

Evaluating Synthetic Data as a Proxy for Real Clinical Data in Machine Learning Models: A
Comparative Study on Postpartum Hemorrhage Prediction

by

Kam Sharma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Public Health Sciences

Waterloo, Ontario, Canada, 2024

© Kam Sharma 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.
I understand that my thesis may be made electronically available to the public.

Abstract

Introduction:

This thesis investigates the use of synthetic data as a proxy for real clinical data in predictive modeling, focusing on postpartum hemorrhage (PPH). Synthetic data offers a solution to privacy concerns by providing data that mimics real patient data without compromising patient information. The goal is to develop and validate predictive models for PPH using synthetic data and comparing it to the real data, thereby assessing the feasibility and effectiveness of synthetic data in clinical settings.

Methods:

Synthetic data was generated using Generative Adversarial Networks (GANs) from MDClone® to replicate the statistical properties of real clinical data from Ottawa Hospital. The data underwent a thorough cleaning and preparation process, followed by feature selection. Machine learning and statistical models, including logistic regression, decision trees, random forests, and support vector machines, were developed and trained on the synthetic data and then the pipeline was run on the real data at Ottawa Hospital. Model performance was evaluated using precision, recall, F1-score, accuracy, and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

Results:

The synthetic data closely mirrored the real data in statistical properties, with low Hellinger distances for most variables. Machine learning models trained on synthetic data demonstrated high performance, with comparable results to those trained on real data. Key predictors for PPH were determined which included the administration of certain medication and clinical parameters. The comparative analysis showed minimal discrepancies between model outputs from synthetic and real data, validating the use of synthetic data for predictive modeling.

Discussion:

The findings indicate that synthetic data can effectively be used to develop predictive models for PPH, and addressing data accessibility. The study highlights the potential of synthetic data to enhance predictive modeling in healthcare, providing a viable alternative to real data without compromising accuracy. The integration of synthetic data in clinical research can facilitate broader data availability, fostering innovation while adhering to privacy regulations.

Conclusion:

This research demonstrates the viability of synthetic data in predictive modeling for PPH, with models trained on synthetic data showing high performance comparable to those trained on real data. The study contributes to the theoretical understanding of synthetic data utility and offers practical

implications for improving patient outcomes and optimizing healthcare resources. Future research should focus on expanding the use of synthetic data in other clinical areas and further validating its effectiveness in diverse healthcare settings.

Acknowledgements

The journey to completing this thesis has been an adventure filled with challenges, growth, and countless moments of insight. It is with deep appreciation and heartfelt gratitude that I acknowledge those who have accompanied and supported me along the way.

First and foremost, I owe an immeasurable debt of gratitude to my advisor, Dr. Helen Chen. Her wisdom, patience, and unwavering belief in my potential have been the guiding lights that kept me on course. Her ability to challenge my thinking while offering steadfast support has not only shaped this thesis but has also profoundly influenced my development as a scholar.

I am equally thankful to my committee members, Dr. Zahid Butt and Dr. Maura Grossman, whose thoughtful critiques and inspiring suggestions pushed me to refine my research. Their diverse perspectives added layers of depth to my work that I could not have achieved alone.

I also want to express my profound gratitude to my friends and colleagues and the entire cohort of 2024—whose humor, friendship, and encouragement were the perfect antidotes to the stress of graduate life. The shared late-night study sessions and the collective sighs of relief after exams are memories I will cherish.

To my family—words fall short when trying to express the depth of my gratitude. Your constant encouragement, the sacrifices you made, and your belief in me gave me the strength to persevere even when the road was tough. Thank you for being my unwavering support system.

Finally, I would like to acknowledge all those whose names do not appear here but who have nevertheless left an indelible mark on my journey. Whether through a word of encouragement, a shared laugh, or a simple gesture of kindness, your contributions have not gone unnoticed. This thesis is as much a reflection of your support as it is of my effort. Thank you for walking this path with me.

Table of Contents

Author’s Declaration.....	ii
Abstract.....	iii
Acknowledgements.....	v
List of Figures	ix
List of Tables	x
List of Abbreviations	xi
Chapter 1 Introduction	1
1.1 Synthetic Health Data.....	1
1.2 Postpartum Hemorrhage.....	1
1.1.2 Treatment for Postpartum Hemorrhage	2
1.3 Organization of Thesis	3
Chapter 2.....	5
Literature Review.....	5
2.1 Synthetic Data in Healthcare	5
2.1.1 Generating Synthetic Health Data	5
2.1.2 Types of Synthetic Health Data.....	6
2.1.3 Applications of Synthetic Data in Healthcare	7
2.1.4 Advantages and Challenges of Synthetic Data in Healthcare	8
2.2 Synthetic Data Utility Assessment.....	9
2.3 Models Used for Evaluation.....	9
2.4 Synthetic Data Fidelity Assessment.....	10
2.4.1 Hellinger Distance Metrics	10
2.5 Evaluating Model Performance.....	11
2.6 PPH Predictive Models	15
2.6.1 Traditional Statistical Models for PPH Prediction	15
2.6.2 Machine Learning Models for PPH Prediction.....	16
2.6.3 Combining Traditional and ML Models for PPH Prediction	16
2.7 Summary and Implications.....	17
2.8 Study Rational	18
2.9 Gaps in Literature.....	18
2.9.1 Addressing Gaps in PPH Predictive Modeling - Statistical vs ML Models	18
2.9.2 Addressing Gaps in PPH Predictive Modeling – Integration of Synthetic Data	20
2.9.3 Addressing Gaps in PPH Predictive Modeling – Focus on the Canadian Population	21
2.10 Contributions and Implications	21
2.10.1 Theoretical Implications	21
2.10.2 Practical Implications	22
Chapter 3.....	23

Methods.....	23
3.1 Study Design	23
3.2 Step 1: Data Cleaning and Model Development	23
3.2.1 Initial Meeting and Data Acquisition	23
3.2.2 Data Cleaning and Preparation	24
3.2.3 Feature Selection	24
3.2.4 Model Development	25
3.3 Step 2: Comparison of Outputs from Synthetic and Real Data.....	25
3.3.1 Sending the Pipeline to the Data Analyst	25
3.3.2 Running the Models on the Real Data	26
3.3.3 Outcome Comparison and Analysis	26
3.4 Step 3: Factors Associated with Postpartum Hemorrhage	26
3.5 The Dataset.....	27
3.5.1 Missingness and Censored Data	27
3.5.2 Handling the Missing/Censored Data	28
Chapter 4.....	30
Results.....	30
4.1 EDA and Distribution.....	30
4.2 Model Outputs.....	36
4.3 Decision Tree Outputs.....	41
Chapter 5	46
Discussion.....	46
5.1 Cohort Characteristics, EDA, and Distribution of Data between Real and Synthetic Datasets	46
5.1.1 American Society of Anesthesiologists (ASA) score.....	46
5.1.2 Age	47
5.1.3 Campus	47
5.1.4 Gestational Age	47
5.1.5 Medical Flags	48
5.1.6 Hemoglobin	50
5.1.7 Other Conditions.....	51
5.1.8 Drugs Flag	52
5.1.9 Distributions and Bi-Variant Correlations of Data Between Real and Synthetic Datasets.....	53
5.2 Pearson Correlations	54
5.2.1 Overview of Correlations	54
5.2.2 Correlations with PPH Positive Value.....	55
5.3 Cross-Validation.....	56
5.4 Model Performance	57
5.5 Random Forest Best Decision Tree Analysis.....	58
5.5.1 Synthetic Data Decision Tree Analysis	59
5.5.2 Real Data Decision Tree Analysis	60

5.5.3 Similarities and Differences Between Real and Synthetic Decision Tree Pathways	61
5.6 Limitations/Biases and Future Directions	62
5.7 Conclusion.....	63
References.....	65
Appendix A.....	71

List of Figures

Figure 1: Study Design Summary Flowchart.....	23
Figure 2: Synthetic Dataset Missingness	27
Figure 3: Hellinger Distance Across All the Variables	34
Figure 4: Real vs Synthetic Data Pearson Correlation Heatmaps.....	35
Figure 5: Real vs Synthetic Data AUC-ROC Curves	40
Figure 6: Pruned Decision Tree for High-Risk Patient Identification	41
Figure 7: Pathways for Positive and Negative PPH (Synthetic Data)	43
Figure 8: Pathways for Positive and Negative PPH (Real Data)	45

List of Tables

Table 1: Summary of Cohort Characteristics for Synthetic Data (N = 6930)	30
Table 2: Top Variables with the Highest Correlation with PPH Positive.....	35
Table 3: Variables Kept in the Model Based on Feature Selection done on the Synthetic Dataset and Literature	36
Table 4: 5-Fold Cross Validated AUC Scores for Real and Synthetic Dataset	37
Table 5: Performance of Models Trained on Synthetic Dataset	37
Table 6: Performance of Models Trained on Real Data	38

List of Abbreviations

PPH Postpartum hemorrhage

ML Machine Learning

SVM Support Vector Machines

AI Artificial Intelligence

GANs Generative Adversarial Networks

HIPAA Health Insurance Portability and Accountability

PIPEDA Personal Information Protection and Electronic Documents Act

GDPR General Data Protection Regulation

EMR Electronic Medical Records

EDA Exploratory Data Analysis

SMOTE Synthetic Minority Over-Sampling Technique

AUC Area Under the Curve

ROC Receiver Operating Characteristic

ASA Score The American Society of Anesthesiologists Score

Chapter 1 Introduction

1.1 Synthetic Health Data

Synthetic health data refers to artificially generated data that mimics real health data in its structure, statistical properties, and the way it can be used for analysis, without containing any actual patient health information [1]. Access to individual-level health data is essential for developing robust predictive models and conducting thorough secondary analyses. However, privacy concerns and regulatory constraints often limit data accessibility. Synthetic data generation (SDG) offers a promising solution by providing data that mimics real patient data but does not contain any real patient information, thereby mitigating privacy risks [2]. This type of data is beneficial for a variety of purposes, including research, training, testing of health information systems, and development of AI and machine learning models, especially in situations where privacy concerns and data sharing regulations limit the availability of real health data. In this thesis, we explore the utility of synthetic data in developing machine learning models to predict Postpartum Hemorrhage (PPH).

1.2 Postpartum Hemorrhage

PPH is a serious yet preventable condition that poses a significant threat to life following childbirth. PPH is characterized by excessive bleeding (> 1000 mL) from the birth canal, occurring within the first 24 hours after a cesarian delivery [3]. Alarmingly, the prevalence of PPH is increasing, now accounting for 8% of maternal deaths in developed countries and 20% in developing nations [4]. This upward trend includes Canada, where PPH-related mortality has risen by 22% from 5.1% in 2003 to 6.2% in 2010 [5]. This condition demands prompt attention and management due to its potentially severe consequences.

Uterine atony is a primary factor contributing to PPH, a condition characterized by insufficient contraction of the uterus following childbirth, leading to severe bleeding [6]. This condition can stem from various complications, including tears or injuries to the uterus, which are particularly prevalent following cesarian sections [7]. Furthermore, complications such as retained placental tissue or premature detachment of the placenta from the uterine wall can also result in significant bleeding [6]. Additionally, blood clotting disorders, as well as certain medications administered during childbirth, can exacerbate bleeding post-delivery [6, 7]. Key risk factors include advanced maternal age (above 35 years), multiple pregnancies, the use of general anesthesia, and conditions such as hypertension or preeclampsia, which is characterized by high blood pressure

during pregnancy [8–10]. Recognizing these risk factors early is crucial for managing PPH, as it often occurs without warning.

1.1.2 Treatment for Postpartum Hemorrhage

Historically, Oxytocin has been the go-to medication for the treatment of PPH. In recent times, Carbetocin, a synthetic analogue of Oxytocin, has emerged as a highly effective alternative in reducing PPH incidents following c-section deliveries. Both Oxytocin and Carbetocin are classified as uterotonic agents, a group of medications that promote uterine contractions to prevent and manage excessive bleeding after childbirth [11]. It is now widely used in healthcare settings for its efficacy in preventing and treating PPH.

When comparing Oxytocin and Carbetocin, in the management of PPH, several key differences stand out. Oxytocin, a well-established PPH prevention and treatment medication, is favoured for its cost-effectiveness and proven efficacy. It can be administered intravenously or intramuscularly, providing versatility in different clinical settings [12]. However, Oxytocin's short half-life necessitates continuous monitoring to sustain its effects, which can be a drawback in managing uterine tone over time [12].

On the other hand, Carbetocin, a more recent and more expensive addition to PPH management, especially following c-sections, offers the advantage of a prolonged half-life. This leads to sustained uterine contractions, which may be beneficial in preventing PPH with a single intramuscular injection [13]. The longer activity profile of Carbetocin can reduce the need for repeated doses or continuous infusion, often required with Oxytocin. However, these benefits come at a higher cost, with Carbetocin being almost 20 times more expensive than Oxytocin [14]. This cost differential is crucial, particularly in resource-limited healthcare settings with significant budget constraints.

The initial cost for a single dose of Carbetocin may be higher, but a cost-benefit analysis conducted by the UK National Health Services indicates that using Carbetocin over Oxytocin to prevent PPH during c-sections can lead to a decrease of 30 (88 vs 58) PPH events [15]. This translated to monetary savings of £27,518. Further, probabilistic sensitivity analysis suggests that Carbetocin is 91.5% more likely to provide better health outcomes, with a 69.4% probability of being both more cost-effective and more efficacious compared to Oxytocin [15].

Currently, as a precaution, many doctors are administering Carbetocin instead of Oxytocin despite the higher cost, due to its proven efficacy and ease of administration. This widespread use of Carbetocin highlights the need for a more targeted approach to PPH prevention. By predicting who is

most at risk of developing PPH, healthcare providers can better determine who absolutely needs Carbetocin and who can safely receive Oxytocin.

1.3 Organization of Thesis

This thesis explores the viability of synthetic data as a proxy for real clinical data in predictive modeling, with a specific focus on predicting postpartum hemorrhage (PPH). The thesis is structured into several key sections, each addressing a critical aspect of the research.

Chapter 1 introduces the problem of PPH and the need for accurate predictive models in clinical settings. It outlines the significance of using synthetic data to protect patient privacy while ensuring robust model development. The research objectives and the primary goals of evaluating synthetic data's utility are also detailed.

Chapter 2 is a comprehensive review of existing literature on PPH prediction, the use of synthetic data in healthcare, and various predictive modeling techniques is presented. This review provides a foundation for understanding the current state of research and identifies gaps that this study aims to address.

Chapter 3 describes the data sources, including both real and synthetic datasets, and the processes used to prepare these datasets for analysis. It details the feature selection methods employed, including recursive feature elimination and backward elimination, and explains the rationale behind choosing specific variables for the predictive models. The chapter also outlines the modeling techniques used, including logistic regression, decision trees, random forests, support vector machines (SVM).

Chapter 4 presents the findings from the exploratory data analysis, distribution comparisons, and feature correlations. The performance of various predictive models is evaluated using metrics including Precision, Recall, F1-Score, Accuracy and Area Under the Curve Receiver Operating Characteristic Curve (AUC-ROC). Decision tree analysis is used to identify key pathways leading to higher PPH risk, with detailed interpretations of the splits and their clinical implications.

Chapter 5 interprets the results in the context of the research objectives. It addresses the similarities and differences between real and synthetic data distributions, feature correlations, and model performances. The potential of synthetic data as a reliable proxy for real data is assessed, highlighting the strengths and limitations identified in the thesis. Implications for clinical practice and future research are also discussed.

Chapter 6 outlines the limitations of the thesis, including data quality issues, potential biases in feature selection, and challenges in synthetic data generation. It proposes future research directions to address these limitations and enhance the reliability and applicability of synthetic data in predictive modeling.

Chapter 7 summarizes the key findings of the thesis, emphasizing the viability of synthetic data in predictive modeling for PPH. It reiterates the importance of accurate and privacy-preserving data in healthcare research and underscores the potential benefits of integrating synthetic data methodologies with clinical expertise.

By systematically addressing each aspect of the research, this thesis provides a comprehensive evaluation of synthetic data's utility in predictive modeling, contributing valuable insights to the field of healthcare analytics and patient outcome prediction.

Chapter 2

Literature Review

2.1 Synthetic Data in Healthcare

2.1.1 Generating Synthetic Health Data

Artificial intelligence (AI), which refers to the simulation of human intelligence in machines that are programmed to think and learn like humans, is frequently used to produce synthetic health data. In fact, the use of AI models, particularly machine learning algorithms, is central to the generation of synthetic health data [16]. AI models can analyze complex and large datasets, learning from them, and then generating new data points that mimic the statistical properties and patterns of the original data without directly copying it. This capability is particularly important in healthcare, where preserving patient privacy is paramount [17].

Several methods are employed to generate synthetic data, each with its strengths and weaknesses. Common approaches include sequential synthesis using Decision Trees, conditional Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs) [2, 18]. The generation process begins with a thorough data analysis phase, where real health datasets are meticulously examined. This initial step is crucial for understanding data's 'intrinsic structure, identifying any data wrangling or censorship requirements. Such a deep dive into the real data lays the foundation for the subsequent stages of the synthetic data generation process.

Among the AI techniques employed, GANs stand out as a particularly innovative method. GANs consist of two neural networks, the generator and the discriminator, which are trained simultaneously through a competitive process [18]. The generator's goal is to produce synthetic data that is indistinguishable from real data, while the discriminator's role is to distinguish between the generator's synthetic data and the actual data [18]. This competition drives both networks to improve continuously, resulting in the generation of high-quality synthetic data that closely resembles the real data in statistical characteristics but does not carry any risk of revealing personal health information. MDClone®, a data analytics platform designed to enhance data access, analysis, and utilization in the healthcare sector can also uses GANs to generate synthetic health data [19]. This data is not tied to any real individual but retains the statistical properties of the original data. The synthesized data should be indistinguishable from real data in terms of analytical use but completely devoid of any personal health information, thereby safeguarding patient confidentiality [19].

This complex process of generating synthetic health data underscores the balance between utility and privacy, enabling the use of data in healthcare research and development while adhering to ethical standards and privacy regulations. Through this approach, researchers and developers can access a wealth of data that mirrors real-world complexity and variability, essential for training robust AI models and conducting comprehensive health research, without the ethical and privacy concerns associated with using real patient data.

2.1.2 Types of Synthetic Health Data

Machine learning, a subset of artificial intelligence, uses algorithms and statistical models to analyze large datasets, identify patterns, and make predictions [20]. This approach improves performance on specific tasks through experience, making it particularly beneficial in healthcare for applications such as predictive modeling and algorithm training.

Synthetic health data retains the statistical properties and relationships of real-world data, making it useful for predictive modeling, algorithm training, and software testing while. By analyzing complex datasets, machine learning models can learn from them and produce new data points that mimic the original data's statistical properties and patterns, thereby playing a crucial role in generating synthetic health data. There are several types of synthetic health data, each with distinct characteristics and applications.

Simulated data is created using mathematical models and simulations that replicate the behavior and characteristics of real-world health data. These simulations are based on known medical and biological processes and are often used in scenarios where specific conditions or rare events need to be studied without the availability of sufficient real-world data. Simulated data is particularly useful for training machine learning models, conducting epidemiological studies, and testing healthcare software [1].

Anonymization and de-identification are both common techniques used for privacy protection, but they have distinct meanings and implications for data privacy. De-identified data involves removing or obfuscating personally identifiable health information (PHI) from real health records to protect patient privacy. Although it reduces the risk of re-identification, there remains a potential for individuals to be identified if enough data points are available. De-identified data is commonly used in research studies, clinical trials, and healthcare analytics, allowing researchers to analyze patient data while complying with privacy regulations like the Health Insurance Portability

and Accountability Act (HIPAA), Personal Information Protection and Electronic Documents Act (PIPEDA), and General Data Protection Regulation (GDPR) [1, 21].

Anonymization, on the other hand, is the process of transforming data so that it cannot be linked back to an individual by any means, making re-identification virtually impossible [22]. This involves removing both direct and indirect identifiers or altering them sufficiently to prevent re-identification. Techniques for anonymization include data aggregation, data masking, generalization, and suppression. Anonymized data is suitable for public data releases, open data initiatives, and situations where complete privacy is mandatory, ensuring that privacy risks are minimized to the fullest extent [22]. Unlike de-identification, anonymization adheres to stricter privacy protection standards set by regulations like GDPR, which often have specific criteria for what constitutes anonymized data. Thus, while de-identification reduces the risk of re-identification, it does not eliminate it, whereas anonymization ensures that the data cannot be linked back to any individual, providing a higher level of privacy protection [22].

Fully synthetic data, first proposed by Rubin in 1993 and later developed by Raghunathan et al. in 2003, is completely artificial and does not contain any real data [23]. This type of data offers strong privacy control due to the absence of real data, but it has lower analytic value because it may not capture all the complexities and nuances of real-world data [1].

Hybrid/Augmented synthetic data combines elements of real and synthetic data. For example, real patient data might be augmented with synthetic data to create a larger and more diverse dataset. This approach is useful for enhancing the robustness and generalizability of machine learning models. It allows researchers to leverage the strengths of both real and synthetic data while mitigating the limitations of each [1].

2.1.3 Applications of Synthetic Data in Healthcare

Synthetic health data has been increasingly utilized across various healthcare research areas, primarily due to its ability to maintain patient privacy while providing high-quality data for analysis.

Synthetic data is extensively used in simulation studies and predictive analytics, allowing researchers to expand sample sizes or add variables not present in original datasets, thus enhancing the robustness of their models. For instance, disease-specific hybrid simulations and microsimulations for policy testing have leveraged synthetic data to validate models and improve prediction accuracy [1]. Researchers also use synthetic data to develop and test new hypotheses, methodologies, and algorithms without risking patient privacy, particularly useful in the early stages of research when

access to real data may be restricted [1]. In epidemiology and public health research, synthetic data supports studies by providing large datasets that reflect real-world data's statistical properties, enabling the study of population health trends without needing access to confidential records [1]. Additionally, synthetic data aids in the development and testing of health IT systems, including Electronic Health Records (EHR) and clinical decision support systems, ensuring they work correctly before deployment [1, 24]. Organizations can also publicly release synthetic datasets to encourage collaborative research and innovation while ensuring patient privacy is maintained, facilitating broader access to high-quality data for the research community [1].

2.1.4 Advantages and Challenges of Synthetic Data in Healthcare

The advantages of using synthetic data include enhanced privacy protection, increased data availability, compliance with regulatory requirements, and support for innovation and testing. Synthetic data provides a high level of privacy protection as it does not correspond to real individuals, significantly reducing the risk of re-identification, making it suitable for sharing and collaborative research [24, 25]. Synthetic data helps researchers comply with stringent data protection regulations, such as HIPAA, PIPEDA, and GDPR, by mitigating privacy concerns and facilitating easier data sharing across institutions and countries, fostering international collaboration [25]. Additionally, synthetic data supports the development and testing of new algorithms, predictive models, and healthcare technologies in a controlled environment, enabling researchers to refine their approaches before applying them to real-world data [24, 25]

However, using synthetic data also presents challenges, such as ensuring that synthetic data accurately reflects the statistical properties and relationships of real data, which is crucial for drawing reliable conclusions from research findings [25]. The acceptance of synthetic data in clinical practice remains limited due to concerns about its reliability as a substitute for real data in decision-making processes where accuracy is critical [25]. Generating high-quality synthetic data can be resource-intensive, requiring advanced computational tools and expertise, posing a barrier for some research institutions with limited resources [25]. Lastly, despite minimizing re-identification risks, synthetic data can still pose a potential disclosure risk, especially if unique or rare patient attributes are present in the dataset, necessitating ongoing evaluation and refinement of synthetic data generation methods to ensure complete privacy protection [25]

2.2 Synthetic Data Utility Assessment

Khaled El Emam is a prominent figure in the field of health data privacy and the use of synthetic data. His work in evaluating synthetic data, particularly in healthcare, involves rigorous methodologies to ensure that synthetic datasets can be used safely for research. He highlights that evaluating the utility of synthetic data involves a comprehensive comparison of the results obtained from synthetic datasets with those from real datasets [26]. This evaluation process ensures that the synthetic data is a valid proxy for the real data and can be effectively used for research and analysis. Several key metrics are used in this process.

Replicability assesses the ability of synthetic data to replicate the results of analyses performed on real data. High replicability indicates that the synthetic data maintains the same relationships and patterns as the real data, which is crucial for ensuring that conclusions drawn from synthetic data are consistent with those from real data [2]

Decision agreement evaluates whether the conclusions drawn from analyses of synthetic data are consistent with those drawn from real data. It checks if the synthetic data produces the same direction and statistical significance as the real data, ensuring that the decisions made based on synthetic data are valid and reliable [2]

2.3 Models Used for Evaluation

Machine learning and Statistical models are pivotal in predicting outcomes and making data-driven decisions. Among these models, Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression are widely used due to their interpretability and effectiveness.

A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It splits the data into subsets based on the value of input features, creating branches until reaching a decision or a prediction [27]. The Gini impurity measures the degree of impurity or misclassification in a node. A Gini impurity of 0 indicates perfect purity, where all elements belong to a single class [28]. A higher Gini impurity indicates more mixed classes. Nodes with lower Gini values (closer to 0) indicate that the model can better distinguish between patients who are likely to develop PPH and those who are not. These nodes have a more homogenous class distribution. Conversely, nodes with higher Gini values (closer to 0.5) indicate greater impurity, meaning that the node contains a more mixed class distribution, making it harder for the model to distinguish between patients who are likely and unlikely to develop PPH [28]. The pathways and Gini impurity values identified in these decision trees provide clear criteria for assessing PPH risk.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees [29]. This model improves predictive accuracy and controls over-fitting by averaging multiple trees' results. Random Forest is robust against overfitting and can handle large datasets with higher dimensionality [29]. Additionally, it provides insights into feature importance, helping to identify which features contribute most to the prediction [29].

Support Vector Machine (SVM) is a supervised learning model that analyzes data for classification and regression analysis [30]. It constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. SVM uses kernel functions to transform data into higher dimensions, making it possible to find a linear separation in a transformed space even if the original space is not linearly separable [30]. SVM aims to maximize the margin between different classes, providing a robust decision boundary.

Logistic Regression is a statistical model that is commonly used for binary classification problems [31]. It models the probability that a given input point belongs to a certain class. The logistic function, also known as the sigmoid function, is used to map predicted values to probabilities [31]. Logistic Regression is interpretable and straightforward, providing coefficients that indicate the strength and direction of the relationship between input features and the target variable. It is effective for binary classification tasks and can be extended to multiclass classification through techniques like one-vs-rest and one-vs-one approaches [31]. Despite its simplicity, Logistic Regression often performs well with linearly separable data and provides a good baseline model for classification problems.

2.4 Synthetic Data Fidelity Assessment

2.4.1 Hellinger Distance Metrics

To ensure that the synthetic data was a valid proxy for the real data, the Hellinger distance metrics was conducted. The Hellinger distance is a measure of the difference between two probability distributions [26]. By comparing the distributions of the synthetic and real data, the analyst could quantify how closely the synthetic data matched the real data. A low Hellinger distance indicated that the synthetic data accurately represented the distribution of the real data, validating its use in training the predictive models.

$$HD(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2}$$

(1)

- HD(P, Q) Hellinger distance between the P and Q distributions.
- P denotes the distribution of the “real data” and Q denotes the distribution of the synthetic data.
- p_i and q_i they are the probabilities of the i th event or outcome in the two distributions P and Q, respectively.
- $1/\sqrt{2}$: This is a scaling factor, included to ensure that the Hellinger distance falls in the range $[0, 1]$, with 0 indicating identical distributions and 1 indicating completely dissimilar distributions.

2.5 Evaluating Model Performance

In evaluating the performance of predictive models for PPH, several key metrics were utilized.

Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify all relevant instances, specifically the proportion of actual positives correctly identified by the model [32].

Recall is defined as the ratio of true positive (TP) predictions to the sum of true positives and false negatives (FN). Mathematically, it is expressed as:

$$Recall = \frac{TP}{TP + FN}$$

(2)

- **TP (True Positives)**: The number of positive instances correctly predicted by the model.
- **FN (False Negatives)**: The number of positive instances incorrectly predicted as negative by the model.

Recall focuses on the model's ability to identify positive instances correctly. It is particularly important in situations where missing positive instances is costly or undesirable, such as in medical

diagnoses, fraud detection, and other critical applications [32]. High recall indicates that the model successfully identifies most of the positive cases, though it may come at the expense of precision, meaning it might also identify some negative instances as positive [32].

Precision measures the accuracy of the positive predictions made by the model, indicating the proportion of true positive predictions out of all positive predictions made [32].

Precision is defined as the ratio of true positive (TP) predictions to the sum of true positives and false positives (FP) [32]. Mathematically, it is expressed as:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- **TP (True Positives):** Positive instances correctly predicted.
- **FP (False Positives):** Negative instances incorrectly predicted as positive.

Precision focuses on the accuracy of the model's positive predictions. It is particularly important in situations where the cost of false positives is high, such as in spam detection, where labeling a legitimate email as spam is undesirable. High precision indicates that the model's positive predictions are reliable, though it may come at the expense of recall, meaning it might miss some positive instances [32].

F1-Score provides a single metric that balances both precision and recall, offering a more comprehensive view of the model's performance, especially in cases where there is an uneven class distribution [33].

The F1-Score is defined as the harmonic mean of precision and recall. Mathematically, it is expressed as:

$$F1 - Score = 2 \times \frac{Precision + Recall}{Precision \times Recall} \quad (4)$$

- Precision is the ratio of true positive (TP) predictions to the sum of true positives and false positives (FP). It measures the accuracy of the positive predictions made by the model.
- Recall is the ratio of true positive predictions to the sum of true positives and false negatives (FN). It measures the model's ability to correctly identify all relevant instances.

The F1-Score provides a balance between precision and recall, ensuring that both metrics are taken into account [33]. It is particularly useful in scenarios where the class distribution is

imbalanced, and a high score indicates that the model has a good balance between identifying positive instances and minimizing false positives. High F1-Score suggests that the model is performing well in terms of both precision and recall, making it a valuable metric in applications such as medical diagnoses, information retrieval, and binary classification tasks [33].

Accuracy is a metric used to evaluate the performance of a classification model. It measures the overall correctness of the model's predictions by determining the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy is defined as the ratio of correctly predicted instances (both true positive and true negative) to the total number of instances. Mathematically, it is expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

- TP (True Positives): The number of positive instances correctly predicted by the model.
- TN (True Negatives): The number of negative instances correctly predicted by the model.
- FP (False Positives): The number of negative instances incorrectly predicted as positive by the model.
- FN (False Negatives): The number of positive instances incorrectly predicted as negative by the model.

Variables:

- TP (True Positives): Correctly predicted positive cases.
- TN (True Negatives): Correctly predicted negative cases.
- FP (False Positives): Incorrectly predicted positive cases.
- FN (False Negatives): Incorrectly predicted negative cases.

Accuracy provides a straightforward measure of the proportion of correct predictions made by the model out of all predictions [34]. It is particularly useful when the class distribution is balanced. However, in scenarios with imbalanced classes, accuracy can be misleading, as it does not account for the distribution of false positives and false negatives [34]. High accuracy indicates that the model performs well overall, but additional metrics such as precision, recall, and F1-Score are often necessary to get a more nuanced understanding of model performance in cases of class imbalance [34].

AUC-ROC provides an aggregate measure of performance across all classification thresholds, reflecting the model's ability to distinguish between positive and negative classes [35]. AUC-ROC is defined as the area under the ROC curve, which is a plot of the true positive rate (TPR or recall) against the false positive rate (FPR) at various threshold settings. Mathematically, it is expressed as the integral of the ROC curve [34].

$$AUC\ ROC = \int_0^1 TPR(FPR)d(FPR)$$

(6)

In this context:

- TPR (True Positive Rate or Recall): The ratio of true positive (TP) predictions to the sum of true positives and false negatives (FN). It measures the ability of the model to correctly identify positive instances.

$$TPR = \frac{TP}{TP + FN}$$

(7)

- FPR (False Positive Rate): The ratio of false positive (FP) predictions to the sum of false positives and true negatives (TN). It measures the likelihood of the model incorrectly identifying negative instances as positive.

$$FPR = \frac{FP}{FP + TN}$$

(8)

Variables:

- TP (True Positives): The number of positive instances correctly predicted by the model.

- FP (False Positives): The number of negative instances incorrectly predicted as positive by the model.
- FN (False Negatives): The number of positive instances incorrectly predicted as negative by the model.
- TN (True Negatives): The number of negative instances correctly predicted by the model.

The AUC-ROC metric provides a single scalar value that summarizes the performance of the model across all possible thresholds [34]. AUC values range from 0 to 1, with a value of 1 indicating perfect discrimination between positive and negative instances and a value of 0.5 suggesting no discriminative power (equivalent to random guessing) [34]. High AUC values indicate that the model has a good measure of separability, meaning it is effective in distinguishing between the classes. This metric is particularly useful for evaluating models on imbalanced datasets, where accuracy might be misleading [34].

All these metrics were computed automatically using standard functions available in machine learning libraries such as scikit-learn, ensuring consistency and accuracy in the evaluation process.

2.6 PPH Predictive Models

Various predictive models have been developed in the past 10 years to identify women at risk of developing PPH. These models utilize a range of methodologies, from traditional statistical analyses to advanced machine learning techniques. Despite the advancements, the current state of PPH predictive models reveals several challenges and limitations, particularly in their clinical applicability and generalizability.

2.6.1 Traditional Statistical Models for PPH Prediction

Traditional predictive models for PPH typically rely on variations of logistic regression. These models use predefined clinical and demographic variables to estimate the risk of PPH. For instance, the study by Koopmans et al. developed models incorporating antepartum and intrapartum variables to predict PPH in women with gestational hypertension or mild preeclampsia [36]. While these models provided moderate discrimination (AUC of 0.59 and 0.64), their clinical utility is limited by moderate predictive power and reliance on specific patient cohorts. This limitation suggests a need for models that can handle more complex data.

2.6.2 Machine Learning Models for PPH Prediction

Recent advancements have seen the application of machine learning (ML) to predict PPH, offering improved accuracy and the ability to handle large datasets with numerous variables. Ranjbar et al. conducted a systematic review of ML-based PPH prediction models, identifying several algorithms that showed promising results [37]. For example, the Gradient Boosting Model achieved an AUC of 0.93 in one study, significantly outperforming traditional statistical methods. These ML models identified numerous predictors, such as maternal weight, fetal macrosomia, and hematocrit levels, which were critical in enhancing predictive accuracy.

Similarly, Liu et al. developed a predictive model using a combination of Light Gradient Boost and Logistic Regression [38]. This model incorporated both clinical and demographic factors, providing a robust tool for predicting severe PPH in a large cohort of vaginal deliveries.

Various machine learning models such as Random Forest, Decision Trees, and SVM have also been applied to predict PPH. These models have shown varying degrees of success. For instance, the Random Forest model demonstrated a high performance with an AUC of 0.92, and the Decision Tree model had a moderate performance with an AUC of 0.64. Support Vector Machines, while promising, often require careful tuning of parameters to achieve optimal performance. These models leverage the power of large datasets and complex variable interactions, enhancing the predictive accuracy and reliability of PPH risk assessments [37].

2.6.3 Combining Traditional and ML Models for PPH Prediction

The integration of traditional statistical methods with ML techniques can significantly enhance predictive performance in healthcare. Traditional methods, such as logistic regression, are user-driven and focus on inferring relationships between variables, which makes them highly interpretable and valuable for understanding underlying patterns in data. However, these methods often struggle with high-dimensional data and complex interactions among variables. On the other hand, ML models, such as Random Forests, and SVM excel in handling large datasets with numerous predictors and capturing non-linear relationships without needing predefined models [39, 40]

By combining the strengths of both approaches, researchers can create hybrid models that leverage the interpretability of traditional methods and the predictive power of ML techniques. For example, a hybrid model might use logistic regression for initial feature selection and relationship inference, followed by an ML model for refining predictions and handling complex interactions. This

integration can lead to more robust and accurate predictive models that are both interpretable and powerful [39, 40].

2.7 Summary and Implications

Synthetic health data emerges as a crucial tool in the context of predictive modeling, offering a means to generate high-quality, privacy-preserving datasets essential for training and validating models. For this thesis, synthetic health data generated by MDClone®, which utilizes GANs to create data that maintains the statistical properties and correlations found in real datasets, will be used. This approach allows us to generate data that is analytically indistinguishable from real data while being devoid of any personal health information, thereby safeguarding patient confidentiality.

The integration of synthetic health data into predictive modeling for PPH presents a groundbreaking opportunity to overcome the limitations of real-world data, such as accessibility and privacy concerns. Rigorous methodologies are essential to ensure that synthetic datasets can safely and effectively be used for research. Evaluation metrics, including fidelity, replicability, and decision agreement, ensure that synthetic data maintains high utility, meaning it can be used effectively for research without significant loss of accuracy or reliability compared to real data.

With the foundation of robust synthetic data, the current landscape of PPH predictive models shows a significant evolution from traditional statistical methods to more sophisticated machine learning techniques. Traditional statistical models, while interpretable, often struggle with high-dimensional data and complex interactions, which are better handled by machine learning models. However, the integration of both approaches holds promise for developing more robust and accurate predictive models that are both interpretable and powerful.

Overall, the combination of advanced predictive models and synthetic health data, particularly using MDClone®'s GAN-generated data, presents a promising avenue for enhancing PPH prediction and management. This approach can potentially optimize resource allocation, improve patient outcomes, and drive innovation in healthcare while maintaining strict privacy standards. By addressing the identified gaps in the literature, such as the need for comprehensive validation studies and the integration of synthetic data, this thesis aims to develop more accurate, generalizable, and privacy-preserving predictive models that can be effectively used in clinical practice. The findings from this research will not only contribute to the field of maternal healthcare

but also set a precedent for the broader adoption of synthetic health data in medical research, fostering innovation and improving healthcare outcomes globally.

2.8 Study Rational

The purpose of this thesis is to develop a predictive model for PPH and evaluate the usability of synthetic health data in comparison to real health data. By leveraging advanced machine learning techniques, the thesis aims to enhance early detection and management of PPH. Additionally, it will assess the effectiveness of synthetic data in predictive modeling, providing insights into its potential as a viable alternative to real patient data while maintaining patient privacy and adhering to regulatory constraints. Furthermore, this thesis will aim to predict who is most at risk of developing PPH and would benefit most from receiving Carbetocin instead of Oxytocin, optimizing resource allocation and potentially reducing healthcare costs.

This thesis aims to contribute to the existing body of knowledge by developing and validating predictive models for postpartum hemorrhage (PPH) that leverage both real and synthetic health data, addressing key gaps in model generalizability and data integration. By tailoring these models to the Canadian healthcare context, this research will enhance early detection and management of PPH, thereby improving patient outcomes and optimizing healthcare resource allocation. Additionally, the synthetic data utility assessment will demonstrate that synthetic health data, such as that generated by MDClone® using GANs, can be just as reliable as real data for medical and clinical research. This is crucial as real-world data is often difficult to access due to privacy concerns and regulatory restrictions. Proving the reliability of synthetic data will provide a viable alternative for researchers, expanding data availability and fostering innovation while maintaining strict privacy standards.

2.9 Gaps in Literature

The current literature on predictive modeling for PPH identifies several key gaps that need to be addressed to improve the accuracy and applicability of these models.

2.9.1 Addressing Gaps in PPH Predictive Modeling - Statistical vs ML Models

The current literature on PPH predictive modeling highlights several critical gaps that need to be addressed to improve model performance and applicability.

The reasons for the poorly performing traditional statistical models for PPH prediction are multifaceted. First, PPH is influenced by a wide range of factors, including maternal, fetal, and labor-related variables. Traditional statistical methods like logistic regression are limited in handling the

complex interactions between these variables, especially when there are many predictors [41]. This complexity reduces the effectiveness of traditional models, leading to fewer studies using these methods. Second, traditional models require high-quality, consistent data with well-defined variables[41]. In many cases, clinical data can be inconsistent or incomplete, limiting the ability of traditional statistical methods to produce reliable models. Third, the rise of machine learning and artificial intelligence offers more sophisticated tools for modeling complex phenomena like PPH. These methods can handle larger datasets and more complex interactions, leading researchers to favor them over traditional statistical methods [42]. Finally, traditional models often have moderate predictive power, as evidenced by the modest AUC values reported by Koopman et al., above. This performance gap further discourages the use of traditional methods when more powerful machine learning techniques are available [36].

The application of AI and ML techniques offers more sophisticated tools for modeling complex phenomena like PPH. These methods can handle larger datasets and more complex interactions, leading researchers to favor them over traditional statistical methods [42]. ML models, such as Random Forest, and Decision Trees are designed to handle large datasets with many variables, capturing complex interactions between them. These models can also manage multicollinearity, which poses significant challenges for traditional models, leading to more stable estimates and better predictive performance [42].

Using ML models offers several advantages over traditional statistical methods, particularly in the context of complex and multifactorial conditions like PPH. Traditional statistical methods, such as logistic regression, often struggle with high-dimensional data where the number of predictors is large relative to the number of observations. In contrast, ML models like Random Forest, and Decision Trees are designed to handle large datasets with many variables, capturing complex interactions between them. Traditional models typically assume linear relationships between predictors and the outcome, which may not adequately capture the true complexity of medical data [41]. ML models can capture non-linear relationships and interactions between variables without needing to specify these relationships in advance, resulting in higher predictive accuracy [42]. For instance, Gradient Boosting Models and Random Forests have shown higher AUC values in PPH prediction [38].

Furthermore, ML models often include mechanisms for automatic feature selection, reducing the risk of overfitting and improving model performance [43]. They are also designed to scale efficiently with data size, making them suitable for large datasets commonly encountered in healthcare. Multicollinearity, which can pose significant challenges for traditional statistical models, leading to unstable estimates and difficulties in interpretation, is effectively managed by ML models like Random Forests and ensemble methods [42]. These models provide more nuanced predictive insights, allowing for better risk stratification and personalized predictions, which are crucial in healthcare settings for individualized risk assessments.

Moreover, ML techniques are widely used across various disciplines, facilitating the transfer of knowledge and methodologies between fields, leading to continuous improvements in predictive modeling. For example, studies in PPH prediction have shown that ML models such as Random Forest and Gradient Boosting achieve higher AUC values and better predictive performance compared to traditional logistic regression models [38]. These ML models can incorporate a wide range of predictors, including clinical, demographic, and even genomic data, providing a more comprehensive risk assessment for PPH.

2.9.2 Addressing Gaps in PPH Predictive Modeling – Integration of Synthetic Data

Another critical gap found in the literature is the lack of integration of synthetic health data in predictive modeling. While synthetic data has shown promise in expanding data availability and ensuring privacy, its application in PPH predictive modeling is still limited. Most studies continue to rely on real-world data, which can be difficult to access due to privacy concerns and other restrictions. Additionally, there is a need for more research on the effective combination of synthetic and real data to enhance model robustness and accuracy.

This thesis aims to address this gap by integrating fully synthetic health data into the predictive modeling process for PPH. We will use MDClone®'s GAN-generated synthetic health data to create datasets that replicate the statistical characteristics of real-world PPH data, ensuring that the synthetic data is representative and useful for model training and validation. The thesis will rigorously validate the predictive models developed using fully synthetic data, comparing their performance with those trained on real data to evaluate the effectiveness of synthetic data in enhancing model robustness and accuracy.

2.9.3 Addressing Gaps in PPH Predictive Modeling – Focus on the Canadian Population

Many predictive models for PPH have been developed and validated in international contexts where population demographics, healthcare delivery systems, and resource availability significantly differ from those in Canada. These differences can lead to models that do not accurately reflect the risk factors and healthcare environment faced by Canadian patients. For example, differences in racial and ethnic compositions, socioeconomic factors, and access to healthcare services can all influence the incidence and management of PPH. Models developed in countries with different healthcare systems may not account for these unique Canadian factors, resulting in lower predictive accuracy and reliability when applied to Canadian populations.

This thesis aims to address this gap by focusing specifically on the Canadian population, using data from Canadian healthcare institutions to develop and validate predictive models for PPH. Data from Ottawa Hospital was used to ensure that the models are based on the specific demographic and clinical factors relevant to the Canadian context. This will help improve the accuracy and relevance of the predictive models for Canadian patients. Secondly, engaging with healthcare providers across Canada will be crucial for gathering comprehensive data and validating the models. Collaboration with Canadian institutions will also help ensure that the models are practical and implementable within the Canadian healthcare system.

2.10 Contributions and Implications

2.10.1 Theoretical Implications

The thesis aims to advance theoretical understanding in several ways. First, by validating ML models across diverse clinical settings, the thesis will contribute to the development of more generalizable predictive models for PPH. This effort will enhance the theoretical framework around model generalizability, highlighting the importance of testing models in varied environments to ensure their broad applicability. Second, the thesis will provide insights into the effective combination of synthetic and real health data, contributing to the theoretical framework for data integration in predictive modeling. This will demonstrate how synthetic data can complement real data to improve model robustness and accuracy, setting new standards for future research in the field.

2.10.2 Practical Implications

The practical implications of this thesis are significant. By providing more accurate and timely predictions of PPH risk, the thesis aims to improve patient outcomes through better-informed clinical decisions. The use of predictive models can help optimize the allocation of medical resources, ensuring that high-risk patients receive appropriate interventions, thereby improving efficiency in healthcare delivery. Furthermore, the findings from this thesis can inform healthcare policies and programs aimed at reducing the incidence and severity of PPH, particularly in the Canadian context. By demonstrating the utility of synthetic health data, the thesis will encourage its broader adoption in medical research, fostering innovation and accelerating the development of effective clinical tools.

In summary, this thesis not only addresses key gaps in the current literature on PPH predictive modeling but also has the potential to transform clinical practices and healthcare policies through the innovative use of synthetic health data and advanced machine learning techniques.

Chapter 3

Methods

3.1 Study Design

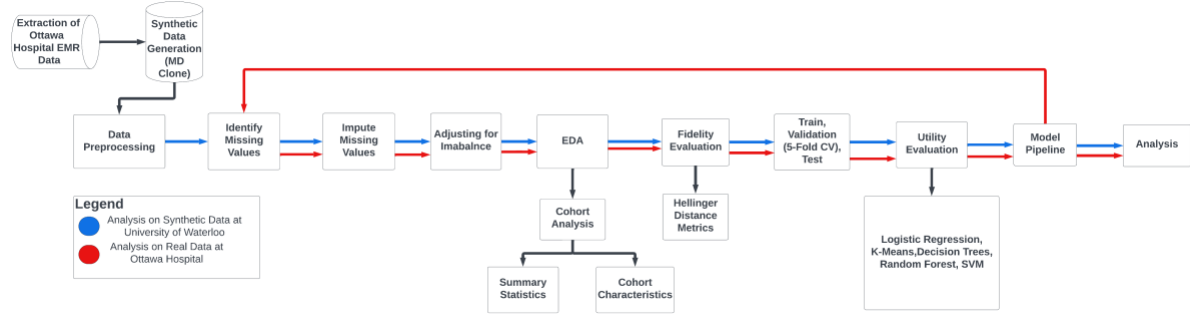


Figure 1: Study Design Summary Flowchart

The design of this thesis involves a comprehensive approach to developing and validating predictive models for PPH using synthetic health data generated from real-world data. The thesis is structured into three main steps: Step 1: Data Cleaning and Model Development, Step 2: Comparison of Outputs from Synthetic and Real Data, and Step 3: Determination of Individuals at the Highest Risk of Developing PPH. All of these major steps can be visualized in Figure 1.

3.2 Step 1: Data Cleaning and Model Development

3.2.1 Initial Meeting and Data Acquisition

The study began with a series of detailed meetings with a data analyst at Ottawa Hospital to discuss the dataset requirements and establish the objectives of the predictive modeling project. During these meetings, the specific variables needed for the study were identified, including demographic information, clinical parameters, and PPH outcomes. The data analyst at Ottawa Hospital played a crucial role in extracting relevant real-world data from the hospital's Electronic Medical Records (EMR) system. This dataset included comprehensive patient information, capturing various aspects such as age, medical history, delivery details, and outcomes related to PPH.

The real-world data, containing valuable patient information and PPH outcomes, was then utilized to create a synthetic dataset. Using MDClone®'s platform, the data analyst generated this synthetic dataset by leveraging GANs. The synthetic data was meticulously crafted to maintain the statistical properties and correlations of the real-world data, ensuring that it accurately mirrored the distribution and relationships found in the original dataset. This step was crucial in preserving the

integrity and utility of the data for subsequent analysis while ensuring patient privacy was uncompromised.

3.2.2 Data Cleaning and Preparation

Upon receiving the synthetic dataset, a thorough data-cleaning process was undertaken. The dataset was first checked for missing values. Given the ability to adjust synthetic data to minimize missingness, the incidence of missing values was low. With less than 10% of the data missing, it was determined that multiple imputation was not necessary. The 10% of the missing data was handled by imputing the missing information by using the mean and mode. This decision was based on standard data analysis practices, ensuring that the integrity and accuracy of the dataset were maintained without introducing potential biases through imputation.

After adjusting for missingness, the data was then normalized to ensure that features were on a comparable scale. Exploratory data analysis (EDA) was performed, which involved visualizing the distributions of various features, identifying any anomalies or outliers, and understanding the relationships between different variables. This step was essential for gaining insights into the dataset, identifying potential issues, and informing subsequent modeling steps.

Addressing the imbalance in the target variable (PPH) was another critical step. With only 10% of patients having PPH as an outcome, the dataset exhibited a severe class imbalance. To achieve better performance of the model, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE is a well-established method that synthesizes new examples for the minority class, thereby creating a more balanced dataset. This adjustment was essential to improve the model's ability to accurately predict PPH outcomes without being biased towards the majority class.

3.2.3 Feature Selection

Feature selection was performed using a combination of methodologies to identify the most relevant variables for the predictive models, ensuring a thorough and robust process. The methodologies included a detailed literature review to identify variables commonly associated with PPH based on existing research, providing a strong theoretical foundation. Iterative backward elimination was employed to remove the least significant variables and reassess model performance, which enhanced the selection process. The threshold was specified to a p-value below 0.05, indicating that the variables are statistically significant predictors. Decision Tree analysis was utilized to capture non-linear relationships and interactions between variables, offering insights into the hierarchical importance of predictors. Recursive Feature Elimination (RFE) systematically removed features and

built models to determine the optimal subset of features, ensuring the final model retained only the most critical predictors. By integrating these approaches, the study ensured a comprehensive and informed selection of features, enhancing the robustness and reliability of the predictive models for accurately identifying the risk of PPH.

3.2.4 Model Development

Various machine learning models were developed and trained on the synthetic data to predict PPH outcomes. The models included logistic regression, Random Forest, Decision Tree, and SVM. These models were chosen to explore different algorithmic approaches and identify the most effective methods for predicting PPH. To adjust for overfitting, several strategies were employed, including cross-validation, regularization techniques, and pruning methods for tree-based models. Cross-validation involved partitioning the dataset into training and validation sets multiple times to ensure the models generalized well to unseen data.

The data was split into training, testing, and validation sets based on standard analysis practices. 80% of the data was used for training, 10% for testing, and 10% for the 5-fold cross-validation. This split ensures that the models are trained on a substantial portion of the data, validated on a separate set to tune hyperparameters and prevent overfitting, and tested on yet another set to evaluate the final performance.

Automatic hyperparameter tuning was performed to optimize model performance, and the model performance was then evaluated using metrics such as, precision, recall, F1 score and AUC-ROC . These steps ensured that the models were not only accurate but also robust and generalizable.

3.3 Step 2: Comparison of Outputs from Synthetic and Real Data

After training and validating the machine learning models on the synthetic data, the next step was to compare the performance of these models when applied to real-world data. This step was crucial to ensure that the synthetic data accurately represented real-world scenarios and that the models could generalize well to actual clinical data.

3.3.1 Sending the Pipeline to the Data Analyst

The entire modeling pipeline, including the trained models and the preprocessing steps, was sent to the data analyst at Ottawa Hospital. This pipeline included all aspects of the data processing and model training steps, ensuring that the models could be applied directly to the real-world data without requiring any additional modifications.

3.3.2 Running the Models on the Real Data

The data analyst at Ottawa Hospital then applied the pipeline to the real-world data extracted from the hospital's EMR system. This step involved running the same machine learning models that were trained on the synthetic data to predict PPH outcomes using the real-world data. This process helped in validating the applicability and effectiveness of the models in a real clinical setting.

3.3.3 Outcome Comparison and Analysis

The final step involved a detailed comparison of the outcomes from the models run on both synthetic and real data. This comparison helped in identifying any discrepancies or areas where the models might need further adjustment. By analyzing the performance metrics and the Hellinger distance, the thesis ensured that the predictive models were robust and could be confidently used in clinical practice to identify patients at risk of PPH.

Through this rigorous comparison process, the thesis aims to demonstrate that synthetic data could effectively be used to develop reliable predictive models, addressing data accessibility and privacy concerns while ensuring the models' applicability in real-world settings.

3.4 Step 3: Factors Associated with Postpartum Hemorrhage

Finally, the outcomes from the models run on both synthetic and real data were compared to identify key predictors for developing PPH. This step focused on determining the individuals at the highest risk of developing PPH based on the combined insights from both datasets. By identifying these high-risk individuals, the thesis aims to provide reliable tools for clinical decision-making, enabling targeted interventions and improving maternal health outcomes.

3.5 The Dataset

The dataset used in this study contains a comprehensive set of 72 variables that capture various aspects of patient demographics, clinical parameters, medication usage, and outcomes related to postpartum hemorrhage (PPH). The dataset includes data from 6930 patients. The data was collected from both the Civic and General campuses of Ottawa Hospital from their obstetrics department on women who have had C-sections. This data collection period spans from July 2019 to December 2022. A detailed data dictionary that defines each variable in the dataset, along with the type of data contained in each column, can be found in Appendix Table 1A: Variables and Data Dictionary.

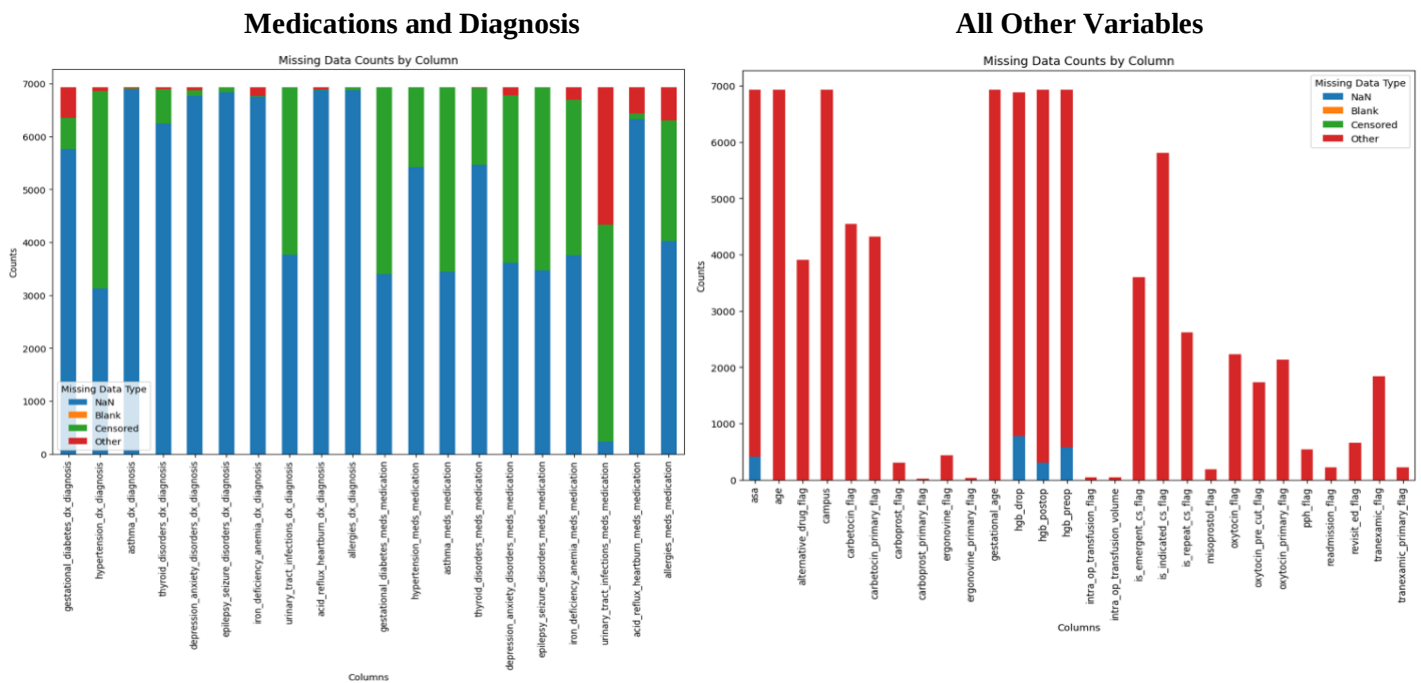


Figure 2: Synthetic Dataset Missingness

3.5.1 Missingness and Censored Data

Upon receiving the synthetic dataset, a thorough data-cleaning process was undertaken to address missing values. Given the ability to adjust synthetic data to minimize missingness, the incidence of missing values was relatively low. The dataset was first checked for missing values across all variables.

Two visual representations were created to highlight the distribution of missing data by column. The first plot focuses on Medication/Diagnosis variables, while the second plot shows the missing

data counts for All Other Variables. These plots categorize missing data into different types: NaN, Blank, Censored, and Other.

For the Medication/Diagnosis graph, the "Other" category indicates that the information is not missing and represents valid data points such as 'yes' (1) or the names of the ICD-10 medications/diagnosis codes. In this context, NaN means that the cell is blank, indicating that the individual did not receive that diagnosis or medication which also means 'no' (0). There were no blank cells in this category. "Censored" means that the cell contains the word 'censored', which signifies that the individual has the diagnosis or is taking the medication, but specific details are withheld for privacy reasons. This makes censored data Missing Not at Random (MNAR), as the missingness is directly related to the nature of the data.

For All Other Variables, NaN means that the cell was blank, indicating missing information. Since these NaN values occur without any specific pattern, they can be considered Missing Completely at Random (MCAR). There were no blank cells and no censored cells in this category. The "Other" category in these variables indicates valid data points such as 'yes' (1) or 'no' (0).

3.5.2 Handling the Missing/Censored Data

In this dataset, NaN values and blank values are treated equivalently. The missing data accounted for less than 10% of the total data. Given this low percentage, multiple imputation was deemed unnecessary. Instead, the missing values were handled by imputing the missing information using the mean (for continuous variables) and mode (for categorical variables). This approach ensures that the integrity and accuracy of the dataset are maintained without introducing potential biases through imputation.

The dataset exhibited an extremely high percentage of censored variables, which required careful consideration. Censored data in this context indicates that a patient had a certain diagnosis or medication, but specific details were withheld to maintain privacy. For example, under the column 'Gestational Diabetes Dx-Diagnosis', if the value is 'censored', it signifies that the patient had a diagnosis related to gestational diabetes, but the exact ICD-10 code is not disclosed due to the small number of patients with that specific code. Therefore, there were three different options to handle this censored data.

One approach involved treating all censored values as 'yes' (or 1). This method was based on the guidance provided by the data analyst at Ottawa Hospital, who explained that censored values could be interpreted as indicating the presence of a condition or medication. Another approach was to remove all instances of censored data from the dataset. This drastically reduced the dataset to

approximately 2000 patients, significantly limiting the dataset's size and its representativeness. A third attempt involved selectively removing patients with multiple censored data points, retaining only those with a single censored data point. Not only did this also drastically reduce the dataset to only about 2000 patients but the resulting models were still very similar to those where all censored data was treated as 'yes'.

Given these outcomes, it was determined that treating all censored values as 'yes' was the most practical and effective approach. This decision was based on the fact that both removing censored data and selectively removing censored data resulted in models that were very similar to those obtained by treating censored values as 'yes'. Given this outcome, it was deemed unnecessary to further adjust for censored data, as the simple approach of converting censored values to 'yes' provided consistent and reliable results.

Chapter 4

Results

4.1 EDA and Distribution

Table 1: Summary of Cohort Characteristics for Synthetic Data (N = 6930)

Variables	Positive PPH Patients, n (%)	Negative PPH Patients, n (%)	All Patients, n (%)
ASA Score			
ASA = 1.0	10 (0.14%)	173 (2.50%)	183 (2.64%)
ASA = 2.0	181 (2.61%)	2665 (38.45%)	2846 (41.06%)
ASA = 3.0	307 (4.43%)	3292 (47.48%)	3599 (51.91%)
ASA = 4.0	36 (0.52%)	236 (3.41%)	272 (3.92 %)
ASA = 5.0	7 (0.10%)	23 (0.33%)	30 (0.43%)
Age Groups			
Age 17-25	47 (0.68%)	505 (7.28%)	552 (7.96%)
Age 26-35	262 (3.76%)	3472 (49.44%)	3689 (53.22%)
Age 36-45	167 (2.41%)	2417 (34.87%)	2584 (37.29%)
Age 46+	8 (0.12%)	97 (1.40%)	105 (1.52%)
Campus = 1	306 (4.42%)	2820 (40.69%)	3126 (45.11%)
Campus = 0	235 (3.39%)	3569 (51.5%)	3804 (54.89%)
Drug Flags			
Alternative Drug Flag = 1	415 (5.99%)	3492 (50.39%)	3905 (56.38%)
Alternative Drug Flag = 0	126 (1.82%)	2897 (41.8%)	3025 (43.62%)
Carbetocin Flag = 1	350 (5.05%)	4198 (60.58%)	4551 (65.63%)
Carbetocin Flag = 0	191 (2.76%)	2191 (31.62%)	2379 (34.37%)
Carbetocin Primary Flag = 1	298 (4.3%)	4020 (58.01%)	4319 (62.31%)
Carbetocin Primary Flag = 0	243 (3.51%)	2369 (34.18%)	2611 (37.69%)
Carboprost Flag = 1	107 (1.54%)	198 (2.86%)	305 (4.4%)
Carboprost Flag = 0	434 (6.26%)	6191 (89.34%)	6625 (95.60%)

Carboprost Primary Flag = 1	10 (0.14%)	12 (0.17%)	22 (0.32%)
Carboprost Primary Flag = 0	531 (7.66%)	6377 (92.02%)	6908 (99.68%)
Ergonovine Flag = 1	121 (1.75%)	311 (4.49%)	432 (6.23%)
Ergonovine Flag = 0	420 (6.06%)	6078 (87.71%)	6498 (93.77%)
Ergonovine Primary Flag = 1	9 (0.13%)	29 (0.42%)	38 (0.55%)
Ergonovine Primary Flag = 0	532 (7.68%)	6360 (91.77%)	6892 (99.45%)
Oxytocin Flag = 1	194 (2.8%)	2039 (29.42%)	2234 (32.22%)
Oxytocin Flag = 0	347 (5.01%)	4350 (62.77%)	4696 (67.78%)
Oxytocin Pre-Cut Flag = 0	404 (5.83%)	4794 (69.18%)	5198 (75.01%)
Oxytocin Pre-Cut Flag = 1	137 (1.98%)	1595 (23.02%)	1732 (24.99%)
Oxytocin Primary Flag = 1	175 (2.53%)	1964 (28.34%)	2138 (30.87%)
Oxytocin Primary Flag = 0	366 (5.28%)	4425 (63.85%)	4792 (69.13%)
Tranexamic Flag = 1	266 (3.84%)	1577 (22.76%)	1844 (26.59%)
Tranexamic Flag = 0	275 (3.97%)	4812 (69.44%)	5086 (73.41%)
Tranexamic Primary Flag = 1	38 (0.55%)	181 (2.61%)	219 (3.16%)
Tranexamic Primary Flag = 0	503 (7.26%)	6208 (89.58%)	6712 (96.84%)
Medical Flags			
Emergent C-Section Flag = 1	293 (4.23%)	3304 (47.68%)	3597 (51.9%)
Emergent C-Section Flag = 0	248 (3.58%)	3085 (44.52%)	3333 (48.1%)
Indicated C-Section Flag = 1	468 (6.75%)	5340 (77.06%)	5808 (83.81%)
Indicated C-Section Flag = 0	73 (1.05%)	1049 (15.14%)	1122 (16.19%)
Repeat C-Section Flag = 1	192 (2.77%)	2430 (35.06%)	2622 (37.84%)
Repeat C-Section Flag = 0	349 (5.04%)	3959 (57.13%)	4308 (62.16%)
Readmission Flag = 1	24 (0.35%)	200 (2.89%)	224 (3.23%)
Readmission Flag = 0	517 (7.46%)	6189 (89.31%)	6706 (96.77%)

Revisit Emergency Department Flag = 1	96 (1.39%)	562 (8.11%)	658 (9.49%)
Revisit Emergency Department Flag = 0	445 (6.42%)	5827 (84.08%)	6272 (90.51%)
Gestational Age			
Gestational Age = 23-29	28 (0.4%)	137 (1.99%)	165 (2.39%)
Gestational Age = 30-35	110 (1.59%)	998 (14.41%)	1107 (15.99%)
Gestational Age = 36-41	300 (4.34%)	3639 (52.72%)	3939 (57.06%)
Gestational Age = 42+	2 (0.03%)	26 (0.38%)	28 (0.4%)
Hemoglobin Drop			
Hemoglobin Drop $\leq$ -30	43 (0.62%)	298 (4.31%)	341 (4.93%)
Hemoglobin Drop -29 to -10	176 (2.55%)	2041 (29.58%)	2217 (32.13%)
Hemoglobin Drop -9 to 0	68 (0.99%)	630 (9.13%)	698 (10.12%)
Hemoglobin Drop $\geq$ 1	5 (0.07%)	30 (0.43%)	35 (0.51%)
Hemoglobin Post-Operation			
Postoperative Hemoglobin $\leq$ 100	103 (1.49%)	1032 (14.93%)	1135 (16.42%)
Postoperative Hemoglobin 101-120	103 (1.49%)	967 (13.99%)	1070 (15.48%)
Postoperative Hemoglobin 121-140	50 (0.72%)	538 (7.78%)	588 (8.5%)
Postoperative Hemoglobin $\geq$ 141	5 (0.07%)	27 (0.39%)	32 (0.46%)
Hemoglobin Pre-Operation			
Preoperative Hemoglobin $\leq$ 100	7 (0.1%)	37 (0.53%)	44 (0.63%)
Preoperative Hemoglobin 101-120	19 (0.27%)	198 (2.86%)	217 (3.13%)
Preoperative Hemoglobin 121-140	71 (1.03%)	927 (13.55%)	998 (14.58%)
Preoperative Hemoglobin $\geq$ 141	4 (0.06%)	37 (0.54%)	41 (0.59%)
Other Conditions			
Gestational Diabetes Diagnosis = 1	50 (0.72%)	527 (7.6%)	577 (8.33%)
Gestational Diabetes Diagnosis = 0	491 (7.09%)	5862 (84.59%)	6353 (91.67%)
Hypertension Diagnosis = 1	208 (3.0%)	2076 (29.96%)	2284 (32.96%)

Hypertension Diagnosis = 0	333 (4.81%)	4313 (62.24%)	4646 (67.04%)
Thyroid Disorder Diagnosis = 1	1 (0.01%)	39 (0.56%)	40 (0.58%)
Thyroid Disorder Diagnosis = 0	540 (7.79%)	6350 (91.63%)	6890 (99.42%)
Urinary Tract Infection Diagnosis = 1	4 (0.06%)	25 (0.36%)	29 (0.42%)
Urinary Tract Infection Diagnosis = 0	537 (7.75%)	6364 (91.83%)	6901 (99.58%)
Gestational Diagnosis Medication = 1	360 (5.19%)	3169 (45.73%)	3529 (50.92%)
Gestational Diagnosis Medication = 0	181 (2.61%)	3220 (46.46%)	3401 (49.08%)
Hypertension Medication = 1	164 (2.37%)	1351 (19.49%)	1515 (21.86%)
Hypertension Medication = 0	377 (5.44%)	5038 (72.7%)	5415 (78.14%)
Asthma Medication = 1	354 (5.11%)	3130 (45.17%)	3484 (50.27%)
Asthma Medication = 0	187 (2.7%)	3259 (47.03%)	3446 (49.73%)
Thyroid Disorder Medications = 1	161 (2.32%)	1305 (18.83%)	1466 (21.15%)
Thyroid Disorder Medication = 0	380 (5.48%)	5084 (73.36%)	5464 (78.85%)
Depression/Anxiety Disorder Medication = 1	334 (4.82%)	2989 (43.13%)	3323 (47.95%)
Depression/Anxiety Disorder Medication = 0	207 (2.99%)	3400 (49.06%)	3607 (52.05%)
Epilepsy/Seizure Disorder Medication = 1	352 (5.08%)	3104 (44.79%)	3456 (49.87%)
Epilepsy/Seizure Disorder Medication = 0	189 (2.73%)	3285 (47.4%)	3474 (50.13%)
Iron Deficiency/Anemia Medication = 1	335 (4.83%)	2844 (41.04%)	3179 (45.87%)
Iron Deficiency/Anemia Medication = 0	206 (2.97%)	3545 (51.15%)	3751 (54.13%)
Urinary Tract Infection Medication = 1	531 (7.66%)	6173 (89.08%)	6704 (96.74%)
Urinary Tract Infection Medication = 0	10 (0.14%)	216 (3.12%)	226 (3.26%)
Allergy Medication = 1	277 (4.0%)	2627 (37.91%)	2904 (41.9%)
Allergy Medication = 0	264 (3.81%)	3762 (54.29%)	4026 (58.1%)

Table 1 represents the distribution of various patient characteristics and medical conditions, stratified by whether the patients had a positive or negative experience with PPH.

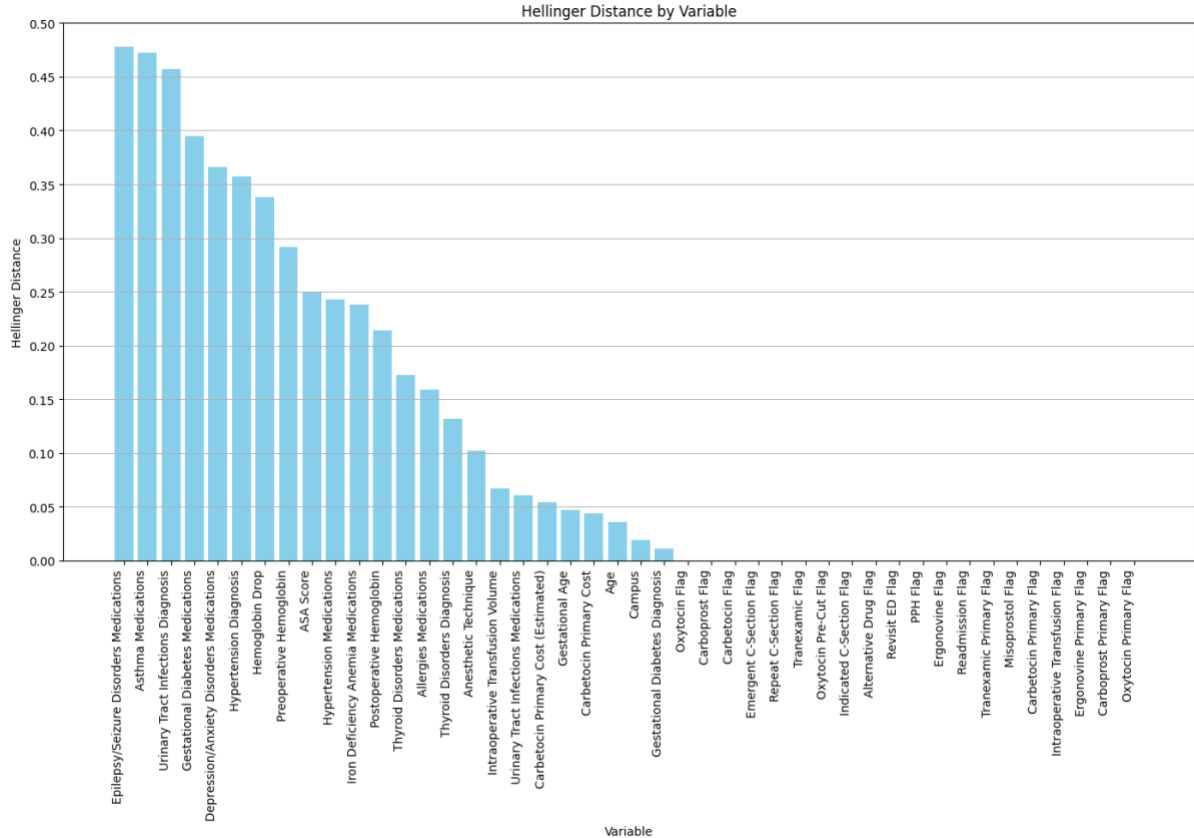


Figure 3: Hellinger Distance Across All the Variables

Figure 3 illustrates the Hellinger Distance for each variable, comparing the distribution differences between synthetic and real datasets. The y-axis represents the Hellinger Distance, with higher values indicating greater distributional differences. Variables are listed along the x-axis, ordered from highest to lowest Hellinger Distance. Notably, 'Epilepsy/Seizure Disorders Medications' and 'Asthma Medications' exhibit the highest distances, reflecting significant discrepancies between the synthetic and real datasets. Conversely, variables such as 'Oxytocin Primary Flag' and 'Ergonovine Primary Flag' show minimal differences, suggesting that the synthetic data closely mirrors the real data for these features.

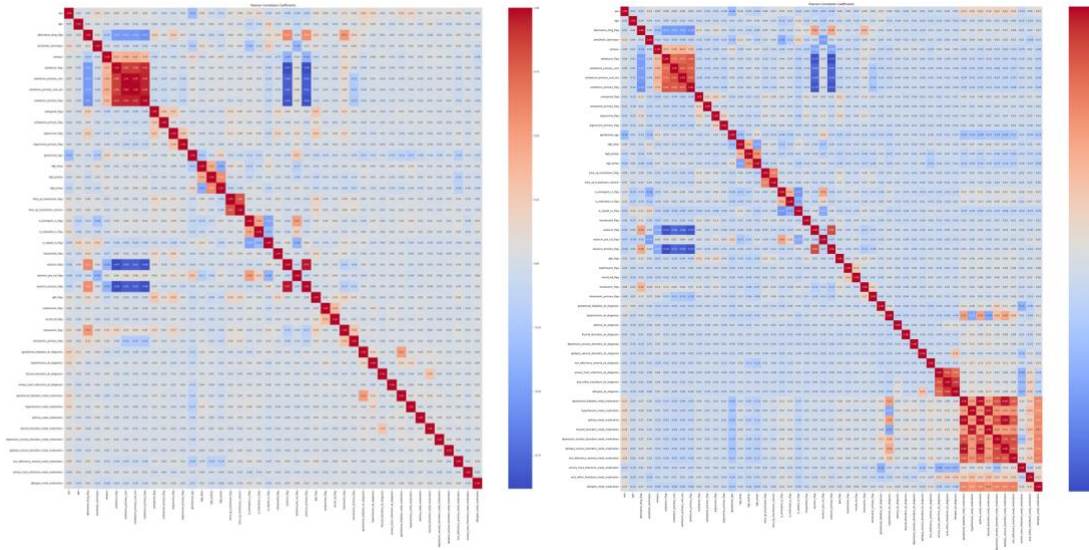


Figure 4: Real vs Synthetic Data Pearson Correlation Heatmaps

In Figure 4 the left heatmap represents the correlations between variables in the synthetic dataset, while the right heatmap shows the correlations in the real dataset. Notably, the synthetic data exhibits higher correlations in the diagnosis and medication variables compared to the real data. There also seems to be a generally slightly higher correlation between the variables in the real dataset as opposed to the synthetic dataset.

Table 2: Top Variables with the Highest Correlation with PPH Positive

Features	Real	Synthetic
Carboprost Flag	0.27	0.22
Ergonovine Flag	0.25	0.19
Tranexamic Acid Flag	0.19	0.15
Intraoperative Transfusion Flag	0.18	0.15
Alternative Drug Flag	0.15	0.12
Misoprostol Flag	0.15	0.11
Intraoperative Transfusion Volume	0.15	0.11
Carboprost Primary Flag	0.10	-
Revisit Emergency Department Flag	0.10	-
Iron Deficiency Anemia Medication Flag	0.09	0.09
Gestational Diabetes Medication Flag	-	0.09
Urinary Tract Infections Diagnosis	-	0.09

Table 2 presents the top features correlating with PPH in both the real and synthetic datasets. In the real dataset, the most influential features include "Carboprost Flag" (0.27), "Ergonovine Flag" (0.25), and "Tranexamic Flag" (0.19). These features indicate the administration of certain medications and

procedures related to PPH management. Similarly, in the synthetic dataset, the top features are "Carboprost Flag" (0.22), "Ergonovine Flag" (0.19), and "Intraoperative Transfusion Flag" (0.15). However, some discrepancies are noted. For instance, "Carboprost Primary Flag" and "Revisit Emergency Department Flag," which appear in the real dataset, do not appear in the synthetic dataset's top features. Conversely, "Gestational Diabetes Medications" and "Urinary Tract Infections Diagnosis" are significant in the synthetic dataset but not in the real dataset.

4.2 Model Outputs

Table 3: Variables Kept in the Model Based on Feature Selection done on the Synthetic Dataset and Literature

Actual Name	Data Type
ASA Score	Numeric
Age	Numeric
Alternative Drug Flag	Binary (1 or 0)
Campus	Binary (1 or 0)
Carbetocin Flag	Binary (1 or 0)
Carbetocin Primary Flag	Binary (1 or 0)
Carboprost Flag	Binary (1 or 0)
Carboprost Primary Flag	Binary (1 or 0)
Ergonovine Flag	Binary (1 or 0)
Ergonovine Primary Flag	Binary (1 or 0)
Gestational Age	Numeric
Hemoglobin Drop	Numeric
Postoperative Hemoglobin	Numeric
Preoperative Hemoglobin	Numeric
Intraoperative Transfusion Flag	Binary (1 or 0)
Intraoperative Transfusion Volume	Numeric
Emergent C-Section Flag	Binary (1 or 0)
Indicated C-Section Flag	Binary (1 or 0)
Repeat C-Section Flag	Binary (1 or 0)
Oxytocin Flag	Binary (1 or 0)
Oxytocin Pre-Cut Flag	Binary (1 or 0)
Oxytocin Primary Flag	Binary (1 or 0)
Readmission Flag	Binary (1 or 0)
Revisit Emergency Department Flag	Binary (1 or 0)
Tranexamic Flag	Binary (1 or 0)
Tranexamic Primary Flag	Binary (1 or 0)
Gestational Diabetes Diagnosis	Binary (1 or 0)
Hypertension Diagnosis	Binary (1 or 0)
Thyroid Disorders Diagnosis	Binary (1 or 0)
Urinary Tract Infections Diagnosis	Binary (1 or 0)

Gestational Diabetes Medications	Binary (1 or 0)
Hypertension Medications	Binary (1 or 0)
Asthma Medications	Binary (1 or 0)
Thyroid Disorders Medications	Binary (1 or 0)
Depression/Anxiety Disorders Medications	Binary (1 or 0)
Epilepsy/Seizure Disorders Medications	Binary (1 or 0)
Iron Deficiency Anemia Medications	Binary (1 or 0)
Urinary Tract Infections Medications	Binary (1 or 0)
Allergies Medications	Binary (1 or 0)

Table 3 lists the variables retained in the model after feature selection. Each variable is accompanied by its actual name and data type, based on the analysis of the synthetic dataset and literature review.

Table 4: 5-Fold Cross Validated AUC Scores for Real and Synthetic Dataset

Model	Real Mean AUC Score	Synthetic Mean AUC Score
Logistic Regression	0.85	0.81 (+/- 0.014)
Decision Tree	0.93	0.92 (+/- 0.0053)
Random Forest	0.99	0.98 (+/- 0.013)
SVM	0.97	0.97 (+/- 0.010)

Table 4 shows the mean AUC scores from 5-fold cross-validation for different machine learning models applied to both real and synthetic datasets. The Logistic Regression model achieved a mean AUC score of 0.85 for real data and 0.81 (+/- 0.014) for synthetic data. The Decision Tree model had mean scores of 0.93 for real data and 0.92 (+/- 0.0053) for synthetic data. The Random Forest model performed exceptionally well, with mean AUC scores of 0.99 for real data and 0.98 (+/- 0.013) for synthetic data. The SVM model also demonstrated high performance, with a mean AUC score of 0.97 for both datasets, with synthetic data having a standard deviation of +/- 0.010. Generally, there should be a standard deviation for the cross-validation, but we were only able to get it for the synthetic data, not the real data.

Table 5: Performance of Models Trained on Synthetic Dataset

Model	Precision	Recall	F1-Score	Accuracy	AUC-ROC
Logistic Regression	0.79	0.73	0.76	0.77	0.81
Decision Tree	0.91	0.92	0.92	0.92	0.93
Random Forest	0.98	0.92	0.95	0.95	0.98
SVM	0.94	0.85	0.89	0.90	0.97

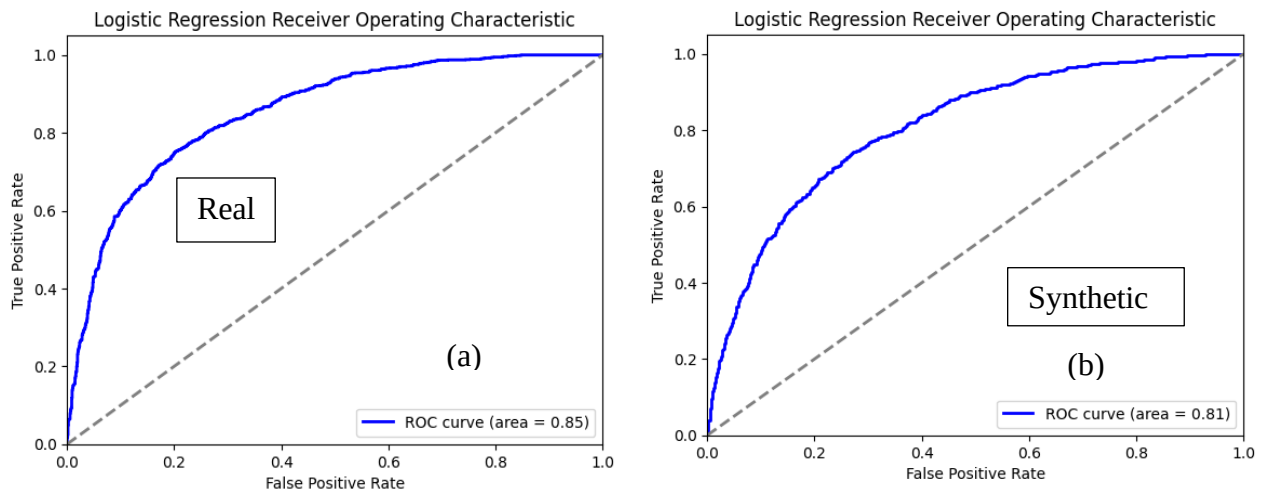
Table 5 shows the performance metrics of various machine learning models in predicting PPH using the synthetic dataset. Metrics reported include Precision, Recall, F1-Score, Accuracy, and AUC-ROC

for Logistic Regression, Decision Tree, Random Forest, and SVM with the highest performance noted in the Random Forest model.

Table 6: Performance of Models Trained on Real Data

Model	Precision	Recall	F1-Score	Accuracy	AUC-ROC
Logistic Regression	0.73	0.70	0.72	0.73	0.85
Decision Tree	0.91	0.94	0.92	0.93	0.92
Random Forest	0.99	0.94	0.96	0.96	0.98
SVM	0.96	0.86	0.91	0.93	0.97

Table 6: Performance of Models Trained on Real Data shows the performance metrics of various machine learning models in predicting PPH using the real dataset. Metrics reported include Precision, Recall, F1-Score, Accuracy, and AUC-ROC for Logistic Regression, Decision Tree, Random Forest, and SVM highlighting that Random Forest exhibits the highest performance.



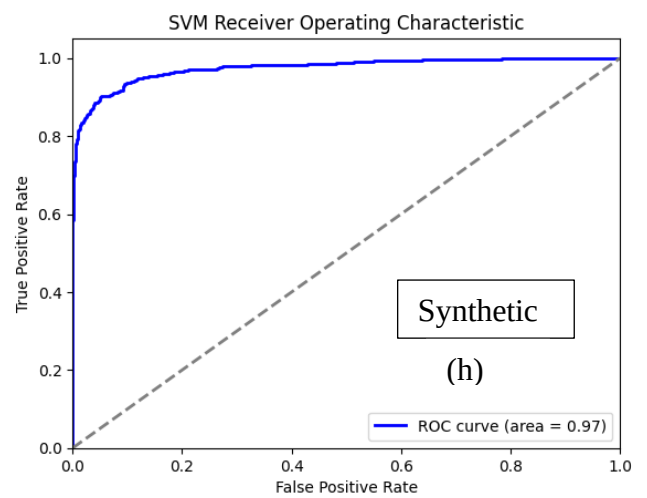
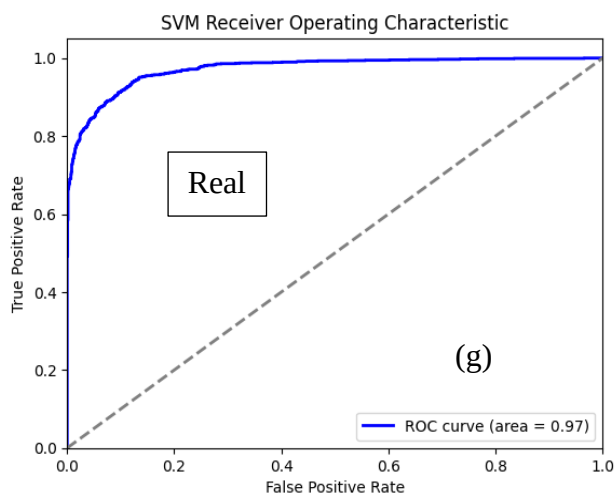
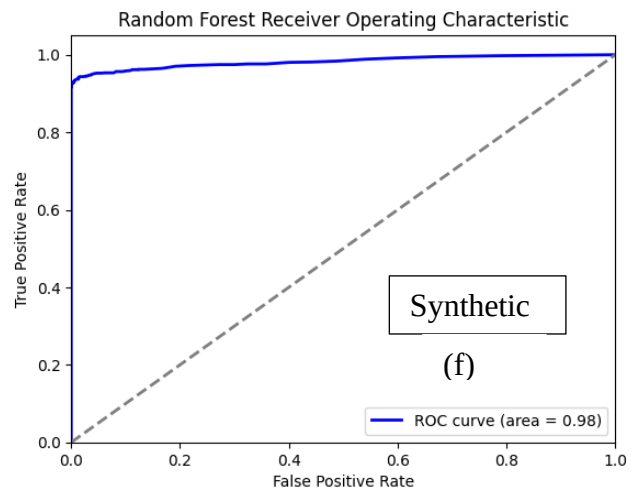
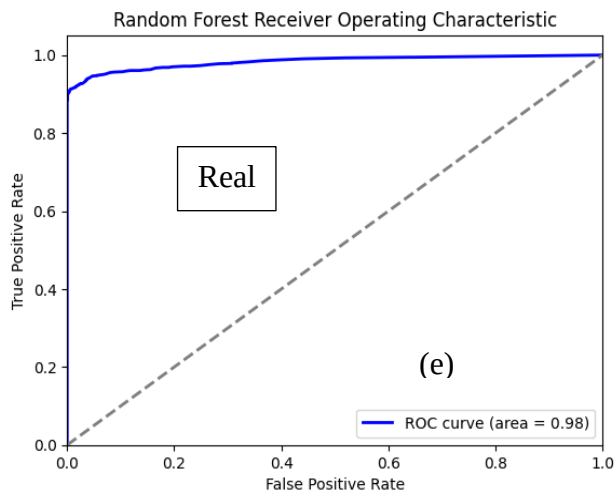
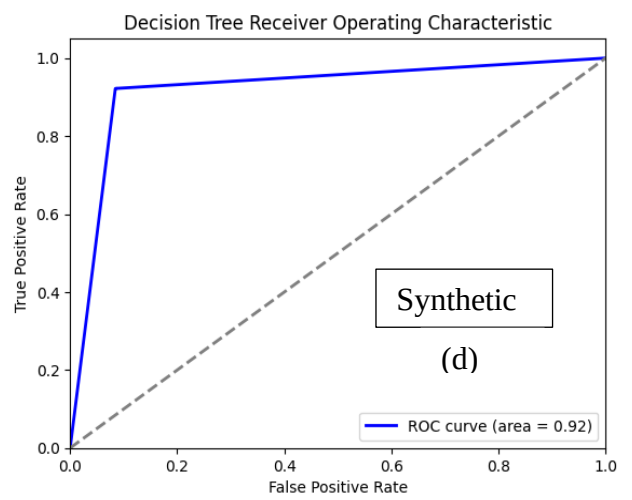
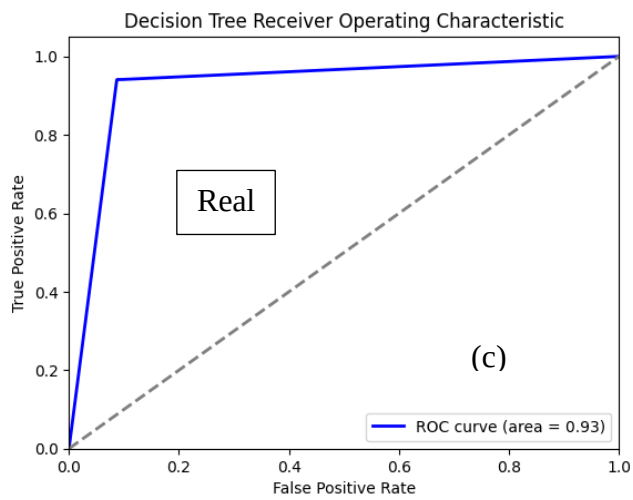


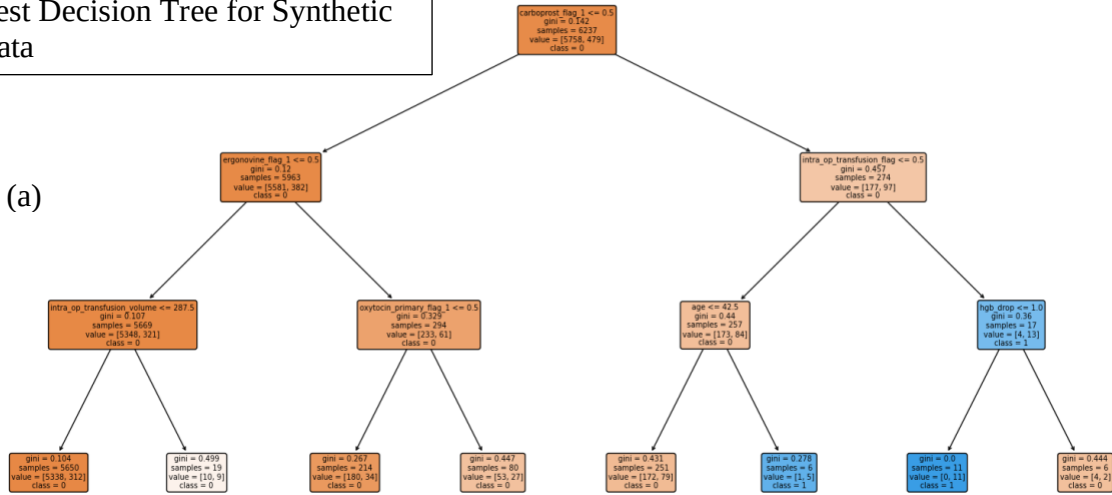
Figure 5: Real vs Synthetic Data AUC-ROC Curves

Figure 5: Real vs Synthetic Data AUC-ROC Curves displays the ROC (Receiver Operating Characteristic) curves for various models, comparing their performance on real data (left column) versus synthetic data (right column). The models and their corresponding AUC (Area Under the Curve) scores are as follows: (a) Logistic Regression on real data with an AUC score of 0.85, (b) Logistic Regression on synthetic data with an AUC score of 0.81, (c) Decision Tree on real data with an AUC score of 0.92, (d) Decision Tree on synthetic data with an AUC score of 0.93, (e) Random Forest on real data with an AUC score of 0.98, (f) Random Forest on synthetic data with an AUC score of 0.98, (g) SVM on real data with an AUC score of 0.97, and (h) SVM on synthetic data with an AUC score of 0.97. The ROC curves provide a visual representation of the models' performance, with higher AUC scores indicating better accuracy.

4.3 Decision Tree Outputs

Best Decision Tree for Synthetic Data

(a)



Best Decision Tree for Real Data

(b)

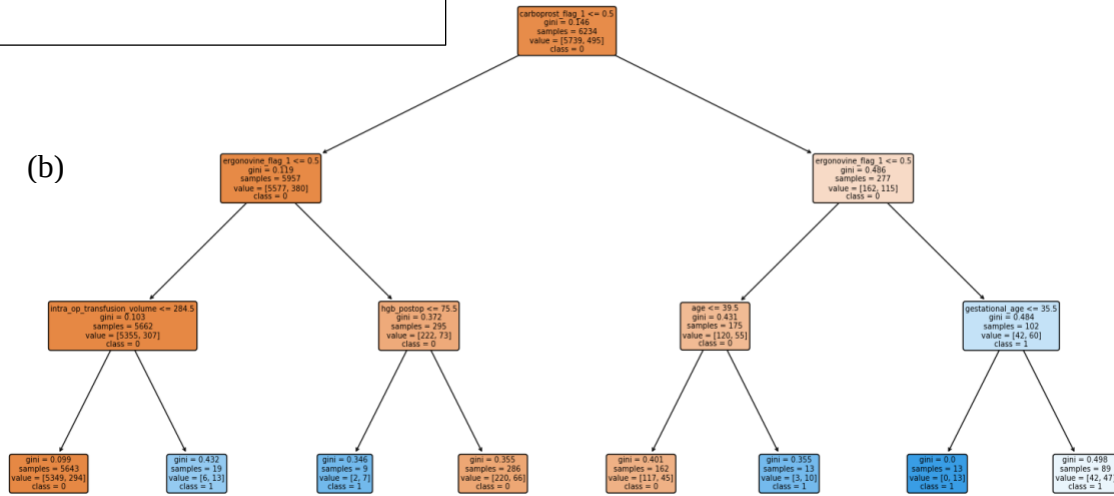


Figure 6: Pruned Decision Tree for High-Risk Patient Identification

The figure displays the best Decision Tree models for both synthetic (a) and real (b) data in predicting PPH outcomes. The top tree represents the model trained on real data, while the bottom tree represents the model trained on synthetic data. Both trees begin with the same root node, "Carboprost Flag," and share a similar structure in subsequent splits, indicating key decision points such as "Ergonovine Flag," "Intraoperative Transfusion Volume," and "Age." Despite slight variations in the

paths and leaf nodes, the overall structure and key decision variables remain consistent between the two models.

1. Pathway 1:

Root Node (a): Carboprost Flag ≤ 0.5 (No)

Left Subtree (b): Ergonovine Flag ≤ 0.5 (No)

Left Subtree of b (d): Intraoperative Transfusion Volume Flag ≤ 267.5

Leaf Node: Class 0

Gini: 0.104, Samples: 5650, Value: [5338, 312]

2. Pathway 2:

Root Node (a): Carboprost Flag ≤ 0.5 (No)

Left Subtree (b): Ergonovine Flag ≤ 0.5 (No)

Left Subtree of b (d): Intraoperative Transfusion Volume Flag > 267.5

Leaf Node: Class 0

Gini: 0.499, Samples: 19, Value: [10, 9]

3. Pathway 3:

Root Node (a): Carboprost Flag ≤ 0.5 (No)

Left Subtree (b): Ergonovine Flag > 0.5 (Yes)

Right Subtree of b (e): Oxytocin Primary Flag ≤ 0.5 (No)

Leaf Node: Class 0

Gini: 0.257, Samples: 214, Value: [180, 34]

4. Pathway 4:

Root Node (a): Carboprost Flag ≤ 0.5 (No)

Left Subtree (b): Ergonovine Flag > 0.5 (Yes)

Right Subtree of b (e): Oxytocin Primary Flag > 0.5 (Yes)

Leaf Node: Class 0

Gini: 0.447, Samples: 80, Value: [53, 27]

5. Pathway 5:

Root Node (a): Carboprost Flag > 0.5 (Yes)

Right Subtree (c): Intraoperative Transfusion Flag ≤ 0.5 (No)

Left Subtree of c (f): Age ≤ 42.4

Leaf Node: Class 0

Gini: 0.431, Samples: 251, Value: [172, 79]

6. Pathway 6:

Root Node (a): Carboprost Flag > 0.5 (No)
 Right Subtree (c): Intraoperative Transfusion Flag $\leq$ 0.5 (Yes)
 Left Subtree of c (f): Age > 42.5
 Leaf Node: Class 1
 Gini: 0.278, Samples: 6, Value: [1, 5]

7. Pathway 7:

Root Node (a): Carboprost > 0.5 (No)
 Right Subtree (c): Intraoperative Transfusion Flag > 0.5 (No)
 Left Subtree of c (f): Hemoglobin Drop $\leq$ 1.0
 Leaf Node: Class 1
 Gini: 0.0, Samples: 11, Value: [0, 11]

8. Pathway 8:

Root Node (a): Carboprost Flag > 0.5 (No)
 Right Subtree (c): Intraoperative Transfusion Flag > 0.5 (No)
 Left Subtree of c (f): Hemoglobin Drop > 1.0
 Leaf Node: Class 0
 Gini: 0.44, Samples: 5, Value: [4, 2]

Figure 7: Pathways for Positive and Negative PPH (Synthetic Data)

1. Pathway 1:

Root Node (a): Carboprost Flag $\leq$ 0.5 (No)
 Left Subtree (b): Ergonovine Flag $\leq$ 0.5 (No)
 Left Subtree of b (d): Intraoperative Transfusion Volume $\leq$ 267.5
 Leaf Node: Class 0
 ○ Gini: 0.099, Samples: 5643, Value: [5349, 294]

2. Pathway 2:

Root Node (a): Carboprost Flag $\leq$ 0.5 (No)
 Left Subtree (b): Ergonovine Flag $\leq$ 0.5 (No)
 Left Subtree of b (d): Intraoperative Transfusion Volume > 284.5
 Leaf Node: Class 1
 ○ Gini: 0.432, Samples: 19, Value: [6, 13]

3. Pathway 3:

Root Node (a): Carboprost Flag $\leq$ 0.5 (No)

Left Subtree (b): Ergonovine Flag > 0.5 (Yes)

Right Subtree of b (e): Hemoglobin Post Operation ≤ 75.5

Leaf Node: Class 1

- Gini: 0.346, Samples: 9, Value: [2, 7]

4. Pathway 4:

Root Node (a): Carboprost Flag ≤ 0.5 (No)

Left Subtree (b): Ergonovine Flag > 0.5 (Yes)

Right Subtree of b (e): Hemoglobin Post Operation > 75.5

Leaf Node: Class 0

- Gini: 0.355, Samples: 286, Value: [220, 66]

5. Pathway 5:

Root Node (a): Carboprost Flag > 0.5 (Yes)

Right Subtree (c): Ergonovine Flag ≤ 0.5 (No)

Left Subtree of c (f): Age ≤ 39.5

Leaf Node: Class 0

- Gini: 0.401, Samples: 162, Value: [117, 45]

6. Pathway 6:

Root Node (a): Carboprost Flag > 0.5 (Yes)

Right Subtree (c): Ergonovine Flag ≤ 0.5 (No)

Left Subtree of c (f): Age > 39.5

Leaf Node: Class 1

- Gini: 0.355, Samples: 13, Value: [3, 10]

7. Pathway 7:

Root Node (a): Carboprost Flag > 0.5 (No)

Right Subtree (c): Ergonovine Flag > 0.5 (No)

Right Subtree of c (g): Gestational Age ≤ 35.5

Leaf Node: Class 1

- Gini: 0.0, Samples: 13, Value: [0, 13]

8. Pathway 8:

Root Node (a): Carboprost Flag > 0.5 (No)

Right Subtree (c): Ergonovine Flag > 0.5 (No)

Right Subtree of c (g): Gestational Age > 35.5

Leaf Node: Class 1

- Gini: 0.498, Samples: 89, Value: [42, 47]

Figure 8: Pathways for Positive and Negative PPH (Real Data)

Chapter 5

Discussion

The primary objective of this thesis is to evaluate the utility of synthetic data as a viable proxy for real data in predictive modeling using clinical data. Specifically, this research aims to assess whether synthetic data can be used to predict PPH outcomes as accurately as the model trained on real dataset. By doing so, the thesis addresses two main purposes: first, to determine the overall applicability of synthetic data in clinical predictive modeling; and second, to develop and validate models for predicting PPH, using both synthetic and real datasets, to understand the effectiveness and limitations of synthetic data in this context. By examining various aspects such as the EDA, distribution of data, Hellinger Distance, Pearson Correlation Coefficient, and Statistical and ML models we aim to understand the strengths and limitations of synthetic data. This discussion will highlight how these factors influence the performance of predictive models and their implications for the broader application of synthetic data in healthcare research. The insights gained from this analysis are critical for assessing the feasibility of using synthetic data in clinical settings, ensuring that predictive models remain robust, accurate, and generalizable.

5.1 Cohort Characteristics, EDA, and Distribution of Data between Real and Synthetic Datasets

Table 1 provides an overview of the distribution of various patient characteristics and medical conditions, stratified by their PPH Diagnosis (Positive or Negative).

5.1.1 American Society of Anesthesiologists (ASA) score

The American Society of Anesthesiologists (ASA) score is a system used to assess and communicate a patient's pre-anesthesia medical comorbidities [44]. It ranges from 1 (healthy) to 5 (moribund). Notably, an ASA score of 2 is automatically applied to individuals who are pregnant, reflecting their physiological changes and potential for complications [44]. Higher ASA scores (3 and above) were associated with a higher incidence of PPH, suggesting that patients with more severe underlying health conditions are at increased risk. Specifically, ASA 3 patients constituted the majority of PPH cases (51.91%), highlighting the importance of preoperative health status in predicting PPH risk.

5.1.2 Age

Patients aged 17-25 account for 7.96% of the total cohort, with 0.68% in the PPH Positive group and 7.28% in the PPH Negative group. Younger women generally have a lower risk of PPH compared to older age groups, likely due to fewer comorbid conditions and generally better overall health, which aligns with the lower prevalence of PPH in this age group [45]. Patients aged 26-35 represent the largest age group, accounting for 53.22% of the cohort, with 3.76% in the PPH Positive group and 49.44% in the PPH Negative group. This age group typically represents a peak in reproductive activity, leading to a higher number of pregnancies and deliveries [46]. While generally healthy, women in this age range may still encounter obstetric complications that can increase the risk of PPH, consistent with their higher prevalence of PPH. Patients aged 36-45 make up 37.29% of the cohort, with 2.41% in the PPH Positive group and 34.87% in the PPH Negative group. Advanced maternal age (35 and older) is associated with higher risks of pregnancy complications, including PPH, due to factors such as decreased uterine tone, higher incidence of comorbidities, and increased likelihood of interventions such as C-sections [45]. The relatively high prevalence of PPH in this group reflects these risk factors. Patients aged 46+ are the smallest age group, representing 1.52% of the cohort, with 0.12% in the PPH Positive group and 1.40% in the PPH Negative group. Pregnancies in women aged 46 and older, while rare, carry significant risks and are often associated with higher rates of complications, including PPH. However, the overall number of pregnancies in this age group is low, reflecting their small representation in the cohort.

5.1.3 Campus

The distribution of PPH cases across different campuses shows variability, which is most likely because the general campus (campus = 1) tends to handle more severe cases of pregnancies, which might contribute to the higher incidence of PPH observed in this setting.

The data shows that most PPH cases occurred at full-term gestation (36-41 weeks). This aligns with the general population distribution but highlights that even at full-term, significant risks for PPH remain. Both preoperative and postoperative hemoglobin levels were lower in patients experiencing PPH, underscoring the importance of monitoring and managing hemoglobin levels to prevent severe outcomes.

5.1.4 Gestational Age

Patients with a gestational age of 23-29 weeks account for 2.39% of the total cohort, with 0.4% in the PPH Positive group and 1.99% in the PPH Negative group. Extremely preterm births (23-29 weeks) are associated with higher maternal and neonatal complications, including potentially PPH. These

pregnancies often involve medical interventions and complications such as placental issues or preeclampsia, which can increase the risk of PPH [47]. However, the overall incidence of deliveries in this age group is low due to the higher likelihood of miscarriages or medical interventions before this gestational period. Patients with a gestational age of 30-35 weeks represent 15.99% of the cohort, with 1.59% in the PPH Positive group and 14.41% in the PPH Negative group. Late preterm births (30-35 weeks) still carry significant risks for both the mother and the baby. The risk of PPH is elevated in this group due to potential complications like uterine atony, preeclampsia, and the need for medical interventions such as C-sections [48]. The higher percentage in the PPH Negative group is consistent with the greater number of births in this gestational age range, but the elevated risk of PPH remains a concern. Patients with a gestational age of 36-41 weeks make up 57.06% of the cohort, with 4.34% in the PPH Positive group and 52.72% in the PPH Negative group. Full-term births (36-41 weeks) are generally considered to be the safest period for delivery, with the lowest risk of complications. However, PPH can still occur due to factors like prolonged labor, uterine atony, or delivery-related trauma [48]. The majority of births fall within this range, and the distribution between PPH Positive and PPH Negative groups reflects the overall lower but still present risk of PPH at full term. The gestational age distribution of PPH risk in the cohort aligns well with existing literature. Extremely preterm births have a high risk of complications, including PPH, due to the critical nature of these pregnancies. Late preterm births also carry significant risks but are more common, which is reflected in the data. Full-term births are associated with the lowest risk of complications, including PPH, but due to the higher volume of births in this category, the absolute numbers remain significant.

5.1.5 Medical Flags

Emergent C-sections are typically performed when there is an immediate threat to the life of the mother or baby, such as in cases of placental abruption, uterine rupture, or severe preeclampsia. These conditions are also known risk factors for PPH [49]. Patients with an Emergent C-Section Flag = 1 account for 51.9% of the total cohort, with 4.23% in the PPH Positive group and 47.68% in the PPH Negative group. Patients without an Emergent C-Section Flag = 0 constitute 48.1% of the cohort, with 3.58% in the PPH Positive group and 44.52% in the PPH Negative group. The data reveals that there isn't a significant difference in the total number of people who had an emergent C-section compared to those who had a non-emergent or planned C-section, with the distribution being almost 50-50. Interestingly, a slightly higher percentage of people who had emergent C-sections were in the PPH positive group. This is surprising because one might expect that emergent C-sections, often

performed in life-threatening situations, would be associated with a much higher incidence of PPH compared to planned C-sections. However, the data suggests otherwise, indicating that the PPH rates are relatively similar between the two groups. This could be due to medical teams being highly vigilant and well-prepared to manage potential complications, such as PPH, during emergent C-sections. Another possibility is that patients undergoing planned C-sections might have been identified as high-risk beforehand, prompting healthcare providers to take extra precautions, which could include proactive measures to prevent PPH. Regardless, the unexpected similarity in PPH rates between emergent and non-emergent C-sections warrants further investigation to better understand the factors at play.

Indicated C-sections are planned due to known risk factors like previous C-section, breech presentation, or maternal health issues. While planned, these C-sections can still be associated with higher risk of complications, including PPH, especially if the underlying conditions are severe. Patients with an Indicated C-Section Flag = 1 represent 83.81% of the cohort, with 6.75% in the PPH group and 77.06% in the non-PPH group. Patients without an Indicated C-Section Flag = 0 make up 16.19% of the cohort, with 1.05% in the PPH group and 15.14% in the non-PPH group. The higher percentage of indicated C-sections in the PPH group suggests that even planned, medically indicated C-sections are associated with a higher risk of developing PPH. This might reflect underlying health conditions or complications that necessitate a planned C-section and concurrently increase the risk of PPH. It aligns with the understanding that patients with known risk factors requiring planned C-sections may still face significant complications.

Repeat C-sections are usually planned and occur in a controlled environment, which can help mitigate the risk of PPH. However, the surgical history of multiple C-sections can increase the risk of complications such as placenta accreta, which can lead to PPH [50]. Patients with a Repeat C-Section Flag = 1 account for 37.84% of the cohort, with 2.77% in the PPH group and 35.06% in the non-PPH group. Patients without a Repeat C-Section Flag = 0 constitute 62.16% of the cohort, with 5.04% in the PPH group and 57.13% in the non-PPH group. The lower percentage of repeat C-sections in the PPH group indicates that PPH is more likely to occur in first-time or non-repeat C-section deliveries, but this does not match with literature. A explanation for this could be because repeat C-sections are managed more carefully with a preoperative plan.

PPH can lead to complications such as anemia, infection, and the need for further surgical intervention, which increases the likelihood of readmission. Patients with a Readmission Flag = 1 represent a small portion of the cohort, accounting for 3.23%, with 0.35% in the PPH group and 2.89% in the non-PPH group. Patients without a Readmission Flag = 0 make up 96.77% of the cohort,

with 7.46% in the PPH group and 89.31% in the non-PPH group. The slightly higher readmission rate in the PPH group is consistent with the literature, as patients who experience PPH are more likely to suffer from complications requiring additional medical attention, thus necessitating readmission. This highlights the need for close monitoring and follow-up care for patients who have experienced PPH. Patients with a Readmission Flag = 1 represent a small portion of the cohort, accounting for 3.23%, with 0.35% in the PPH Positive group and 2.89% in the PPH Negative group. Patients without a Readmission Flag = 0 make up 96.77% of the cohort, with 7.46% in the PPH Positive group and 89.31% in the PPH Negative group. The low readmission rate in both groups is a positive indicator of the overall quality of post-operative care and patient management. However, the slightly higher readmission rate in the PPH Positive group suggests that patients with PPH might experience more complications requiring readmission. This is consistent with the understanding that PPH can lead to complications such as anemia, infection, and the need for further surgical intervention, thus increasing the likelihood of readmission.

5.1.6 Hemoglobin

Patients experiencing a significant drop in hemoglobin (≤ -30 g/L) account for 4.93% of the total cohort, with 0.62% in the PPH Positive group and 4.31% in the PPH Negative group. A large drop in hemoglobin levels is a critical indicator of severe blood loss, which is often associated with PPH [51]. Moderate drops in hemoglobin (-29 to -10 g/L) represent 32.13% of the cohort, with 2.55% in the PPH Positive group and 29.58% in the PPH Negative group, indicating significant bleeding albeit less severe. Minor drops in hemoglobin (-9 to 0 g/L) account for 10.12% of the cohort, with 0.99% in the PPH Positive group and 9.13% in the PPH Negative group, reflecting minor blood loss. Patients with no drop or an increase in hemoglobin (≥ 1 g/L) account for 0.51% of the cohort, with 0.07% in the PPH Positive group and 0.43% in the PPH Negative group, which is not generally associated with PPH.

Postoperative hemoglobin levels also provide valuable insights. Patients with low postoperative hemoglobin levels (≤ 100 g/L) represent 16.42% of the cohort, with 1.49% in the PPH Positive group and 14.93% in the PPH Negative group, indicating significant blood loss during surgery. Moderate postoperative hemoglobin levels (101-120 g/L) make up 15.48% of the cohort, with 1.49% in the PPH Positive group and 13.99% in the PPH Negative group. Higher postoperative hemoglobin levels (121-140 g/L) account for 8.5% of the cohort, with 0.72% in the PPH Positive group and 7.78% in the PPH Negative group, suggesting minimal blood loss. Very high postoperative hemoglobin levels ($\geq$

141 g/L) represent 0.46% of the cohort, with 0.07% in the PPH Positive group and 0.39% in the PPH Negative group, which are generally not indicative of PPH.

Preoperative hemoglobin levels help understand patient vulnerability to PPH. Patients with low preoperative hemoglobin levels (≤ 100 g/L) account for 0.63% of the cohort, with 0.1% in the PPH Positive group and 0.53% in the PPH Negative group, indicating a predisposition to severe outcomes if PPH occurs. Patients with moderately low preoperative hemoglobin levels (101-120 g/L) constitute 3.13% of the cohort, with 0.27% in the PPH Positive group and 2.86% in the PPH Negative group. Normal preoperative hemoglobin levels (121-140 g/L) make up 14.58% of the cohort, with 1.03% in the PPH Positive group and 13.55% in the PPH Negative group, indicating a lower risk of severe outcomes. High preoperative hemoglobin levels (≥ 141 g/L) account for 0.59% of the cohort, with 0.06% in the PPH Positive group and 0.54% in the PPH Negative group, which is not typically associated with an increased risk of PPH.

5.1.7 Other Conditions

Patients with gestational diabetes represent 8.33% of the cohort, with 0.72% in the PPH Positive group. Literature suggests that effective management of blood sugar levels in gestational diabetes can mitigate some of the PPH risks [52]. Hypertension is observed in 32.96% of the cohort, with 3.0% in the PPH Positive group. Hypertension is a known risk factor for PPH due to its association with preeclampsia and other hypertensive disorders of pregnancy, but well-managed cases show reduced severe outcomes [36].

Regarding medication use, 50.92% of the cohort are on gestational diabetes medication, with 5.19% in the PPH Positive group. Proper management of gestational diabetes through medication reduces complications. Hypertension medication is used by 21.86% of the cohort, with 2.37% in the PPH Positive group, reflecting effective management of hypertension. Asthma medication is used by 50.27% of the cohort, with 5.11% in the PPH Positive group, indicating good control over the condition. Thyroid disorder medication is used by 21.15% of the cohort, with 2.32% in the PPH Positive group, and effective management appears to reduce PPH risk. Depression/anxiety medication is used by 47.95% of the cohort, with 4.82% in the PPH Positive group, and epilepsy/seizure disorder medication is used by 49.87% of the cohort, with 5.08% in the PPH Positive group. Proper management of these conditions is crucial for preventing complications. Iron deficiency/anemia medication is used by 45.87% of the cohort, with 4.83% in the PPH Positive group, highlighting the importance of managing anemia to reduce PPH risk. Urinary tract infection (UTI) medication is used by 96.74% of the cohort, with 7.66% in the PPH Positive group, reflecting the routine treatment of

this condition during pregnancy. Allergy medication is used by 41.9% of the cohort, with 4.0% in the PPH Positive group, and allergy management does not significantly impact PPH outcomes.

Less prevalent conditions include thyroid disorders, diagnosed in 0.58% of the cohort, with 0.01% in the PPH Positive group, and UTIs, diagnosed in 0.42% of the cohort, with 0.06% in the PPH Positive group. These conditions show minimal impact on PPH incidence within this cohort.

5.1.8 Drugs Flag

Comparing the primary flags for Carbetocin, Oxytocin, Carboprost, Ergonovine, and Tranexamic, it becomes clear that Carboprost, Ergonovine and Tranexamic were not commonly used as primary medications. Specifically, Carboprost was the primary medication in only 0.14% of PPH cases, Ergonovine in 0.13% of PPH cases, and Tranexamic in 0.55% of PPH cases. These figures suggest that Carboprost, Ergonovine, and Tranexamic are more likely to be employed as secondary treatments in severe PPH cases, where first-line treatments such as Oxytocin or Carbetocin may have proven insufficient. This aligns with existing literature, which indicates that these medications are often utilized as second-line uterotonics in more severe cases, reflecting their role in managing complicated PPH scenarios rather than serving as initial preventive measures. This aligns with literature indicating that these medications are often used as second-line uterotonics in more severe cases where initial treatments may not have been sufficient [53].

For Oxytocin, the data reveals a noteworthy trend. Only 2.8% of PPH cases and 29.42% of non-PPH cases took Oxytocin. Conversely, 5.01% of PPH cases, compared to 62.77% of non-PPH cases, did not take Oxytocin. This suggests that Oxytocin administration is associated with a lower incidence of PPH, as a higher percentage of PPH cases were observed among those not receiving Oxytocin.

Examining the different types of Oxytocin administration provides further insight. The Oxytocin Pre-Cut Flag, indicating Oxytocin use before delivery, was present in 1.98% of PPH cases and 23.02% of non-PPH cases, whereas 5.83% of PPH cases and 69.18% of non-PPH cases did not take Oxytocin before incision. This indicates that prophylactic use of Oxytocin before delivery is associated with reduced PPH incidence. The Oxytocin Primary Flag, indicating Oxytocin as the primary medication to prevent PPH, was present in 2.53% of PPH cases and 28.34% of non-PPH cases, compared to 5.28% of PPH cases and 63.85% of non-PPH cases that did not take Oxytocin as a primary form of prevention. This further underscores that Oxytocin usage, particularly as a primary medication, is linked to a lower occurrence of PPH. When compared to literature prophylactic use of Oxytocin provides the best result in preventing PPH, with the lowest incidence of PPH among the

cases [54]. This is followed by the use of Oxytocin as the primary medication, which is slightly less effective but still significantly preventive.

Shifting focus to Carbetocin, the Carbetocin Flag indicates that Carbetocin was used in 5.05% of PPH cases and 60.58% of non-PPH cases, while 2.76% of PPH cases and 31.62% of non-PPH cases did not use Carbetocin. The significant difference in non-PPH cases suggests a strong preventive effect of Carbetocin. 4.3% of PPH cases and 58.01% of non-PPH cases used Carbetocin as a primary form of preventative medication, compared to 3.51% of PPH cases and 34.18% of non-PPH cases not using Carbetocin as a primary preventative medication. This stark difference highlights Carbetocin's effectiveness in preventing PPH, as the high usage in non-PPH cases suggests that Carbetocin significantly reduces the risk of developing PPH.

When comparing Carbetocin and Oxytocin, Oxytocin appears to be associated with a lower incidence of PPH. Only 2.8% of PPH cases had received Oxytocin, compared to 29.42% of non-PPH cases, suggesting that Oxytocin administration is linked to better prevention of PPH. Conversely, 5.05% of PPH cases had received Carbetocin, while 60.58% of non-PPH cases had received it. This significant difference in usage between PPH and non-PPH cases for Carbetocin suggests that, while it is widely used, its effectiveness may not be as strong as Oxytocin in preventing PPH. The high usage rate of Carbetocin among non-PPH patients implies it might be administered more frequently than necessary.

In conclusion, while both drugs are effective in reducing the incidence of PPH, Carbetocin is used more extensively than Oxytocin. However, Oxytocin seems to provide better results in preventing PPH, as indicated by the lower percentage of PPH cases among those administered the drug. The larger difference between PPH and non-PPH cases for Carbetocin suggests that it might be used more frequently than necessary, highlighting a potential area for optimizing PPH prevention strategies by possibly favoring Oxytocin, which appears more effective in certain clinical scenarios.

5.1.9 Distributions and Bi-Variant Correlations of Data Between Real and Synthetic Datasets

The Hellinger Distance is a measure used to quantify the similarity between two probability distributions, ranging from 0 to 1. A Hellinger Distance of 0 indicates that the distributions are identical, while a value of 1 signifies that the distributions are completely different [26]. In practical applications, there are no universally accepted absolute thresholds for what constitutes a "good" or "bad" Hellinger Distance. The interpretation of these values can depend significantly on the specific context and requirements of the analysis. Thus, while certain thresholds may be suggested in various

studies or applications, their appropriateness can vary based on the particular use case and domain-specific considerations.

Figure 3 shows that variables such as 'Epilepsy/Seizure Disorders Medications' (0.478), 'Asthma Medications' (0.472), and 'Urinary Tract Infections Diagnosis' (0.457) had Hellinger Distance values significantly above the acceptable threshold, suggesting substantial differences between the real and synthetic datasets. 'Gestational Diabetes Medications' (0.395), 'Depression/Anxiety Disorders Medications' (0.366), 'Hypertension Diagnosis' (0.357), 'Hemoglobin Drop' (0.338), 'Preoperative Hemoglobin' (0.292), and 'Postoperative Hemoglobin' (0.214), mainly pertain to medications and diagnoses for comorbidities which seems odd. Therefore, when going back to check the missingness values from Figure 2, these variables all had significantly high censorship or NaN values. This high level of censorship seems to be what contributes to their significant Hellinger Distances, indicating that the synthetic dataset may not effectively capture the nuances present in the real data for these specific variables. These variables, mainly medications and diagnoses for comorbidities, are critical in PPH prediction. Their significant differences could impact the accuracy and generalizability of predictive models when using synthetic data.

5.2 Pearson Correlations

5.2.1 Overview of Correlations

Figure 4 shows the complete correlations between all the variables in both the real and synthetic datasets.

Notably, the synthetic data exhibits higher correlations in the diagnosis and medication variables compared to the real data, especially evident in the bottom right-hand side of the synthetic data heatmap. This higher correlation could be due to the high censorship in this part of the data which was then had to be adjusted for. This is a potential artifact of the synthetic data generation process, which unfortunately needs to be further investigated by MDClone®.

Additionally, there appears to be a generally slightly higher correlation between the variables in the real dataset as opposed to the synthetic dataset. This discrepancy suggests that the real data captures more nuanced relationships between variables that the synthetic data does not fully replicate. These slight differences in correlation patterns could impact the performance of predictive models, emphasizing the importance of using real data for the most accurate results when possible.

Overall, while the synthetic data provides a close approximation of the real data, some discrepancies in correlation strength suggest areas for further refinement in synthetic data generation

processes to ensure even more accurate representation of real-world relationships. This is critical for the reliability and validity of predictive models trained on synthetic data, particularly in sensitive applications such as healthcare.

5.2.2 Correlations with PPH Positive Value

The highest feature correlations with PPH are compared between the real and synthetic datasets to evaluate the consistency of relationships identified in Table 2. For instance, the Carboprost Flag shows the highest correlation with PPH at 0.27 in the real dataset and 0.22 in the synthetic dataset. Similarly, the Ergonovine Flag has a correlation of 0.25 in the real data and 0.19 in the synthetic data. Other features like the Tranexamic Acid Flag (0.19 real, 0.15 synthetic), Intraoperative Transfusion Flag (0.18 real, 0.15 synthetic), and Alternative Drug Flag (0.15 real, 0.12 synthetic) also demonstrate notable correlations, albeit modest.

Carboprost, Ergonovine Flag, and Tranexamic Acid Flag, which had the highest correlations, are secondary drugs administered in addition to the default/primary drug Carbetocin or primary drug Oxytocin. When examining the administration patterns, it was found that Carboprost and Ergonovine were almost never given as primary drugs and were only administered to approximately 10% of patients as secondary drugs alongside the two primary preventative medication; either Carbetocin or Oxytocin. Tranexamic acid was more liberally used as a secondary medication, given to about 33% of patients, and as a primary form of medication to approximately 5% of patients. The high correlation observed with these drugs is significant because only about 10% of the patients developed PPH. This suggests that while there is not a high correlation between these alternative drugs being taken as primary treatments and the development of PPH, it may be worthwhile to further investigate the relationship between PPH outcomes and these alternative medications used as secondary preventative measures.

However, when referencing Table 2, some significant discrepancies are evident as well. For example, the Carboprost Primary Flag and Revisit Emergency Department Flag, which are present in the real dataset with Pearson correlation coefficients of 0.10 each, do not appear in the synthetic dataset. Conversely, the synthetic dataset includes features like the Gestational Diabetes Medication Flag and Urinary Tract Infections Diagnosis with correlations of 0.09 each, which are not top features in the real dataset. Even though the variables missing in the real data and synthetic correlations have very small correlation coefficients, they may still be important. Carboprost administered as a primary drug instead of as a secondary drug could be vital to prevent PPH. For the variable Revisit

Emergency Department Flag, it is likely not as important because PPH generally will develop within 1 day of giving birth, during which most women are in the hospital anyway [55]. Therefore, if a patient were to revisit the emergency department after giving birth, it most likely isn't related to the development of PPH. Gestational Diabetes Medication Flag and Urinary Tract Infections Diagnosis in the real data set correlations could be potential artifacts due to the way the high censorship of these features were adjusted for.

This analysis indicates that the synthetic data captures some but not all the important relationships present in the real data. The differences suggest potential limitations in the synthetic data's ability to fully replicate the real data's predictive patterns. This can impact the model's ability to accurately predict PPH when using synthetic data, as some critical correlations might be underrepresented or missed entirely.

The overall Pearson correlation coefficients for the features in both the real and synthetic datasets are relatively low, with the highest being 0.27 for the Carboprost Flag in the real dataset. This suggests that even the strongest predictors of PPH in the datasets have a modest linear relationship with the outcome. Generally, in many fields, a Pearson correlation coefficient above 0.3 is considered a weak to moderate correlation, above 0.5 is moderate, and above 0.7 is considered strong [56]. The coefficients observed in this thesis fall below these thresholds, indicating weak linear relationships.

Several factors could contribute to these low correlation coefficients. Firstly, PPH is likely influenced by a multitude of factors, both clinical and non-clinical, which interact in complex ways. Simple linear relationships may not fully capture the intricacies of these interactions, hence the low correlation values. Secondly, the presence of high censorship in the synthetic and how that data was handled could have obscured the true relationships between features and the outcome potentially reducing the apparent strength of some correlations.

Moreover, the relationships between predictors and PPH might be non-linear or involve interactions that are not captured by simple Pearson correlations. Decision trees and other machine-learning algorithms were used to capture these complex relationships.

5.3 Cross-Validation

Table 4 presents the 5-Fold Cross Validated AUC Scores for various machine learning models applied to both real and synthetic datasets, providing a comprehensive comparison of their predictive performance. The key observation from this table is that the mean AUC scores for both real and synthetic data are quite high across all models, indicating strong overall model performance.

The Random Forest model demonstrates exceptional performance with near-perfect scores, achieving a mean AUC score of 0.99 for real data and 0.98 (± 0.013) for synthetic data. This slight discrepancy underscores that synthetic data maintains complex relationships, making it highly effective for sophisticated models like Random Forest. The SVM model follows closely, showing consistent high performance with both datasets, with mean AUC scores of 0.97 for both real and synthetic data. The low standard deviation for synthetic data (± 0.010) indicates minimal variability and stable performance across different folds, implying reliable results.

The Decision Tree model performs well with both datasets, with mean AUC scores of 0.93 for real data and 0.92 (± 0.0053) for synthetic data. The minimal difference in scores highlights the synthetic data's effectiveness in replicating real data characteristics. The Logistic Regression model shows a slight performance drop with synthetic data, with a mean AUC score of 0.85 for real data and 0.81 (± 0.014) for synthetic data. This minor difference suggests that the synthetic dataset closely approximates the real dataset but may not capture all intricacies.

In summary, the mean AUC scores are high for all models across both real and synthetic datasets, demonstrating robust model performance. The standard deviations for the synthetic data are low, indicating minimal variability and stable performance across different folds. These metrics suggest that synthetic data generated by MDClone® is a reliable substitute for real data, maintaining high utility for training predictive models while addressing privacy concerns.

5.4 Model Performance

When comparing the model performance metrics between the synthetic model performance in Table 5 and real model performance in Table 6, it is evident that the Random Forest and Decision Tree models consistently demonstrate superior performance in predicting PPH.

For the synthetic dataset, the Random Forest model performed the best with an F1 score of 0.95, an accuracy of 0.95, and an AUC-ROC of 0.98. The Decision Tree model followed closely with an F1 score of 0.92, an accuracy of 0.92, and an AUC-ROC of 0.93. These models excel due to their ability to handle complex, non-linear relationships between features and outcomes. Additionally, these tree-based models can manage missing data more effectively and are less sensitive to the distribution of the input data, which is advantageous given the variability observed between the real and synthetic datasets.

Logistic Regression and SVM models also perform adequately on the synthetic dataset. Logistic Regression achieved an accuracy of 0.77, an F1 score of 0.76, and an AUC-ROC of 0.81. The SVM model, which can handle non-linear relationships through kernel functions, shows better adaptability.

It achieves an accuracy of 0.90, an F1 score of 0.89, and an AUC-ROC of 0.97. Despite its adaptability, SVM's performance is slightly lower than that of the Random Forest model.

For the real dataset, the Random Forest model consistently demonstrates superior performance in predicting PPH. It attained an F1 score of 0.96, an accuracy of 0.96, and an AUC-ROC of 0.98. The Decision Tree model also performed well with an F1 score of 0.92, an accuracy of 0.93, and an AUC-ROC of 0.92. These high F1 scores reflect the models' ability to maintain a balance between precision (the ability to correctly identify positive cases) and recall (the ability to capture all relevant positive cases). The Random Forest model benefits from its ensemble nature, where multiple decision trees are combined to improve overall accuracy and reduce overfitting. This model's robustness makes it well-suited for capturing intricate patterns in the real dataset, leading to high precision and recall values.

Logistic Regression and SVM models also perform adequately on the real dataset. Logistic Regression provides an overall accuracy of 0.73, an F1 score of 0.72, and an AUC-ROC of 0.85. The SVM model achieves an accuracy of 0.93, an F1 score of 0.91, and an AUC-ROC of 0.97.

Comparing the best models from both datasets, the Random Forest model shows similar high performance in both the synthetic and real datasets with an AUC-ROC of 0.98, indicating that the model's predictive power is consistent across different types of data. The Decision Tree model also shows robust performance, though slightly lower than the Random Forest model.

Overall, the comparison between the models trained on real data and those trained on synthetic data reveals that both datasets provide consistent results, with the Random Forest model showing the highest performance across all metrics. The slight differences in precision, recall, F1-score, accuracy, and AUC-ROC between the real and synthetic datasets suggest that the synthetic data is a reliable proxy for real data in training predictive models for PPH.

5.5 Random Forest Best Decision Tree Analysis

Figure 6, Figure 7, and Figure 8 illustrate the pruned decision trees and their pathways derived from both synthetic and real datasets. Analyzing the best decision tree models for both datasets reveals both similarities and differences in how these models prioritize and split on various features, shedding light on the consistency and reliability of synthetic data in predicting PPH. These figures can also help highlight which variables are most important in predicting PPH.

In the context of predicting and understanding Postpartum Hemorrhage (PPH), it is more beneficial to analyze the pathways that end with class 1 (indicating a higher risk of PPH). The reason for this is that the goal is to identify and understand the factors and conditions that lead to the development of

PPH. By focusing on the pathways that end with class 1, we can gain insights into the specific variables and their combinations that are most indicative of PPH risk, which is crucial for developing effective preventive strategies and interventions. Analyzing pathways that end with class 0 (indicating no PPH) can also be valuable as it highlights factors and conditions associated with a lower risk of PPH.

5.5.1 Synthetic Data Decision Tree Analysis

When analyzing pathway 6 from Figure 7 it can be seen that patients who received Carboprost (Carboprost Flag > 0.5), did not receive an intraoperative transfusion (Intraoperative Transfusion Flag ≤ 0.5), and were older than 42.5 years are at higher risk of developing PPH. This pathway has a Gini impurity of 0.278, indicating some degree of impurity, but it is relatively low, suggesting that the majority of samples in this node belong to a single class. Specifically, out of 6 samples, 5 developed PPH, giving a value of [1, 5]. This means that one sample did not develop PPH, while five did. The small sample size of 6, while statistically limited, highlights the importance of age as a significant risk factor. Despite the administration of Carboprost, older patients are still at high risk, suggesting that Carboprost alone is insufficient as a preventive measure for older patients. This is important as Carboprost is identified as a second-line uterotonic, making it more effective in conjunction with other first-line uterotonic agents like Carbetocin or Oxytocin [53]. This underscores the need for careful monitoring and additional preventive strategies in this demographic.

Similarly, Pathway 7 seen in Figure 7 reveals that patients who received Carboprost and an intraoperative transfusion but had a hemoglobin drop of ≤ 1.0 are at higher risk of developing PPH. This pathway has a Gini impurity of 0.0, indicating perfect purity, which means that all samples in this node belong to a single class. Specifically, out of 11 samples, all 11 developed PPH, giving a value of [0, 11]. The zero impurity and consistent outcome across all samples in this pathway strongly suggest that despite the administration of both Carboprost and a transfusion, patients with minimal hemoglobin drop can still be at high risk. This highlights the critical role of hemoglobin levels in predicting PPH risk. The fact that all samples resulted in PPH emphasizes the need for a comprehensive approach to managing and preventing PPH that goes beyond single interventions. Even with proactive measures like Carboprost administration and transfusions, monitoring hemoglobin levels remains crucial for identifying high-risk patients and implementing timely interventions [51].

5.5.2 Real Data Decision Tree Analysis

Figure 8 illustrates the pathways from the real data decision tree. This section will focus on the factors that potentially influence the development of PPH. By identifying and understanding these critical variables and their interactions, we aim to highlight the most significant predictors of PPH risk.

When interpreting Pathway 2, it can be seen that individuals who did not receive Carboprost (Carboprost Flag ≤ 0.5) or Ergonovine (Ergonovine Flag ≤ 0.5) but had an intraoperative transfusion volume greater than 284.5 mL developed PPH. This pathway has a Gini impurity of 0.432, indicating moderate impurity, with 19 samples, of which 13 developed PPH (value = [6, 13]). This pathway indicates that the absence of second-line uterotonic agents like Carboprost and Ergonovine did not counteract the effect of substantial blood loss during birth, necessitating a blood transfusion greater than 284.5 mL. PPH management specifies that once the amount of blood lost during a C-section exceeds 750-1000 mL, a blood transfusion needs to be given to the patient [57]. Depending on the severity of the blood loss, the patient may be stabilized with just one unit of whole blood, which is approximately 500 mL, or in severe cases, the MTP (Massive Transfusion Protocol) needs to be implemented where patients could receive up to 10 units of blood to maintain life function [57].

Similarly, Pathway 3 reveals that patients who did not receive Carboprost (Carboprost Flag ≤ 0.5) but received Ergonovine (Ergonovine Flag > 0.5) and had postoperative hemoglobin levels ≤ 75.5 g/L developed PPH. This suggests that despite the administration of Ergonovine, the significant drop in hemoglobin levels post-operation indicates that severe blood loss might have occurred during birth. Additionally, this could signal that if PPH hasn't already occurred, it is likely imminent, as a postoperative hemoglobin lower than 100 g/L indicates extensive blood loss or anemia [58]. This pathway has a Gini impurity of 0.346, indicating some degree of impurity but relatively low, with 9 samples, of which 7 developed PPH (value = [2, 7]). The low postoperative hemoglobin highlights the critical need for close monitoring and potentially additional interventions to manage severe blood loss effectively.

Transitioning to Pathway 6, we observe that patients who did not receive Carboprost (Carboprost Flag > 0.5) or Ergonovine (Ergonovine Flag ≤ 0.5) and were older than 39.5 years of age are at higher risk of developing PPH. This pathway has a Gini impurity of 0.355, indicating moderate impurity, with 13 samples, of which 10 developed PPH (value = [3, 10]). This highlights that older age is a significant risk factor even when Carboprost is administered. The small sample size, while limited, emphasizes the importance of age as a risk factor. This finding is expected as an older maternal age (generally > 35 years of age) increases the incidence of PPH [59].

When interpreting Pathways 7 and 8 together, a consistent theme emerges regarding the significant risk factors for PPH. Pathway 7 shows that patients who did not receive Carboprost (Carboprost Flag > 0.5) or Ergonovine (Ergonovine Flag > 0.5) and had a gestational age ≤ 35.5 weeks are at higher risk of developing PPH. This pathway has a Gini impurity of 0.0, indicating perfect purity, meaning all samples in this node belong to a single class. Out of 13 samples, all 13 developed PPH (value = [0, 13]). This highlights that preterm birth is a significant risk factor for PPH. The pure node and small sample size emphasize the high risk associated with preterm gestational age. A clear link exists between preterm birth and an increased risk of PPH.

Similarly, Pathway 8 reveals that patients who did not receive Carboprost (Carboprost Flag > 0.5) or Ergonovine (Ergonovine Flag > 0.5) and had a gestational age > 35.5 weeks also developed PPH. This pathway has a Gini impurity of 0.498, indicating almost maximum impurity, which means the samples are nearly equally split between the classes, with a slight majority belonging to the PPH class. Out of 89 samples, 47 developed PPH (value = [42, 47]). This pathway underscores that even patients with full-term gestations are at a significant risk of PPH when Carboprost and Ergonovine are not administered.

Comparing Pathways 7 and 8, it becomes evident that the administration of Carboprost and Ergonovine is crucial in mitigating the risk of PPH, regardless of gestational age. Both preterm and full-term patients without these medications exhibit a high incidence of PPH, emphasizing the importance of these second-line uterotonic agents in PPH prevention strategies.

5.5.3 Similarities and Differences Between Real and Synthetic Decision Tree Pathways

Analyzing the decision tree pathways derived from both synthetic and real datasets reveals several similarities and differences that are crucial for understanding the predictive factors of PPH. Both datasets highlighted certain common variables, such as the Carboprost Flag, Intraoperative Transfusion Volume, and Hemoglobin levels, which play significant roles in determining the risk of PPH.

In terms of similarities, the importance of the Carboprost Flag was evident in both datasets. For the synthetic data, Pathways 6 and 7 indicate that patients without Carboprost administration were at higher risk, highlighting Carboprost as a crucial factor in managing PPH. Similarly, for the real data, Pathways 2 and 3 show that patients who did not receive Carboprost (Carboprost Flag ≤ 0.5) had significant splits leading to PPH outcomes. In terms of similarities, the importance of the Carboprost Flag was evident in both datasets.

Age was also consistently identified as a risk factor in both datasets. In the synthetic data, Pathway 6 indicates that patients aged > 42.5 years are at higher risk. Similarly, in the real data, Pathway 6 highlights that patients aged > 39.5 years are at higher risk of developing PPH. The consistent identification of older maternal age in both datasets emphasizes its significance as a predictor of PPH.

Furthermore, Hemoglobin levels were another key variable in both datasets. In the synthetic data, Pathway 7 demonstrates that patients with a hemoglobin drop ≤ 1.0 are at higher risk of PPH. In the real data, Pathways 3 and 4 show that lower postoperative hemoglobin levels (≤ 75.5 g/L) are associated with higher PPH risks, aligning with the importance of hemoglobin levels in predicting PPH outcomes.

However, there are notable differences between the datasets. Gestational age plays a significant role in the real data, as indicated by Pathways 7 and 8, however this factor is not prominently featured in the synthetic data pathways, suggesting that the synthetic data might not capture the nuances of gestational age as effectively.

Overall, while the synthetic data set shows many similarities to the real data, including identifying key variables like Carboprost Flag, Age, and Hemoglobin levels, the granularity of insights from the synthetic data may be somewhat limited. This limitation is reflected in the fewer pathways leading to PPH-positive leaf nodes in the synthetic data. Despite these differences, the synthetic data remains a valuable tool, providing insights that are broadly consistent with the real data, and can be particularly useful when real data is scarce or sensitive.

5.6 Limitations/Biases and Future Directions

There are several limitations and biases that need to be addressed to ensure the effective use of synthetic data for predictive modeling in healthcare. Firstly, the quality and accuracy of synthetic data are heavily dependent on the algorithms used for its generation. In this study, GAN-generated data might not fully capture the complexity and variability of real-world clinical data, particularly for rare conditions and treatments, which could potentially introduce biases if the generative model does not accurately represent all subpopulations or specific clinical scenarios. To address this limitation, future research should explore the use of other advanced generative models and algorithms, ensuring that the training data is diverse and representative of the target population, covering various clinical scenarios and subpopulations.

Feature selection and engineering pose another limitation, as these processes can introduce biases if certain relevant features are omitted or improperly transformed, impacting the model's performance and its ability to generalize to new data. Selection bias might occur if the features chosen for

modeling are not representative of the underlying factors influencing PPH. This limitation can be addressed by utilizing systematic and unbiased feature selection methods and incorporating domain expertise and clinical knowledge from health care professionals to guide the processes.

The study's model validation and evaluation processes also present limitations. Although standard metrics were used for validation, real-world performance might differ due to unseen factors or data variations, leading to validation bias if the evaluation metrics do not fully capture the model's performance across different settings and populations. To mitigate this, conducting external validation using independent datasets from different healthcare institutions and employing a comprehensive set of evaluation metrics, including sensitivity, specificity, precision, recall, F1 score, and AUC-ROC, can provide a more holistic view of model performance.

High levels of missingness/censorship in the data are another concern, as they can obscure true relationships between features and outcomes, potentially leading to biased models. Censorship bias might occur if the method used to handle missing data are not robust. Robust imputation methods and sensitivity analyses should be employed to handle missing data effectively, and the methods used should be clearly documented and justified to ensure transparency and reproducibility.

5.7 Conclusion

This thesis explored the viability of using synthetic data as a proxy for real clinical data in predictive modeling, specifically for PPH. Through a comprehensive analysis of data distributions, feature correlations, and model performances, the findings highlight both the potential advantages and challenges of employing synthetic data in clinical research.

The evaluation of data distributions using the Hellinger Distance revealed significant differences for key variables related to medications and comorbidities, primarily due to high levels of missingness/censorship in the synthetic data. These discrepancies underscore the need for improving synthetic data generation processes to better capture the complexity of real-world clinical scenarios. Despite these differences, most variables demonstrated acceptable Hellinger Distances, indicating that synthetic data can approximate real data distributions for many clinical features.

Feature correlations with PPH showed some variation between the real and synthetic datasets. Notably, while certain features like the Carboprost Flag, Ergonovine Flag, and Tranexamic Acid Flag were consistently correlated with PPH in both datasets, discrepancies in other features highlighted limitations in the synthetic data's ability to replicate all real-world relationships. Nonetheless, the overall similarity in feature correlations suggests that synthetic data can capture many of the key factors associated with PPH.

Model performance comparisons revealed that tree-based models, particularly the Random Forest and Decision Tree classifiers, exhibited high accuracy and F1 scores for both real and synthetic data. These models outperformed others, such as Logistic Regression and SVM, indicating their robustness in handling the complex interactions of clinical variables. The high degree of similarity in the performance metrics between real and synthetic data models reinforces the potential of synthetic data to serve as a reliable proxy in predictive modeling.

The AUC-ROC scores and 5-fold cross-validated AUC scores further validated the models' generalizability. High AUC-ROC scores for both real and synthetic data models demonstrated strong discriminatory power, while consistent cross-validated AUC scores underscored the models' stability and generalizability.

Decision tree analysis provided insights into the pathways leading to higher PPH risk. The consistent presence of variables such as the Carboprost Flag, Intraoperative Transfusion Volume, Gestational Age, and Hemoglobin levels across both real and synthetic data trees highlighted their significance in PPH prediction. The detailed breakdown of splits, Gini impurity, and sample distributions emphasized the interpretability and clinical relevance of the decision tree models.

Overall, this thesis demonstrates that synthetic data, when generated and validated appropriately, can serve as a viable proxy for real clinical data in predictive modeling. The findings support the utility of synthetic data in developing robust and accurate predictive models for PPH, potentially enhancing clinical decision-making and patient outcomes. However, addressing limitations related to data generation, feature selection, and validation processes is crucial for further improving the reliability and applicability of synthetic data in healthcare research.

By advancing synthetic data methodologies and integrating them with clinical expertise, researchers can leverage the benefits of synthetic data while mitigating potential biases and inaccuracies. This approach holds promise for accelerating medical research, safeguarding patient privacy, and ultimately improving healthcare delivery.

References

1. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. 2023; 2: e0000082.
2. El Kababji S, Mitsakakis N, Fang X, Beltran-Bless AA, Pond G, Vandermeer L, Radhakrishnan D, Mosquera L, Paterson A, Shepherd L, Chen B. Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clinical Cancer Informatics*. 2023 Nov;7:e2300116.
3. Newsome J, Martin JG, Bercu Z, Shah J, Shekhani H, Peters G. Postpartum hemorrhage. *Techniques in vascular and interventional radiology*. 2017 Dec 1;20(4):266-73.
4. Bienstock JL, Eke AC, Hueppchen NA. Postpartum hemorrhage. *New England Journal of Medicine*. 2021 Apr 29;384(17):1635-45.
5. Mehrabadi A, Liu S, Bartholomew S, Hutcheon JA, Kramer MS, Liston RM, Joseph KS, Maternal Health Study Group of the Canadian Perinatal Surveillance System. Temporal trends in postpartum hemorrhage and severe postpartum hemorrhage in Canada from 2003 to 2010. *Journal of Obstetrics and Gynaecology Canada*. 2014 Jan 1;36(1):21-33.
6. Oyelese Y, Ananth CV. Postpartum hemorrhage: epidemiology, risk factors, and causes. *Clinical obstetrics and gynecology*. 2010 Mar 1;53(1):147-56.
7. Edhi MM, Aslam HM, Naqvi Z, Hashmi H. Post partum hemorrhage: causes and management. *BMC research notes*. 2013 Dec;6:1-6.
8. Nyfløt LT, Sandven I, Stray-Pedersen B, Pettersen S, Al-Zirqi I, Rosenberg M, Jacobsen AF, Vangen S. Risk factors for severe postpartum hemorrhage: a case-control study. *BMC pregnancy and childbirth*. 2017 Dec;17:1-9.
9. Oyelese Y, Scorza WE, Mastrolia R, Smulian JC. Postpartum hemorrhage. *Obstetrics and gynecology clinics of North America*. 2007 Sep 1;34(3):421-41.
10. Magann EF, Evans S, Hutchinson M, Collins R, Lanneau G, Morrison JC. Postpartum hemorrhage after cesarean delivery: an analysis of risk factors. *Southern medical journal*. 2005 Jul 1;98(7):681-6.
11. Attilakos G, Psaroudakis D, Ash J, Buchanan R, Winter C, Donald F, Hunt LP, Draycott T. Carbetocin versus oxytocin for the prevention of postpartum haemorrhage following caesarean section: the results of a double-blind randomised

- trial. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2010 Jul;117(8):929-36.
12. Oladapo OT, Okusanya BO, Abalos E, Gallos ID, Papadopoulou A. Intravenous versus intramuscular prophylactic oxytocin for reducing blood loss in the third stage of labour. *Cochrane Database of Systematic Reviews*. 2020(11).
 13. Holleboom CA, Van Eyck J, Koenen SV, Kreuwel IA, Bergwerff F, Creutzberg EC, Bruinse HW. Carbetocin in comparison with oxytocin in several dosing regimens for the prevention of uterine atony after elective caesarean section in the Netherlands. *Archives of gynecology and obstetrics*. 2013 Jun;287:1111-7.
 14. Gallos ID, Coomarasamy A. Carbetocin: worth the extra expense?. *Best Practice & Research Clinical Obstetrics & Gynaecology*. 2019 Nov 1;61:55-65.
 15. Van Der Nelson HA, Draycott T, Siassakos D, Yau CW, Hatswell AJ. Carbetocin versus oxytocin for prevention of post-partum haemorrhage at caesarean section in the United Kingdom: an economic impact analysis. *European Journal of Obstetrics & Gynecology and Reproductive Biology*. 2017 Mar 1;210:286-91.
 16. Chen RJ, Lu MY, Chen TY, Williamson DF, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*. 2021 Jun;5(6):493-7.
 17. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. *Journal of big Data*. 2022 Jan 6;9(1):3.
 18. Torfi A, Fox EA, Reddy CK. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences*. 2022 Mar 1;586:485-500.
 19. Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, Loh F, Kollef MH, Maddox T, Evanoff B, Dror H. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA open*. 2020 Dec 1;3(4):557-66.
 20. Tyagi AK, Chahal P. Artificial intelligence and machine learning algorithms. In: *Challenges and applications for implementing machine learning in computer vision 2020* (pp. 188-219). IGI Global.
 21. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *Journal of medical Internet research*. 2019 May 31;21(5):e13484.

22. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *Journal of medical Internet research*. 2019 May 31;21(5):e13484.
23. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*. 2003 Mar 1;19(1):1.
24. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*. 2023 Oct 9;6(1):186.
25. Kokosi T, Harron K. Synthetic data in medical research. *BMJ medicine*. 2022;1(1).
26. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*. 2022 Apr 7;10(4):e35734.
27. Kingsford C, Salzberg SL. What are decision trees?. *Nature biotechnology*. 2008 Sep;26(9):1011-3.
28. Yuan Y, Wu L, Zhang X. Gini-impurity index analysis. *IEEE Transactions on Information Forensics and Security*. 2021 Apr 30;16:3154-69.
29. Rigatti SJ. Random forest. *Journal of Insurance Medicine*. 2017 Jan 1;47(1):31-9.
30. Noble WS. What is a support vector machine?. *Nature biotechnology*. 2006 Dec 1;24(12):1565-7.
31. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. *Logistic regression*. New York: Springer-Verlag; 2002 Aug.
32. Miao J, Zhu W. Precision–recall curve (PRC) classification trees. *Evolutionary intelligence*. 2022 Sep;15(3):1545-69.
33. Yedidia A. Against the F-score. URL: <https://adamyedidia.files.wordpress.com/2014/11/fscore.pdf>. 2016 Dec 8.
34. Tharwat A. Classification assessment methods. *Applied computing and informatics*. 2021 Jan 4;17(1):168-92.
35. Narkhede S. Understanding auc-roc curve. *Towards data science*. 2018 Jun 26;26(1):220-7.
36. Koopmans CM, Van der Tuuk K, Groen H, Doornbos JP, De Graaf IM, Van Der Salm PC, Porath MM, Kuppens SM, Wijnen EJ, Aardenburg R, Van Loon AJ.

- Prediction of postpartum hemorrhage in women with gestational hypertension or mild preeclampsia at term. *Acta obstetrica et gynecologica Scandinavica*. 2014 Apr;93(4):399-407.
37. Ranjbar A, Ghamsari SR, Boujarzadeh B, Mehrnoush V, Darsareh F. Predicting risk of postpartum hemorrhage using machine learning approach: A systematic review. *Gynecology and Obstetrics Clinical Medicine*. 2023 Jul 12.
 38. Liu J, Wang C, Yan R, Lu Y, Bai J, Wang H, Li R. Machine learning-based prediction of postpartum hemorrhage after vaginal delivery: combining bleeding high risk factors and uterine contraction curve. *Archives of Gynecology and Obstetrics*. 2022 Oct;306(4):1015-25.
 39. Rajula HS, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*. 2020 Sep 8;56(9):455.
 40. Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischer T. Machine learning and conventional statistics: making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*. 2022 Mar;30(3):753-7.
 41. Menard S. Logistic regression: From introductory to advanced concepts and applications. Sage publications; 2009 Apr 29.
 42. Bennett M, Hayes K, Kleczyk EJ, Mehta R. Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. *arXiv preprint arXiv:2201.02469*. 2022 Jan 7.
 43. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*. 2022 Jun 27;2:927312.
 44. Guide AR. American Society of Anesthesiologists. *Code*. 2002;99140:12.
 45. Pubu ZM, Bianba ZM, Yang G, CyRen LM, Pubu DJ, Lang KZ, Zhen B, Zhaxi QZ, Nyma ZG. Factors affecting the risk of postpartum hemorrhage in pregnant women in Tibet health facilities. *Medical science monitor: international medical journal of experimental and clinical research*. 2021;27:e928568-1.
 46. Bretherick KL, Fairbrother N, Avila L, Harbord SH, Robinson WP. Fertility and aging: do reproductive-aged Canadian women know what they need to know?. *Fertility and sterility*. 2010 May 1;93(7):2162-8.

47. Whelan A, Ciomperlik H, Ghose I, Mendez-Figueroa H, Wagner S, Wiley R. Postpartum Hemorrhage: Preterm Versus Term Pregnancies [ID: 1356883]. *Obstetrics & Gynecology*. 2023 May 1;141(5S):102S-3S.
48. Morgan AS, Mendonça M, Thiele N, David AL. Management and outcomes of extreme preterm birth. *bmj*. 2022 Jan 10;376.
49. Ekin A, Gezer C, Solmaz U, Taner CE, Dogan A, Ozeren M. Predictors of severity in primary postpartum hemorrhage. *Archives of gynecology and obstetrics*. 2015 Dec;292:1247-54.
50. Du L, Feng L, Bi S, Zhang L, Tang J, Zhong L, Zhou X, Tan H, Huang L, Lin L, Zeng S. Probability of severe postpartum hemorrhage in repeat cesarean deliveries: a multicenter retrospective study in China. *Scientific Reports*. 2021 Apr 19;11(1):8434.
51. Michaan N, Madar D, Kern G, Cohen A, Many A, Yogev Y, Grisaru D, Laskov I. Average maternal hemoglobin decline after first elective cesarean section compared to higher order cesarean sections. *International Journal of Gynecology & Obstetrics*. 2023 Jan;160(1):79-84.
52. Lucas IM, Barr EL, Barzi F, Longmore DK, Lee IL, Kirkwood M, Whitbread C, Connors C, Boyle JA, Simon D, Goodrem A. Gestational diabetes is associated with postpartum hemorrhage in Indigenous Australian women in the PANDORA study: A prospective cohort. *International Journal of Gynecology & Obstetrics*. 2021 Nov;155(2):296-304.
53. Jones AJ, Federspiel JJ, Eke AC. Preventing postpartum hemorrhage with combined therapy rather than oxytocin alone. *American journal of obstetrics & gynecology MFM*. 2023 Feb 1;5(2):100731.
54. Salati JA, Leathersich SJ, Williams MJ, Cuthbert A, Tolosa JE. Prophylactic oxytocin for the third stage of labour to prevent postpartum haemorrhage. *Cochrane Database of Systematic Reviews*. 2019(4).
55. Rath WH. Postpartum hemorrhage—update on problems of definitions and diagnosis. *Acta obstetrica et gynecologica Scandinavica*. 2011 May;90(5):421-8.
56. Cohen I, Huang Y, Chen J, Benesty J, Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. *Noise reduction in speech processing*. 2009:1-4.
57. Kogutt BK, Vaught AJ. Postpartum hemorrhage: Blood product management and massive transfusion. In *Seminars in perinatology* 2019 Feb 1 (Vol. 43, No. 1, pp. 44-50). WB Saunders.

58. Frass KA. Postpartum hemorrhage is related to the hemoglobin levels at labor: Observational study. Alexandria Journal of Medicine. 2015 Dec 1;51(4):333-7.
59. Liu CN, Yu FB, Xu YZ, Li JS, Guan ZH, Sun MN, Liu CA, He F, Chen DJ. Prevalence and risk factors of severe postpartum hemorrhage: a retrospective cohort study. BMC pregnancy and childbirth. 2021 Apr 26;21(1):332.

Appendix A

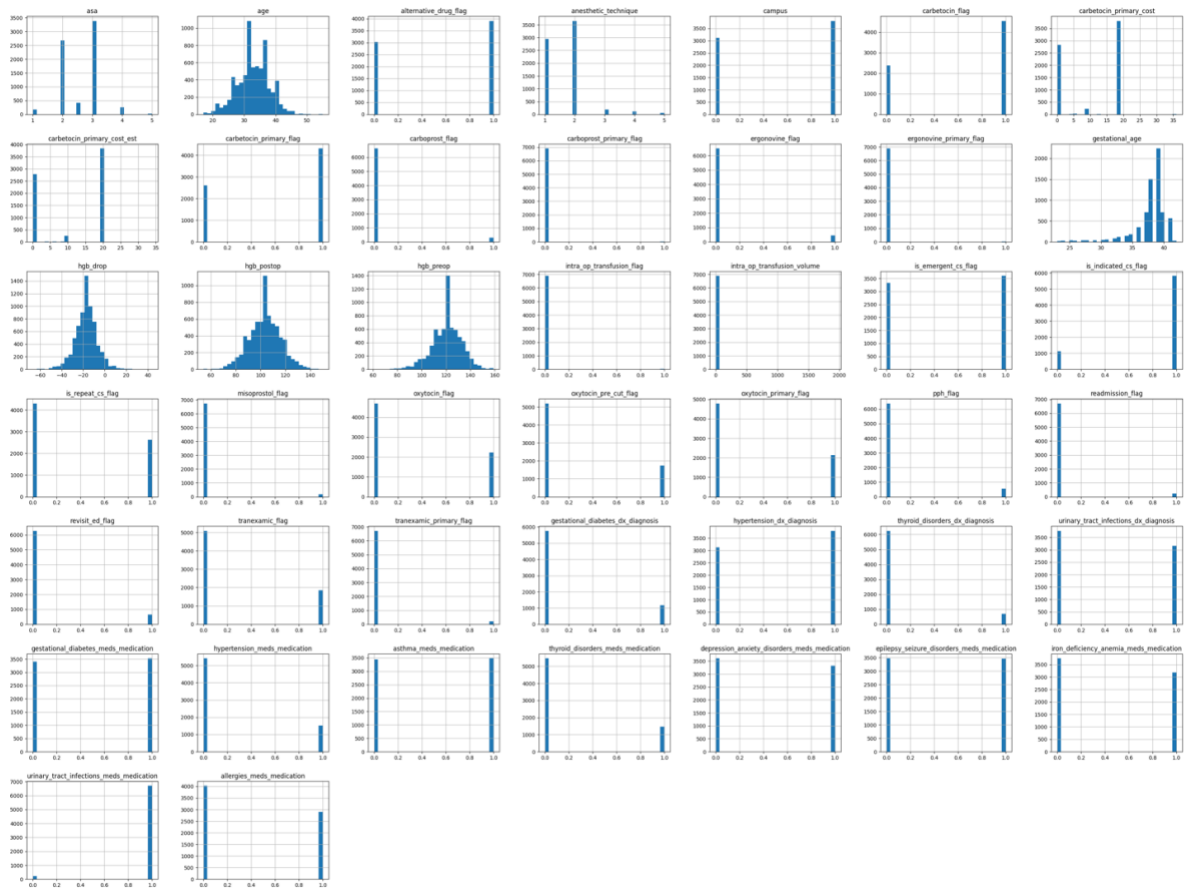
Appendix Tables 1A: Variables and Data Dictionary

Variable Name	Actual Name	Description	Data Type
Patient_ID	Patient ID	Unique identifier for each patient	Integer
asa	ASA Classification	ASA score - subjective assessment of a patient's overall health that is based on five classes (I to V).	Integer
age	Age	Patient age at admission	Integer
alternative_drug_flag	Alternative Drug Used	If any alternative drugs are given post-cut, then 1, else 0. Alternative drugs are Misoprostol, Tranexamic, Carboprost, Ergonovine, and Oxytocin	Binary (0, 1)
alternative_drug_primary_flag	Primary Alternative Drug Used	If any alternative drug is primary, then 1, else 0	Binary (0, 1)
anesthetic_technique	Anesthetic Technique	Primary anaesthetic technique (Epidural, Spinal, General, Combined Spinal and Epidural (CSE), Combined general with regional central/neuraxial (epidural or spinal)	1, 2, 3, 4, 5
campus	Hospital Campus	Campus (Civic/General)	C, G
carbetocin_cost	Carbetocin Cost	Total cost of Carbetocin	Integer
carbetocin_cost_estimated	Estimated Carbetocin Cost	Estimated cost of Carbetocin	Integer
carbetocin_dosage	Carbetocin Dosage	Total post-cut Carbetocin dosage	Integer
carbetocin_flag	Carbetocin Used	If Carbetocin is given post-cut, then 1, else 0	Binary (0, 1)
carbetocin_primary_cost	Primary Carbetocin Cost	If Carbetocin is primary, then total cost of the primary Carbetocin, else 0	Integer
carbetocin_primary_cost_est	Estimated Primary Carbetocin Cost	Estimated primary cost of Carbetocin	Integer

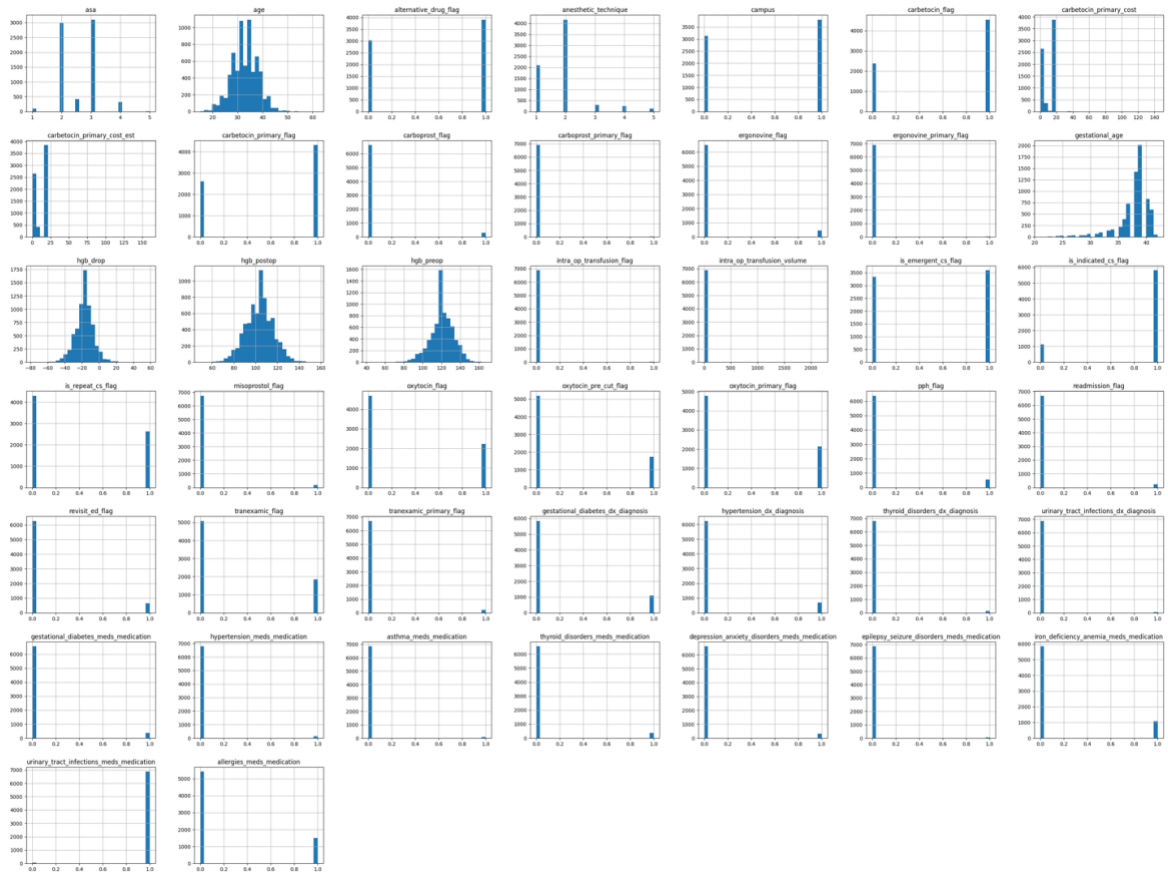
carbetocin_primary_dosage	Primary Carbetocin Dosage	If Carbetocin is primary, then the dosage of the primary Carbetocin, else 0	Integer
carbetocin_primary_flag	Primary Carbetocin Used	If Carbetocin is primary, then 1, else 0	Binary (0, 1)
carboprost_flag	Carboprost Used	If Carboprost is given post-cut, then 1, else 0	Binary (0, 1)
carboprost_primary_flag	Primary Carboprost Used	If Carboprost is primary, then 1, else 0	Binary (0, 1)
ergonovine_flag	Ergonovine Used	If Ergonovine is given post-cut, then 1, else 0	Binary (0, 1)
ergonovine_primary_flag	Primary Ergonovine Used	If Ergonovine is primary, then 1, else 0	Binary (0, 1)
first_anesthesiologist_epic	First Anesthesiologist EPIC ID	First anesthesiologist ID based on EPIC	Integer
gestational_age	Gestational Age	Gestational age in weeks	Integer
hgb_drop	Hemoglobin Drop	Difference between post-op hemoglobin and pre-op hemoglobin	Integer
hgb_postop	Postoperative Hemoglobin	Post-op hemoglobin	Integer
hgb_preop	Preoperative Hemoglobin	Pre-op hemoglobin	Integer
intra_op_transfusion_flag	Intraoperative Transfusion	If any intra-op blood transfusion, then 1, else 0	Binary (0, 1)
intra_op_transfusion_volume	Intraoperative Transfusion Volume	If any intra-op blood transfusion, then the total volume of intra-op transfusion, else 0	Integer
is_emergent_cs_flag	Emergent Cesarean Section	If the case is an emergent C-section, then 1, else 0	Binary (0, 1)
is_indicated_cs_flag	Indicated Cesarean Section	If the case is an indicated C-section, then 1, else 0	Binary (0, 1)
is_repeat_cs_flag	Repeat Cesarean Section	If the case is a repeat C-section, then 1, else 0	Binary (0, 1)

medication_cost	Medication Cost	Total cost of Carbetocin and alternative drugs. This only includes related medications	Integer
misoprostol_flag	Misoprostol Used	If Misoprostol is given post-cut, then 1, else 0	Binary (0, 1)
misoprostol_primary_flag	Primary Misoprostol Used	If Misoprostol is primary, then 1, else 0	Binary (0, 1)
multiple_meds_flag	Multiple Medications Used	If multiple medications are given post-cut, then 1, else 0	Binary (0, 1)
oxytocin_flag	Oxytocin Used	If Oxytocin is given post-cut, then 1, else 0	Binary (0, 1)
oxytocin_pre_cut_flag	Oxytocin Pre-Incision	If any Oxytocin is given pre-cut, then 1, else 0. This indicates that CS was done in labor (emergent CS). Receiving Oxytocin in labor is a risk factor for PPH	Binary (0, 1)
oxytocin_primary_flag	Primary Oxytocin Used	If Oxytocin is primary, then 1, else 0	Binary (0, 1)
post_op_transfusion_flag	Postoperative Transfusion	If any post-op transfusion, then 1, else 0	Binary (0, 1)
post_op_transfusion_volume	Postoperative Transfusion Volume	If any post-op blood transfusion, then the total volume of post-op transfusion, else 0	Integer
postop_los	Postoperative Length of Stay	Total length of stay post operation	Integer
pph_flag	Postpartum Hemorrhage	If the patient had post-partum hemorrhage (PPH), then 1, else 0	Binary (0, 1)
primary_surgeon_epic	Primary Surgeon EPIC ID	Primary surgeon ID based on EPIC	Integer
provider_id_health_record	Provider Health Record ID	Primary Provider ID based on Health Record	Integer
readmission_flag	Readmission	If the patient was readmitted within 42 days of discharge, then 1, else 0. This is all-cause readmission at TOH only	Binary (0, 1)

revisit_ed_flag	Revisit Emergency Department Flag	If the patient visited ED within 42 days of discharge, then 1, else 0. This is all-cause ED visits at TOH only	Binary (0, 1)
total_los	Total Length of Stay	Total length of stay	Integer
hgb_preop	Hemoglobin Before C-section in g/dL	Pre-op hemoglobin	HGB_Preop
hgb_postop	Hemoglobin After C-section in g/dL	Post-op hemoglobin	HGB_Postop
hgb_drop	Change in Hemoglobin Before and After C-section in g/dL	Difference between post-op hemoglobin and pre-op hemoglobin	HGB_Drop
tranexamic_flag	Tranexamic Acid Used	If Tranexamic is given post-cut, then 1, else 0	Binary (0, 1)
tranexamic_primary_flag	Primary Tranexamic Acid Used	If Tranexamic is primary, then 1, else 0	Binary (0, 1)
transfusion_flag	Transfusion	If any blood transfusion, then 1, else 0	Binary (0, 1)
transfusion_volume	Transfusion Volume	If any blood transfusion, then total volume of blood transfusion	Float
gestational_diabetes_dx_diagnosis	Gestational Diabetes Diagnosis	Flag indicating a diagnosis of gestational diabetes	Binary (0, 1)
hypertension_dx_diagnosis	Hypertension Diagnosis	Flag indicating a diagnosis of hypertension	Binary (0, 1)
asthma_dx_diagnosis	Asthma Diagnosis	Flag indicating a diagnosis of asthma	Binary (0, 1)
thyroid_disorders_dx_diagnosis	Thyroid Disorders Diagnosis	Flag indicating a diagnosis of thyroid disorders	Binary (0, 1)
depression_anxiety_disorders_dx_diagnosis	Depression/Anxiety Diagnosis	Flag indicating a diagnosis of depression or anxiety disorders	Binary (0,



Appendix Figure 1A: Data Distribution for the Synthetic Data



Appendix Figure 2A: Data Distribution for the Real Data

The age distribution in both datasets peaks around 30-35 years, while ASA scores mostly range between 2 and 3, reflecting similar preoperative health statuses. The anesthetic technique, campus distribution, and Carbetocin administration patterns also align closely. Costs for Carbetocin cluster around specific values, and gestational ages are primarily between 37 and 39 weeks in both datasets. Hemoglobin levels before and after surgery follow normal distributions. Intraoperative transfusion volumes and the flags for emergent and indicated C-sections, as well as repeat C-sections, show similar distributions. Medication administration flags for Misoprostol, Carboprost, Oxytocin (both pre- and post-cut), and Tranexamic acid reflect consistent patterns. Postoperative length of stay, PPH incidence, readmission rates, and ED revisits within 42 days also exhibit closely matched distributions. These comparisons validate the synthetic dataset's reliability in modeling PPH-related factors. It is important to note that the distribution of diagnosis and medication variables between the synthetic and real datasets did not match as closely as other variables. The

synthetic dataset had a much higher proportion of patients flagged as 'yes' for various diagnoses and medications compared to the real dataset.

Appendix Tables 2A: Demographic Information Summary Statistics

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
ASA	6930	2.58	0.61	1	2	3	3	5
Age	6930	32.85	5.11	17	30	33	36	55
Alternative Drug Flag	6930	0.56	0.50	0	0	1	1	1
Anesthetic Technique	6930	1.79	0.61	1	1	2	2	5
Campus	6930	0.55	0.50	0	0	1	1	1
Carbetocin Flag	6930	0.66	0.47	0	0	1	1	1
Carbetocin Primary Cost	6930	10.32	8.78	0	0	17.98	17.98	35.96
Carbetocin Primary Cost Est	6930	11.51	9.67	0	0	20.00	20.00	33.90
Carbetocin Primary Flag	6930	0.62	0.48	0	0	1	1	1
Carboprost Flag	6930	0.04	0.21	0	0	0	0	1
Carboprost Primary Flag	6930	0.00	0.06	0	0	0	0	1
Ergonovine Flag	6930	0.06	0.24	0	0	0	0	1
Ergonovine Primary Flag	6930	0.01	0.07	0	0	0	0	1
Gestational Age	6930	37.82	2.83	23	37	39	39	42
HGB Drop	6930	-17.38	10.27	-67	-23	-17.38	-12	44
HGB Postop	6930	103.36	12.54	54	95	103.36	112	150
HGB Preop	6930	120.50	12.14	61	114	120.50	128	159
Emergent CS Flag	6930	0.52	0.50	0	0	1	1	1
Indicated CS Flag	6930	0.84	0.37	0	1	1	1	1
Repeat CS Flag	6930	0.38	0.49	0	0	0	1	1
Misoprostol Flag	6930	0.03	0.16	0	0	0	0	1
Oxytocin Flag	6930	0.32	0.47	0	0	0	1	1
Oxytocin Pre-Cut Flag	6930	0.25	0.43	0	0	0	0	1
Oxytocin Primary Flag	6930	0.31	0.46	0	0	0	1	1
PPH Flag	6930	0.08	0.27	0	0	0	0	1
Readmission Flag	6930	0.03	0.18	0	0	0	0	1
Revisit Emergency Department Flag	6930	0.09	0.29	0	0	0	0	1
Tranexamic Flag	6930	0.27	0.44	0	0	0	1	1
Tranexamic Primary Flag	6930	0.03	0.17	0	0	0	0	1
Gestational Diabetes Diagnosis	6930	0.08	0.28	0	0	0	0	1
Hypertension Diagnosis	6930	0.33	0.47	0	0	0	1	1
Asthma Diagnosis	6930	0.00	0.06	0	0	0	0	1

Thyroid Disorders Diagnosis	6930	0.01	0.08	0	0	0	0	1
Depression/Anxiety Diagnosis	6930	0.01	0.09	0	0	0	0	1
Epilepsy/Seizure Disorders Diagnosis	6930	0.00	0.03	0	0	0	0	1
Iron Deficiency Anemia Diagnosis	6930	0.02	0.15	0	0	0	0	1
Urinary Tract Infections Diagnosis	6930	0.00	0.06	0	0	0	0	1
Acid Reflux/Heartburn Diagnosis	6930	0.01	0.08	0	0	0	0	1
Allergies Diagnosis	6930	0.01	0.09	0	0	0	0	1
Gestational Diabetes Medication	6930	0.51	0.50	0	0	1	1	1
Hypertension Medication	6930	0.22	0.41	0	0	0	0	1
Asthma Medication	6930	0.50	0.50	0	0	1	1	1
Thyroid Disorders Medication	6930	0.21	0.41	0	0	0	0	

Appendix Tables 3A: Hellinger Distance Between the Real and Synthetic Data

Variable	Hellinger Distance
Epilepsy/Seizure Disorders Medications	0.478
Asthma Medications	0.472
Urinary Tract Infections Diagnosis	0.457
Gestational Diabetes Medications	0.395
Depression/Anxiety Disorders Medications	0.366
Hypertension Diagnosis	0.357
Hemoglobin Drop	0.338
Preoperative Hemoglobin	0.292
Postoperative Hemoglobin	0.214
ASA Score	0.249
Hypertension Medications	0.243
Iron Deficiency Anemia Medications	0.238
Thyroid Disorders Medications	0.173
Thyroid Disorders Diagnosis	0.132
Anesthetic Technique	0.102
Intraoperative Transfusion Volume	0.067
Gestational Age	0.047
Carbetocin Primary Cost (Estimated)	0.054
Carbetocin Primary Cost	0.044
Age	0.036
Campus	0.019

Allergies Medications	0.159
Alternative Drug Flag	0.000071
Carbetocin Flag	0.000212
Carbetocin Primary Flag	0.000014
Carboprost Flag	0.000216
Carboprost Primary Flag	0.000009
Ergonovine Flag	0.000039
Ergonovine Primary Flag	0.000011
Intraoperative Transfusion Flag	0.000013
Emergent C-Section Flag	0.000147
Indicated C-Section Flag	0.000071
Repeat C-Section Flag	0.000119
Misoprostol Flag	0.000025
Oxytocin Flag	0.000222
Oxytocin Pre-Cut Flag	0.000088
Oxytocin Primary Flag	0.000008
PPH Flag	0.000045
Readmission Flag	0.000028
Revisit Emergency Department Flag	0.000050
Tranexamic Flag	0.000092
Tranexamic Primary Flag	0.000028
Gestational Diabetes Diagnosis	0.011
Urinary Tract Infections Medications	0.061

Appendix Tables 4: Recursive Feature Elimination Ranking for Real and Synthetic Dataset

Rank	Real Dataset Feature	Synthetic Dataset Feature
1	Tranexamic Acid Medication	Carboprost Flag
2	Carboprost Flag	Tranexamic Acid Medication
3	Ergonovine Flag	Iron Deficiency Anemia Medications
4	Preoperative Hemoglobin	Ergonovine Flag
5	Revisit Emergency Department Flag	Thyroid Disorders Medications
6	Gestational Age	Hypertension Medications
7	Hemoglobin Drop	Campus
8	Carbetocin Primary Flag	Gestational Diabetes Medications
9	Carbetocin Flag	Hypertension Diagnosis
10	Urinary Tract Infections Medications	Revisit Emergency Department Flag
11	Intraoperative Transfusion Flag	Gestational Age
12	Misoprostol Flag	Hemoglobin Drop
13	Thyroid Disorders Medications	Preoperative Hemoglobin
14	Anesthetic Technique	Emergent C-Section Flag
15	Carbetocin Primary Cost (Estimated)	Misoprostol Flag

16	Campus	Gestational Diabetes Diagnosis
17	ASA Score	Thyroid Disorders Diagnosis
18	Age	Urinary Tract Infections Medications
19	Carbetocin Primary Cost	Allergies Medications
20	Indicated C-Section Flag	Postoperative Hemoglobin
21	Carboprost Primary Flag	Intraoperative Transfusion Flag
22	Iron Deficiency Anemia Medications	Oxytocin Flag
23	Thyroid Disorders Diagnosis	Carbetocin Flag
24	Oxytocin Flag	Alternative Drug Flag
25	Hypertension Medications	Anesthetic Technique
26	Urinary Tract Infections Diagnosis	Carboprost Primary Flag
27	Gestational Diabetes Diagnosis	Repeat C-Section Flag
28	Hypertension Diagnosis	Epilepsy/Seizure Disorders Medications
29	Gestational Diabetes Medications	Intraoperative Transfusion Volume
30	Depression/Anxiety Disorders Medications	Oxytocin Primary Flag
31	Repeat C-Section Flag	Readmission Flag
32	Alternative Drug Flag	Ergonovine Primary Flag
33	Oxytocin Primary Flag	Oxytocin Pre-Cut Flag
34	Oxytocin Pre-Cut Flag	Indicated C-Section Flag
35	Intraoperative Transfusion Volume	Age
36	Allergies Medications	Carbetocin Primary Flag
37	Asthma Medications	Tranexamic Primary Flag
38	Tranexamic Primary Flag	Carbetocin Primary Cost (Estimated)
39	Ergonovine Primary Flag	ASA Score
40	Epilepsy/Seizure Disorders Medications	Carbetocin Primary Cost
41	Readmission Flag	Depression/Anxiety Disorders Medications
42	Postoperative Hemoglobin	Urinary Tract Infections Diagnosis
43	Emergent C-Section Flag	Asthma Medications

Appendix Tables 5: Iterative Backward Elimination for Real and Synthetic Dataset

Rank	Real Dataset Feature	Synthetic Dataset Feature
1	ASA Score	Campus
2	Anesthetic Technique	Carboprost Flag
3	Campus	Carboprost Primary Flag
4	Carbetocin Flag	Ergonovine Flag
5	Carbetocin Primary Flag	Gestational Age
6	Carboprost Flag	Hemoglobin Drop
7	Carboprost Primary Flag	Preoperative Hemoglobin
8	Ergonovine Flag	Intraoperative Transfusion Flag

9	Gestational Age	Emergent C-Section Flag
10	Hemoglobin Drop	Misoprostol Flag
11	Preoperative Hemoglobin	Revisit Emergency Department Flag
12	Intraoperative Transfusion Flag	Tranexamic Flag
13	Misoprostol Flag	Iron Deficiency Anemia Medications
14	Revisit Emergency Department Flag	-
15	Tranexamic Flag	-
16	Thyroid Disorders Medications	-
17	Iron Deficiency Anemia Medications	-

Appendix Tables 6A: Decision Tree Feature Importance for Real and Synthetic Datasets

Ranking	Real Feature	Synthetic Feature
1	ASA Score	Campus
2	Anesthetic Technique	Carboprost Flag
3	Campus	Carboprost Primary Flag
4	Carbetocin Flag	Ergonovine Flag
5	Carbetocin Primary Flag	Gestational Age
6	Carboprost Flag	Hemoglobin Drop
7	Carboprost Primary Flag	Preoperative Hemoglobin
8	Ergonovine Flag	Intraoperative Transfusion Flag
9	Gestational Age	Emergent C-Section Flag
10	Hemoglobin Drop	Misoprostol Flag
11	Preoperative Hemoglobin	Revisit Emergency Department Flag
12	Intraoperative Transfusion Flag	Tranexamic Acid Flag
13	Misoprostol Flag	Iron Deficiency Anemia Medication Flag

Appendix Tables 7A: Full Model Classification Report for Synthetic Dataset

Model	Metric	PPH Negative	PPH Positive	Accuracy	Macro Avg	Weighted Avg	AUC- ROC
Logistic Regression	Precision	0.76	0.79	0.77	0.77	0.77	0.81
	Recall	0.81	0.73		0.77	0.77	
	F1-Score	0.78	0.76		0.77	0.77	
Decision Tree	Precision	0.92	0.91	0.92	0.92	0.92	0.93
	Recall	0.92	0.92		0.92	0.92	
	F1-Score	0.92	0.92		0.92	0.92	
Random Forest	Precision	0.93	0.98	0.95	0.95	0.95	0.98

	Recall	0.98	0.92		0.95	0.95	
	F1-Score	0.95	0.95		0.95	0.95	
SVM	Precision	0.87	0.94	0.90	0.90	0.90	0.97
	Recall	0.95	0.85		0.90	0.90	
	F1-Score	0.91	0.89		0.90	0.90	
K-Means	Precision	0.45	0.46	0.46	0.45	0.45	-
	Recall	0.34	0.58		0.46	0.45	
	F1-Score	0.38	0.51		0.45	0.45	

Appendix Tables 8A: Full Model Classification Report for Real Data

Model	Metric	PPH Negative	PPH Positive	Accuracy	Macro Avg	Weighted Avg	AUC- ROC
Logistic Regression	Precision	0.74	0.73	0.73	0.73	0.73	0.85
	Recall	0.77	0.70		0.73	0.73	
	F1-Score	0.75	0.72		0.73	0.73	
Decision Tree	Precision	0.94	0.91	0.93	0.93	0.93	0.92
	Recall	0.91	0.94		0.93	0.93	
	F1-Score	0.93	0.92		0.93	0.93	
Random Forest	Precision	0.94	0.99	0.96	0.97	0.96	0.98
	Recall	0.99	0.94		0.96	0.96	
	F1-Score	0.97	0.96		0.96	0.96	
SVM	Precision	0.88	0.96	0.92	0.92	0.92	0.97
	Recall	0.97	0.86		0.91	0.92	
	F1-Score	0.92	0.91		0.92	0.92	
K-Means	Precision	0.45	0.41	0.43	0.43	0.43	-
	Recall	0.45	0.41		0.43	0.43	
	F1-Score	0.45	0.41		0.43	0.43	