Latino applicants, with the latter two differing considerably more. For White applicants, however, the mean density nearly coincides with the 45% line. In other words, the topic distribution of the general model, where all the ethnicities share the same topic distribution, is skewed towards the majority ethnicity, White applicants.

Figure 2.5 provides further evidence. On this heat map, I represent, for each ethnicity, the distance between the topics estimated by the general model and the topics estimated by the ethnicity-specific model. Furthermore, ethnicities are ordered, from left to right, by their number of observations, while topics are ordered, from bottom to top, by the density of the topic. I observe a negative relationship between the distance and the topic density, as well as the number of observations of the ethnicity. This is also reflected in

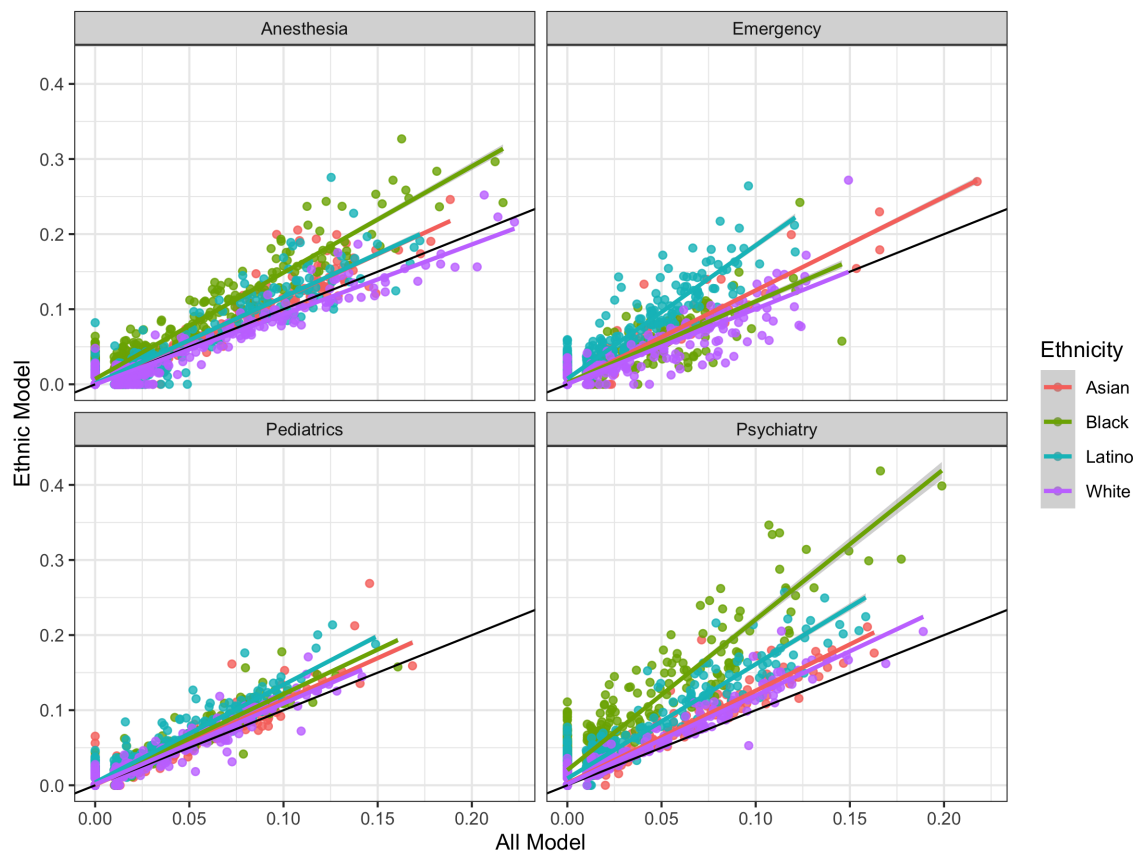


Figure 2.4: Comparison of the density estimated by the general model and the ethnicity-specific models for the four most important and shared topics.

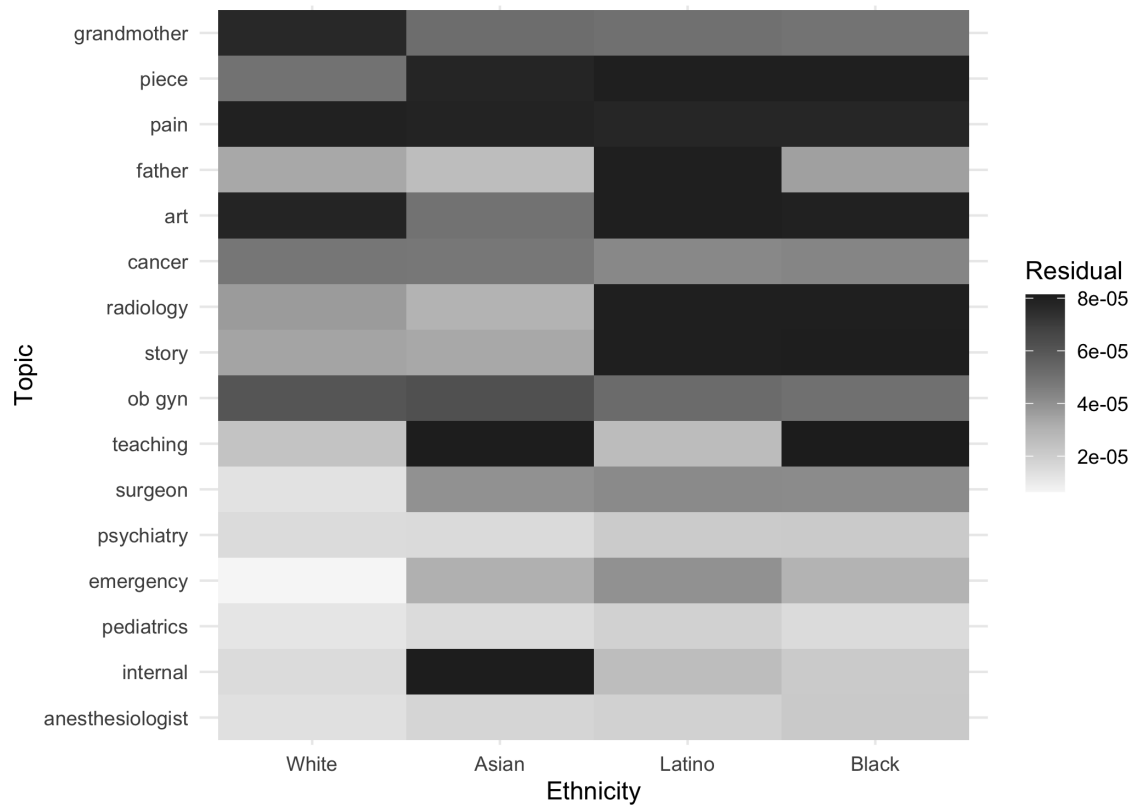


Figure 2.5: Distances between the General Model Topics and the Ethnicity-Specific Model Topics.

the number of overlapping topics between the general model and each ethnicity-specific model, with ten, seven, eight, and ten overlapping topics for Asian, Black, Latino, and White applicants, respectively.

I also compare the distribution of these distances in Figure 2.6 to a distribution that was bootstrapped by sampling, without replacement, two topics from the general model's topic distribution. This bootstrapped distribution is the distribution that we should expect to find when topics are not correlated. Instead, however, we observe a distribution that is rather bi-modal, with one mode at the expected location but also another mode to the left. Furthermore, we observe that the expected mode is higher as the number of samples increases, indicating a stronger overlap with the general model. Thus, the topic distributions of the ethnicity-specific models contain both topics that are shared with the topic distribution of the general model (the first mode) and topics that are specific to that or more ethnicities (the second mode). These shared topics could be leveraged to improve the model for ethnic groups with a low number of samples by learning from the samples of other ethnic groups.

2.4.4 Ethnic Variety Preserves Semantic Relationships

I now compare how different degrees of overlap between the topic distributions of applicants from varying ethnic backgrounds affect the quality of the topics. Specifically, I use the UMass coherence score [Mimno et al., 2011] to evaluate, on a hold-out sample, three topic models: ethnicity-specific, shared, and hybrid LDA. The ethnicity-specific model incorporates a set of topic distributions that is only shared between applicants of the same ethnic background. The shared model utilizes a single topic distribution that is shared by all applicants. The hybrid model encompasses both shared and ethnicity-specific topic distributions. The UMass Coherence score [Mimno et al., 2011] measures the degree to which words of the same topic co-occur within a given context, in our case, a personal statement. This score measures the meaningfulness and interpretability of topics, assuming semantically related words occur together. For each model, topic distributions

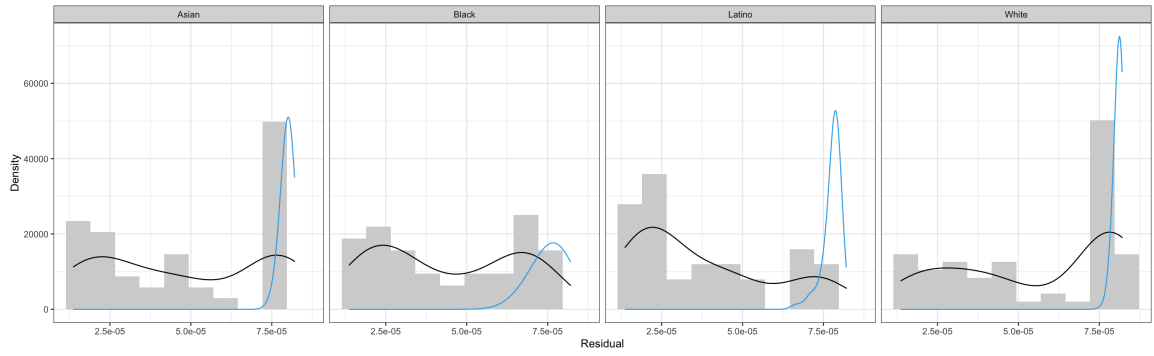


Figure 2.6: Distributions of the Distances between the General Model Topics and the Ethnicity-Specific Model Topics.

with 10, 30, and 50 topics are estimated. We fix the number of ethnicity-specific topics for the hybrid model at 3.

Figure 2.7 compares the UMass score, calculated for the top 30 terms, for each combination of model type and number of topics. The ranking of the three models is consistent across the 10-, 30-, and 50-topic models. First, The hybrid model has the highest coherence score for all models, with -21.9097, -25.6513, and -27.3464 for the 10-, 30-, and 50-topic models, respectively. Second, the ethnicity-specific model with coherence scores of -23.2104, -26.2815, -27.7987, while the shared topic model is last with -23.5472, -26.3143, -27.8864 or the 10-, 30-, and 50-topic models, respectively. Furthermore, I note that the difference in coherence scores between the hybrid model and the ethnicity-specific model is considerably larger than the latter's distance to the shared model. The negative association between the coherence score and the number of topics is consistent with the literature [Stevens et al., 2012].

Figure 2.8 presents the coherence score calculated for the top 5 to 30 terms, normalized by the average score of all three model types. As the number of terms and topics increase, this metric converges to -1 . The hybrid model consistently achieves the highest coherence score, with the exception of the coherence scores calculated for the top 11, 12, 14, and 16 terms of the 30-topic model, where the ethnicity-specific model slightly outperforms the hybrid model.

In Figure 2.9, I compare the ethnicity-specific topic distributions for each hybrid model. The mean densities are positively correlated with sample size and exhibit considerable similarity. For Asian applicants, we observe that the topics “China”/“Chinese” and “India” for each model, while the topics “Pakistan”, “Philippines”, and “Vietnam” each appear once in the 10-, 30-, and 50-topic models, respectively. For Black applicants, the topics “Black”, “Nigeria”, and “Ghana” are consistently observed across all models, although “Ghana” is replaced by “Haiti” in the 10-topic model. Among Latino applicants, Puerto Rico is observed in all three models, while both Hispanic (10, 30) and Spanish-speaking (30, 50) are observed in two models. Colombia is found solely in the 10-topic model.

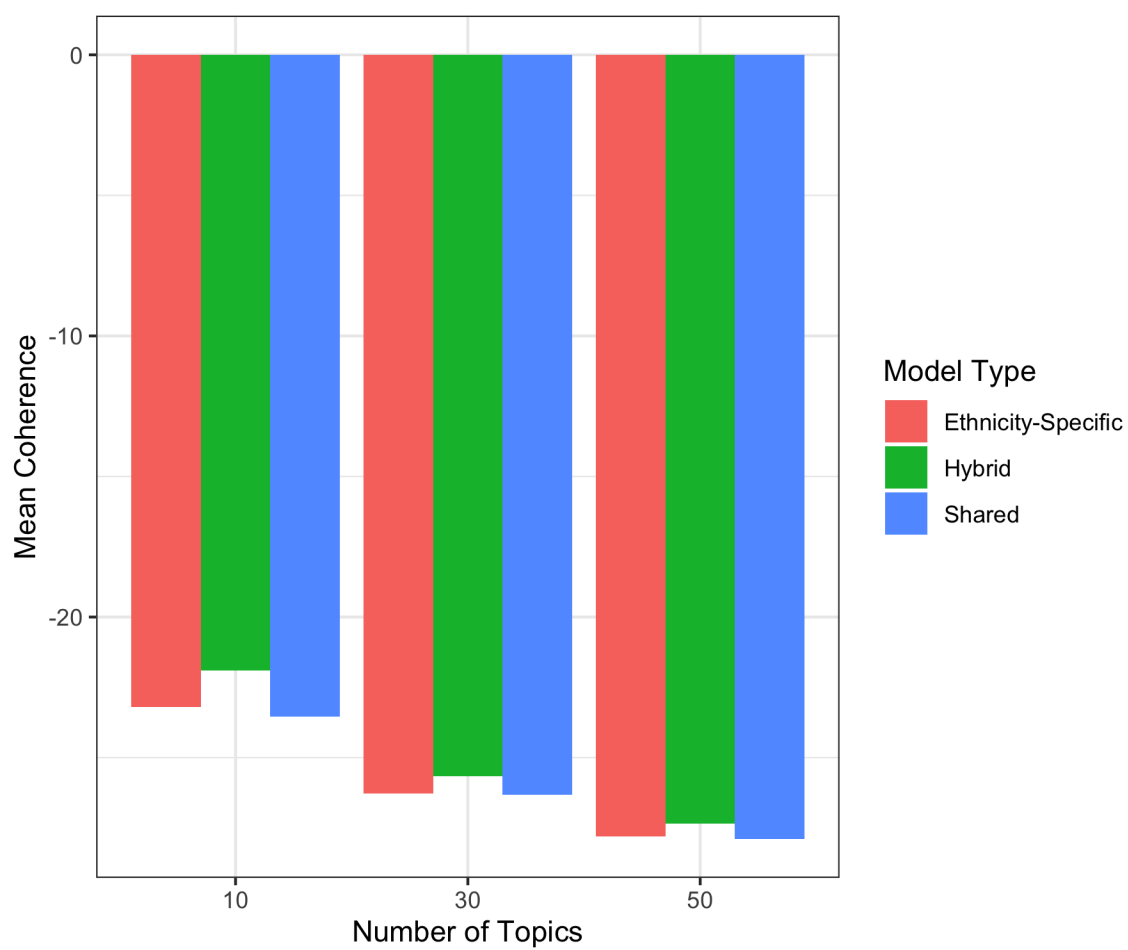


Figure 2.7: Mean coherence scores for the topics of the ethnicity-specific, hybrid, and shared topic models, calculated for the top 30 terms.

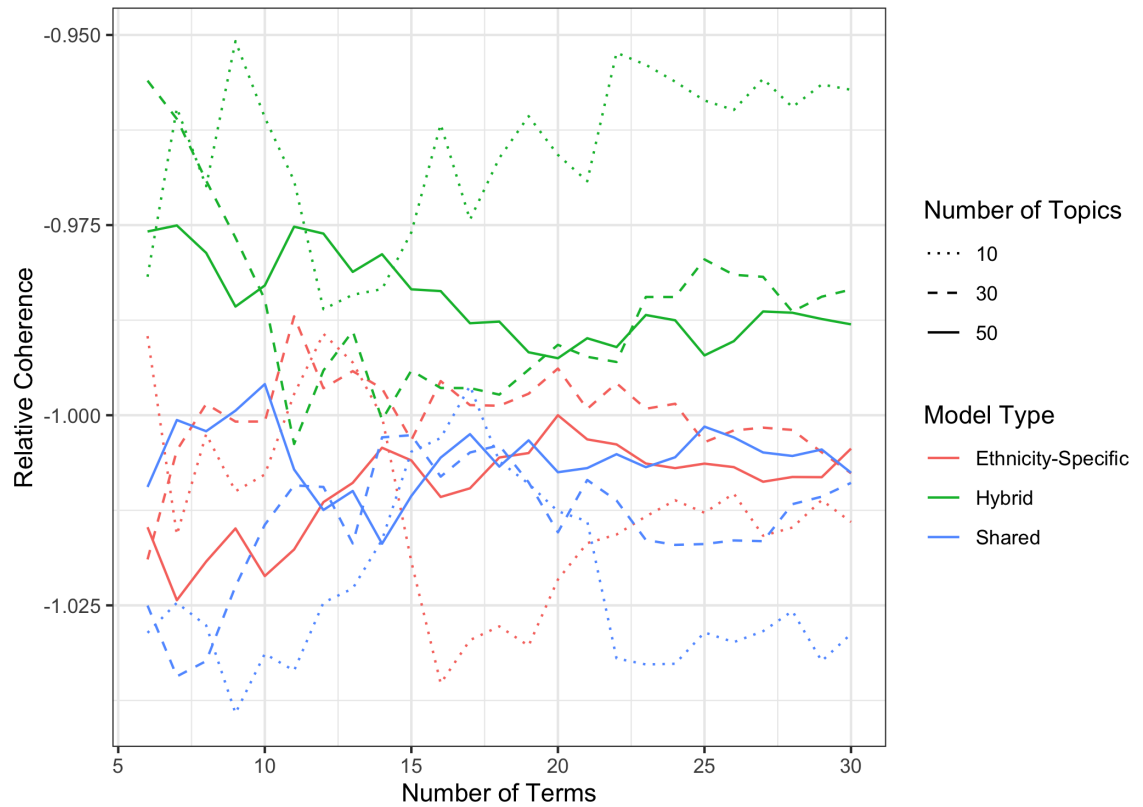


Figure 2.8: Relative mean coherence scores for the topics of the ethnicity-specific, hybrid, and shared topic models, calculated for the top 5 to 30 terms.

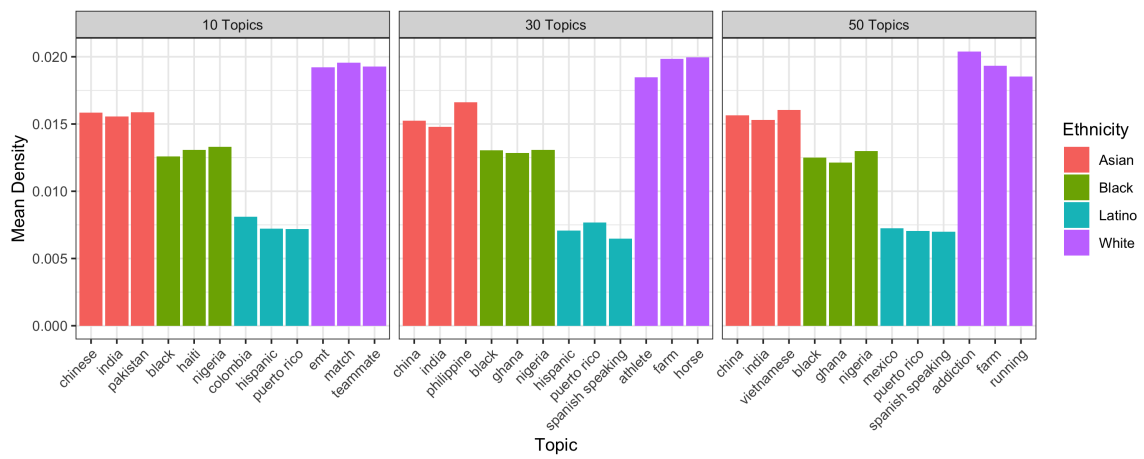


Figure 2.9: Hybrid Model Topic Distributions for All Ethnicities.

Interestingly, for White applicants, no topic is observed three times, which may indicate that the number of topics is not well fit. However, sports-related topics emerge in all three models, featuring terms such as “match” and “teammate” in the 10-topic model, “athlete” in the 30-topic model, and “running” in the 50-topic model. Additionally, rural-related topics like “farm” and “horse” are found in the 30 and 50-topic models. Lastly, both “EMT” (10) and “addiction” (50) appear in one model each. Consistent with the prior word analysis, I find that non-white applicants consistently have topics referring to their geographic locations and identities, while for white applicants, there’s an emphasis on extracurricular activities.

2.5 Conclusion

The tech industry is betting heavily on large language models like ChatGPT as they recognize the potential advantages these models offer in automating text-based tasks such as search [van Dis et al., 2023], medical writing [Kitamura, 2023] and assisting with medical education [Kung et al., 2023] to name a few. Likewise, these language models could offer great utility in the admissions process for GME programs by helping evaluators to assess user-generated text documents like personal statements and limiting the potential for implicit or explicit evaluator bias. However, the benefits of these models are contingent on our understanding and responsible use of the corpora on which they are trained. In particular, we should be considerate of our obligation to preserve the diverse experiences that these personal statements represent. One important driver of such experiences is likely the writer’s ethnic identity, as found in several linguistic studies which have posited a connection between students’ ethnicity and the themes prevalent in their academic writing. [Ivanič, 1998] [Burgess and Ivanič, 2010]. If such an association exists, we can benefit from explicitly communicating the document writer’s ethnic identity to the model to capture the semantic nuances driven by this identity. Notably, the ethnicity-specific semantics of URM applicants, who are under-represented in the historically collected datasets, could be suppressed by the semantic structure of the White majority.

I analyzed a corpus of 35,000 personal statements of Asian, Black, Latino, and White U.S. medical residency applicants and found that significant differences exist between the word frequency distributions of each ethnicity. Likewise, a subsequent topic analysis provides evidence for differences in semantic structure, ascertaining both ethnicity-specific and overlapping topics in the topic distribution. I find that the ethnic diversity in the topic distribution is suppressed when the writer's ethnic identity is not communicated to the model, and this suppression is skewed towards the identities with the smallest representation in the dataset: Black and Latino applicants. An ethnicity-specific model preserves this identity but prevents the model from sharing information between ethnicities and can be constrained by the ethnicity's sample size, posing a problem in particular for Black and Latino applicants who have fewer historic writing samples.

As such, I conduct a comparative study between three models: an ethnicity-specific, shared, and hybrid model. The ethnicity-specific model has a separate topic distribution corresponding to each ethnicity. The shared model has a single topic distribution across all ethnicities. The hybrid model combines the ethnicity-specific and shared models. I find that the topics estimated by the hybrid model consistently outperform the topics estimated by the ethnicity-specific and shared model with respect to the interpretability of the topics. By using an ethnicity-specific topic distribution while sharing the term distribution across all ethnicities, the topic model is able to capture ethnicity-specific differences in the topic distributions of applicants of different ethnic backgrounds while using the information of all applicants to estimate the semantic structure of topics that are shared across ethnicities.

This study has several limitations. First, I analyze topic models to facilitate a medium in which straightforward model comparison can occur. Topic models are just one family of language models, and while the analysis of text corpora remains relevant, care is advised when extrapolating the results of this study to other language models like ChatGPT. Second, I assume that when an applicant self-identifies as more than one ethnic identity, each identity is equally reflected in the semantic structure of the applicant's personal statement. Collecting additional information on the relative importance of the ethnic

identities of multi-ethnic applicants could improve our model. Additionally, in this study, I take an ethnic perspective on identity, but many other elements of identity, such as gender, nationality, religion, or even more granular interpretations of ethnicity, could change our perspective on the corpora of applicants with diverse backgrounds [Chong and Kuo, 2015]. For example, in this study, I treat Black and African American applicants as one ethnic group (i.e., grouping 10th and 1st generation African Americans together) due to the limited number of observations for these applicants, but one could easily argue that, while those groups share experiences related to race, they also have varied cultural backgrounds. Finally, additional attention could be paid to hyperparameter tuning. I used canonical choices and a minimal amount of fitting in order to achieve interpretable results; however, more work could be done to achieve an optimal configuration.

This research highlights the importance of writer identity to the topics (i.e., the semantic structure) with which applicants write their personal statements. It further demonstrates that naive models suppress this diversity. A comparative study shows that the interpretability of topics is improved by using both shared and ethnicity-specific topic distributions. Our approach can be easily adapted for other types of corpora where writer identity is a driver of semantic structure, but also in other applications where preserving semantic diversity (DEI) is of importance. Thus, our work has the potential to promote a topic modeling approach that is more considerate of the diverse writer backgrounds in many other applications where preserving such diversity is of importance.

Appendix

Table 2.1: Hyperparameters for the ELDA models

Model	MinPts	Eps	Perplexity
All	6	0.25	10.037583
Asian	5	0.26	-10.322205
Black and African American	2	0.19	-11.034521
Latino	4	0.20	-10.712639
White	6	0.26	-10.126031

Bibliography

- [Abid et al., 2021] Abid, A., Farooqi, M., and Zou, J. (2021). Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463. [40](#)
- [Aguilera et al., 2020] Aguilera, A., Figueroa, C. A., Hernandez-Ramos, R., Sarkar, U., Cembali, A., Gomez-Pathak, L., Miramontes, J., Yom-Tov, E., Chakraborty, B., Yan, X., et al. (2020). mhealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the diamante study. *BMJ open*, 10(8):e034723. [9](#)
- [Aibana et al., 2019] Aibana, O., Swails, J. L., Flores, R. J., and Love, L. (2019). Bridging the gap: holistic review to increase diversity in graduate medical education. *Academic Medicine*, 94(8):1137–1141. [38](#)
- [Allen et al., 2021] Allen, A., Siefkas, A., Pellegrini, E., Burdick, H., Barnes, G., Calvert, J., Mao, Q., and Das, R. (2021). A digital twins machine learning model for forecasting disease progression in stroke patients. *Applied Sciences*, 11(12):5576. [6](#)
- [Alvero et al., 2020] Alvero, A., Arthurs, N., Antonio, A. L., Domingue, B. W., Gebre-Medhin, B., Giebel, S., and Stevens, M. L. (2020). Ai and holistic review: informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 200–206. [40](#)
- [Aydın and Karaarslan, 2022] Aydın, Ö. and Karaarslan, E. (2022). Openai chatgpt generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*. [40](#)
- [Babal et al., 2019] Babal, J. C., Gower, A. D., Frohna, J. G., and Moreno, M. A. (2019). Linguistic analysis of pediatric residency personal statements: gender differences. *BMC medical education*, 19(1):1–9. [40](#)
- [Baucum et al., 2020] Baucum, M., Khojandi, A., and Vasudevan, R. (2020). Improving deep reinforcement learning with transitional variational autoencoders: A healthcare application. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2273–2280. [6](#)

- [Beaulieu-Jones et al., 2018] Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., and Bauer, C. R. (2018). Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR medical informatics*, 6(1):e8960. [3](#)
- [Bennett and Hauser, 2013] Bennett, C. C. and Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *Artificial intelligence in medicine*, 57(1):9–19. [3](#)
- [Black, 1996] Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *Bmj*, 312(7040):1215–1218. [2](#)
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. [42](#), [44](#), [46](#)
- [Brigl, 2018] Brigl, T. (2018). *Extracting Reliable Topics Using Ensemble Latent Dirichlet Allocation*. PhD thesis, Technische Hochschule Ingolstadt. [42](#)
- [Brysbaert, 2019] Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047. [39](#)
- [Bucholtz and Hall, 2004] Bucholtz, M. and Hall, K. (2004). Language and identity. *A companion to linguistic anthropology*, 1:369–394. [40](#)
- [Bureau, 2022] Bureau, U. S. C. (2022). Selected population characteristics. [40](#)
- [Burgess and Ivanič, 2010] Burgess, A. and Ivanič, R. (2010). Writing and being written: Issues of identity across timescales. *Written communication*, 27(2):228–255. [63](#)
- [Carmody et al., 2021] Carmody, J. B., Rosman, I. S., and Carlson, J. C. (2021). Application fever: reviewing the causes, costs, and cures for residency application inflation. *Cureus*, 13(3). [39](#)

- [Celi et al., 2013] Celi, L. A., Mark, R. G., Stone, D. J., and Montgomery, R. A. (2013). “big data” in the intensive care unit. closing the data loop. *American journal of respiratory and critical care medicine*, 187(11):1157. [2](#)
- [Chakraborty et al., 2013] Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723. [9](#)
- [Chang et al., 2018] Chang, C.-H., Mai, M., and Goldenberg, A. (2018). Dynamic measurement scheduling for adverse event forecasting using deep rl. *arXiv preprint arXiv:1812.00268*. [9](#)
- [Chintagunta et al., 2021] Chintagunta, B., Katariya, N., Amatriain, X., and Kannan, A. (2021). Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR. [40](#)
- [Chong and Kuo, 2015] Chong, V. and Kuo, B. C. (2015). Racial identity profiles of asian-white biracial young adults: Testing a theoretical model with cultural and psychological correlates. *Asian American Journal of Psychology*, 6(3):203. [65](#)
- [Demzik et al., 2021] Demzik, A., Filippou, P., Chew, C., Deal, A., Mercer, E., Mahajan, S., Wallen, E. M., Tan, H.-J., and Smith, A. B. (2021). Gender-based differences in urology residency applicant personal statements. *Urology*, 150:2–8. [40](#)
- [Donders et al., 2006] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091. [19](#)
- [Ertefaie et al., 2016] Ertefaie, A., Shortreed, S., and Chakraborty, B. (2016). Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in medicine*, 35(13):2221–2234. [9](#)

- [Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231. [42](#)
- [Fu et al., 2021] Fu, L.-H., Knaplund, C., Cato, K., Perotte, A., Kang, M.-J., Dykes, P. C., Albers, D., and Collins Rossetti, S. (2021). Utilizing timestamps of longitudinal electronic health record data to classify clinical deterioration events. *Journal of the American Medical Informatics Association*, 28(9):1955–1963. [4](#)
- [Futoma et al., 2018] Futoma, J., Lin, A., Sendak, M., Bedoya, A., Clement, M., O’Brien, C., and Heller, K. (2018). Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks. [9](#), [10](#)
- [Gao et al., 2022] Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., and Pearson, A. T. (2022). Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*. [40](#)
- [Gaweda et al., 2005] Gaweda, A. E., Muezzinoglu, M. K., Aronoff, G. R., Jacobs, A. A., Zurada, J. M., and Brier, M. E. (2005). Reinforcement learning approach to individualization of chronic pharmacotherapy. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 3290–3295. IEEE. [9](#)
- [Ghazizadeh et al., 2020] Ghazizadeh, G., Gheibi, M., and Matwin, S. (2020). Cb-dbscan: A novel clustering algorithm for adjacent clusters with different densities. In *Canadian Conference on Artificial Intelligence*, pages 232–237. Springer. [42](#)
- [Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27. [7](#)

- [Groenwold, 2020] Groenwold, R. H. (2020). Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and prognostic research*, 4(1):1–6. [8](#), [12](#)
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. [6](#)
- [Hoffman et al., 2010] Hoffman, M., Bach, F., and Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23. [42](#)
- [Ivanič, 1998] Ivanič, R. (1998). *Writing and identity*, volume 10. Amsterdam: John Benjamins. [40](#), [63](#)
- [Jagannatha et al., 2018] Jagannatha, A., Thomas, P., and Yu, H. (2018). Towards high confidence off-policy reinforcement learning for clinical applications. In *CausalML Workshop, ICML*. [9](#)
- [Jeblick et al., 2022] Jeblick, K., Schachtner, B., Dextl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., et al. (2022). Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*. [40](#)
- [Johnson et al., 2016] Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9. [5](#), [15](#)
- [Johnson and Khoshgoftaar, 2019] Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54. [40](#)
- [Kahkoska et al., 2020] Kahkoska, A. R., Nguyen, C. T., Jiang, X., Adair, L. A., Agarwal, S., Aiello, A. E., Burger, K. S., Buse, J. B., Dabelea, D., Dolan, L. M., et al. (2020). Characterizing the weight-glycemia phenotypes of type 1 diabetes in youth and young adulthood. *BMJ Open Diabetes Research and Care*, 8(1):e000886. [9](#)

- [Kim et al., 2021] Kim, Y., Suescun, J., Schiess, M. C., and Jiang, X. (2021). Computational medication regimen for parkinson’s disease using reinforcement learning. *Scientific Reports*, 11(1):1–9. [9](#)
- [Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. [6](#), [17](#), [19](#)
- [Kitamura, 2023] Kitamura, F. C. (2023). Chatgpt is shaping the future of medical writing but still requires human judgment. [63](#)
- [Komorowski et al., 2018] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720. [9](#)
- [Krakow et al., 2017] Krakow, E. F., Hemmer, M., Wang, T., Logan, B., Arora, M., Spellman, S., Couriel, D., Alousi, A., Pidala, J., Last, M., et al. (2017). Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using q-learning. *American journal of epidemiology*, 186(2):160–172. [9](#)
- [Kramer et al., 2021] Kramer, O., Even, A., Matot, I., Steinberg, Y., and Bitan, Y. (2021). The impact of data quality defects on clinical decision-making in the intensive care unit. *Computer Methods and Programs in Biomedicine*, 209:106359. [3](#)
- [Kung et al., 2023] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al. (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198. [40](#), [63](#)
- [Levine et al., 2020] Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*. [3](#)

- [Li et al., 2018] Li, L., Komorowski, M., and Faisal, A. A. (2018). The actor search tree critic (astc) for off-policy pomdp learning in medical decision making. *arXiv preprint arXiv:1805.11548*. 7, 9
- [Lin et al., 2018] Lin, R., Stanley, M. D., Ghassemi, M. M., and Nemati, S. (2018). A deep deterministic policy gradient approach to medication dosing and surveillance in the icu. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4927–4931. IEEE. 9
- [Lin et al., 2021] Lin, S., Reminick, J., Love, E., Nwomeh, B., and Krishnaswami, S. (2021). Analysis of the pediatric surgery fellowship application process using the thalamus™ database. *Journal of pediatric surgery*, 56(6):1095–1100. 39
- [Little and Rubin, 2019] Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 7
- [Lizotte et al., 2008] Lizotte, D. J., Gunter, L., Laber, E., and Murphy, S. A. (2008). Missing data and uncertainty in batch reinforcement learning. In *Neural Information Processing Systems (NIPS)*. 4
- [Luengo et al., 2012] Luengo, J., García, S., and Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems*, 32:77–108. 7
- [Marbin et al., 2021] Marbin, J., Rosenbluth, G., Brim, R., Cruz, E., Martinez, A., and McNamara, M. (2021). Improving diversity in pediatric residency selection: using an equity framework to implement holistic review. *Journal of Graduate Medical Education*, 13(2):195–200. 38
- [Marivate et al., 2014] Marivate, V. N., Chemali, J., Brunskill, E., and Littman, M. (2014). Quantifying uncertainty in batch personalized sequential decision making. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 9

- [Martín-Guerrero et al., 2009] Martín-Guerrero, J. D., Gomez, F., Soria-Olivas, E., Schmidhuber, J., Climente-Martí, M., and Jiménez-Torres, N. V. (2009). A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. *Expert Systems with Applications*, 36(6):9737–9742. [9](#)
- [Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26. [42](#)
- [Mimno et al., 2011] Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272. [57](#)
- [Moynahan, 2018] Moynahan, K. F. (2018). The current use of united states medical licensing examination step 1 scores: holistic admissions and student well-being are in the balance. *Academic Medicine*, 93(7):963–965. [39](#)
- [Murphy et al., 2016] Murphy, S. A., Deng, Y., Laber, E. B., Maei, H. R., Sutton, R. S., and Witkiewitz, K. (2016). A batch, off-policy, actor-critic algorithm for optimizing the average reward. *arXiv preprint arXiv:1607.05047*. [9](#)
- [Nehemiah et al., 2021] Nehemiah, A., Roberts, S. E., Song, Y., Kelz, R. R., Butler, P. D., Morris, J. B., and Aarons, C. B. (2021). Looking beyond the numbers: increasing diversity and inclusion through holistic review in general surgery recruitment. *Journal of Surgical Education*, 78(3):763–769. [38](#)
- [Nemati et al., 2016] Nemati, S., Ghassemi, M. M., and Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2978–2981. IEEE. [9](#)
- [NRMP, 2021] NRMP (2021). Results of the 2021 nrmp program director survey. [39](#)

- [NRMP, 2022a] NRMP (2022a). 2022 report on residents executive summary. [40](#)
- [NRMP, 2022b] NRMP (2022b). Results of the 2022 nrmp program director survey. [39](#)
- [Oord et al., 2016] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*. [6](#)
- [OpenAI, 2023] OpenAI (2023). Gpt-4 technical report. [40](#)
- [Osman et al., 2015] Osman, N. Y., Schonhardt-Bailey, C., Walling, J. L., Katz, J. T., and Alexander, E. K. (2015). Textual analysis of internal medicine residency personal statements: themes and gender differences. *Medical Education*, 49(1):93–102. [40](#)
- [Ostapenko et al., 2018] Ostapenko, L., Schonhardt-Bailey, C., Sublette, J. W., Smink, D. S., and Osman, N. Y. (2018). Textual analysis of general surgery residency personal statements: topics and gender differences. *Journal of surgical education*, 75(3):573–581. [40](#)
- [Otugo et al., 2021] Otugo, O., Alvarez, A., Brown, I., and Landry, A. (2021). Bias in recruitment: A focus on virtual interviews and holistic review to advance diversity. *AEM Education and Training*, 5(Suppl 1):S135–S139. [38](#)
- [Parbhoo, 2014] Parbhoo, S. (2014). *A reinforcement learning design for HIV clinical trials*. PhD thesis, University of the Witwatersrand, Faculty of Science, School of Computer Science. [9](#)
- [Park and Bahia, 2022] Park, A. S. and Bahia, J. (2022). Examining the experiences of racialized and indigenous graduate students as emerging researchers. *Sociology of Race and Ethnicity*, 8(3):403–417. [51](#)
- [Peng et al., 2018] Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Li-wei, H. L., Ross, A., Faisal, A., and Doshi-Velez, F. (2018). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In

- AMIA Annual Symposium Proceedings*, volume 2018, page 887. American Medical Informatics Association. [9](#)
- [Phetrittikun et al., 2021] Phetrittikun, R., Suvirat, K., Pattalung, T. N., Kongkamol, C., Ingviya, T., and Chaichulee, S. (2021). Temporal fusion transformer for forecasting vital sign trajectories in intensive care patients. In *2021 13th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5. IEEE. [6](#)
- [Piel et al., 2010] Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J., and Hay, S. I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, 1(1):104. [51](#)
- [Pope et al., 2021] Pope, A. J., Carter, K., and Ahn, J. (2021). A renewed call for a more equitable and holistic review of residency applications in the era of covid-19. *AEM Education and Training*, 5(1):135. [38](#)
- [Prasad et al., 2017] Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. (2017). A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*. [9](#)
- [Raghu et al., 2017a] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. (2017a). Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*. [9](#)
- [Raghu et al., 2017b] Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017b). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR. [9](#)
- [Raghu et al., 2018] Raghu, A., Komorowski, M., and Singh, S. (2018). Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*. [9](#)

- [Rajkomar et al., 2018] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18. [3](#)
- [Řehřek et al., 2011] Řehřek, R., Sojka, P., et al. (2011). Gensim—statistical semantics in python. *Retrieved from genism. org*. [43](#)
- [Rodenburg et al., 2019] Rodenburg, F. J., Sawada, Y., and Hayashi, N. (2019). Improving rnn performance by modelling informative missingness with combined indicators. *Applied Sciences*, 9(8):1623. [12](#)
- [Roggeveen et al., 2021] Roggeveen, L., El Hassouni, A., Ahrendt, J., Guo, T., Fleuren, L., Thorat, P., Girbes, A. R., Hoogendoorn, M., and Elbers, P. W. (2021). Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artificial Intelligence in Medicine*, 112:102003. [9](#)
- [Rubin and Schenker, 1986] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American statistical Association*, 81(394):366–374. [7](#)
- [Schafer and Graham, 2002] Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147. [19](#), [20](#)
- [Schulz et al., 2022] Schulz, A. E., Nussbaum, J. E., Loloi, J., Sankin, A., and Abraham, N. (2022). The impact of holistic review of urology residency applications on selection for interview during the covid-19 pandemic. *Urology*. [39](#)
- [Sejnowski, 2020] Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038. [3](#)
- [Shahar, 1999] Shahar, Y. (1999). Knowledge-based temporal interpolation. *Journal of Experimental & Theoretical Artificial Intelligence*, 11(1):123–144. [8](#)

- [Shahriar et al., 2022] Shahriar, A. A., Puram, V. V., Miller, J. M., Sagi, V., Castañón-Gonzalez, L. A., Prasad, S., and Crichlow, R. (2022). Socioeconomic diversity of the matriculating us medical student body by race, ethnicity, and sex, 2017-2019. *JAMA Network Open*, 5(3):e222621–e222621. [51](#)
- [Shortreed et al., 2011] Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., and Murphy, S. A. (2011). Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1):109–136. [9](#)
- [Srinivasan and Doshi-Velez, 2020] Srinivasan, S. and Doshi-Velez, F. (2020). Interpretable batch irl to extract clinician goals in icu hypotension management. *AMIA Summits on Translational Science Proceedings*, 2020:636. [4](#)
- [Stevens et al., 2012] Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 952–961. [59](#)
- [Sun and Wang, 2021] Sun, Y. and Wang, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *Journal of the american statistical association*, 116(533):421–432. [9](#)
- [Tao et al., 2018] Tao, Y., Wang, L., and Almirall, D. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *The annals of applied statistics*, 12(3):1914. [9](#)
- [Tomkins et al., 2008] Tomkins, S., Liao, P., Yeung, S., Klasnja, P., and Murphy, S. (2008). Intelligentpooling: Practical thompson sampling for mhealth; 2020. *URL https://arxiv.org/abs*. [9](#)
- [Van Buuren and Oudshoorn, 1999] Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO. [20](#)

- [van Dis et al., 2023] van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. L. (2023). Chatgpt: five priorities for research. *Nature*, 614(7947):224–226. [63](#)
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. [6](#), [17](#)
- [Weng et al., 2017] Weng, W.-H., Gao, M., He, Z., Yan, S., and Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. *arXiv preprint arXiv:1712.00654*. [9](#)
- [Yoon et al., 2018] Yoon, J., Jordon, J., and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR. [20](#)
- [Yu et al., 2019] Yu, C., Liu, J., and Zhao, H. (2019). Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC medical informatics and decision making*, 19(2):111–120. [9](#)
- [Yu et al., 2020] Yu, C., Ren, G., and Dong, Y. (2020). Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC medical informatics and decision making*, 20(3):1–8. [9](#), [34](#)
- [Zheng et al., 2021] Zheng, H., Ryzhov, I. O., Xie, W., and Zhong, J. (2021). Personalized multimorbidity management for patients with type 2 diabetes using reinforcement learning of electronic health records. *Drugs*, 81(4):471–482. [9](#)

Vita

Senne Van Steelandt is a Ph.D. candidate in the Department of Business Analytics and Statistics at the Haslam College of Business, University of Tennessee, Knoxville. He obtained his Bachelor of Science in Business Engineering from Ghent University in 2016. In 2018, he obtained his Master of Science in Business Engineering: Data Analytics from the same university, with a Master's dissertation titled "The interplay between traditional media, social media, and music streaming." Senne's research is centered around (deep) generative modeling and reinforcement learning, specifically in the contexts of healthcare and digital marketing applications. He has conducted extensive research on the impact of data quality on generative models, resulting in peer-reviewed papers, which he has presented at various conferences and workshops. His current dissertation, titled "Generative Machine Learning Models in Healthcare Applications", continues this line of research. Senne is dedicated to pursuing a career in applied research and is eager to make a positive and responsible impact at the intersection of generative modeling and task automation.