Developing and Applying a Design Framework to Prepare Electronic Health Record Data for Time-Series Modeling

by

Sean R. Meyer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Design Science)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor Thomas Klumpner, Co-Chair
Assistant Professor Karandeep Singh, Co-Chair
Associate Professor Arvind Rao
Professor Nadine Sarter

Sean R. Meyer
seameyer@umich.edu
ORCID iD: 0000-0003-4360-2952

Table of Contents

List of Figures

List of Tables

Abstract

Well-prepared data for predictive modeling often yields better performance, but preparing data is time-consuming and requires domain expertise. This is especially true in early warning systems, or time-series prediction tasks, where a model is designed to issue predictions at regular intervals even though the underlying data elements are collected at irregular intervals. One general approach to addressing this challenge, which has been adopted by multiple disciplines, is to transform the unevenly collected data into regular intervals and then use data from each time step to predict the outcome in the following time step. Broadly, this is referred to as discrete time survival analysis in statistics; in machine learning, this is called sequence modeling; and in econometrics, this is termed autoregression.

Although much of the emphasis on time-series modeling in the literature is on algorithm selection, data preparation is an equally critical step because clinicians' ability to understand a model's predictions requires at minimum an understanding of the underlying predictors. Although several naive methods exist to convert irregularly collected data into regular data, such as carrying forward the most recent value, recent literature suggests that incorporating multiple summary statistics (e.g., mean, minimum, maximum) and expert-informed lookback periods (e.g., values from the past 6 hours) can help achieve state-of-the-art performance in early warning systems. Transforming data in this fashion is tedious and time-consuming but careful consideration of not only the variables' perceived relevance but also the relevant time points can lead to optimal model performance.

However, there is not a single correct way to prepare time-series data, and the process is heavily reliant on information known only to domain experts. Translating domain expert knowledge into code that can prepare data for modeling is both time-consuming and error-prone due to its highly manual nature. Even when modeling code is published accompanying scientific manuscripts in the literature, data preparation code is

often omitted or not applicable across institutions. This barrier, essentially a design problem, has held up progress in numerous clinical domains where early warning systems are otherwise widely adopted. One such area is maternal care, where simple rule-based early warning systems are widely adopted despite poor performance.

In this work, I first describe the current landscape of maternal early warning systems and survey commonly used data preparation tools in the R language ecosystem. I then examine the current state-of-the-art models for detecting postpartum hemorrhage, a complication consisting of excessive bleeding following childbirth. Finding state-of-the-art models to perform poorly, I describe a “grammar” that translates key design decisions into an R package named wizard, which is short for windowed summarization for autoregressive data preparation. I demonstrate how wizard can be used to replicate prediction tasks in a widely used de-identified dataset, MIMIC-III, which contains clinical data from over 40,000 patients admitted to critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Lastly, I apply wizard to the problem of postpartum hemorrhage, using domain expert knowledge to train models that outperform state-of-the-art models that do not take repeated measurements into account.

Chapter 1
Introduction

1.1 The Daunting Task of Data Preparation

Data preparation for developing clinical predictive models can consume a great deal of manual time and effort [1], [2]. How data are prepared can affect the patterns detected and subsequently model performance [3], [4], but are dependent on clinical application, knowledge of the underlying data, and modeling approach. This body of work takes a design science approach by first describing existing data preparation tools for time-series modeling, and then providing a new grammar to relate common components of time-series data preparation tasks to human-readable functions, akin to the grammar of graphics [5]–[7]. Datasets containing multiple variables and repeated measures are often ill-structured for prediction modeling and require transformation. Existing frameworks and preprocessing software have limitations when applied to clinical data.

In this body of work, I explore data science for clinical modeling from a design perspective. I evaluate existing tools and solutions intended for clinical data preparation and identify the need for a design interface layer for time-series data analysis. I introduce a design grammar, and a software implementation called wizard, to guide the co-development of structured data for modeling, leveraging expertise from clinicians and data scientists. This grammar captures commonly desired attributes in time-series modeling and uses them as human-readable parameters used to build time-based feature sets. I demonstrate how to use this grammar to engineer features using a well-known and widely used de-identified dataset. I then apply the grammar framework to the prediction of postpartum hemorrhage, a severe complication of childbirth, using domain knowledge from clinical experts in obstetrics and obstetric anesthesiology.

1.2 Area of Application

Time-series models are commonly used as components of early warning systems, in which alerts are generated in response to patients exceeding a specified level of risk. In clinical care, one prominent area in which early warning systems are widely used is in the care of patients admitted to the hospital for childbirth, a time during which life-threatening complications can occur. Despite the widespread use of early warning systems in maternal care, such systems generally rely on simple rules to generate alerts, a decision rooted in their historical calculation using pen and paper.

Recent research suggests that such rule-based systems have a low positive predictive value and thus may subject users to excessive alerts, which can lead to alert fatigue [8], [9]. As a result, maternal early warning systems have much to gain through application of a time-series predictive modeling approach.

With this as the primary rationale, prediction of postpartum hemorrhage, or excessive bleeding following childbirth, was selected as the area of focus for the data preparation design work described herein. Postpartum hemorrhage (PPH) is characterized by significant blood loss after childbirth. Maternal death attributed to PPH is among the most preventable causes of maternal mortality [10]. The cornerstone of high-quality care for PPH is early detection and timely intervention [11]. Predicting PPH more accurately than current state-of-the-art systems may be possible but will require preparation of temporal EHR data and application of predictive modeling approaches that consider several factors as compared to only a handful considered by current systems.

1.3 A Designed Approach

Working with clinical domain experts and scientific literature pertaining to PPH, we identify an initial set of clinically relevant predictors for training a model to predict PPH. We then categorize these sets of predictors into 3 categories: baseline predictors, time-varying predictors, and cumulative predictors. Through further elucidation from clinical domain experts, we identify key issues related to the timing of the predictors.

Using this information, we then develop an integration layer, which we term a grammar framework, to facilitate collaboration between an analytical expert who

understands the raw data and a clinician with domain expertise. This framework is intended to constrain the menu of the data preparation options and accelerate the process of data preparation. This is accomplished by supporting a shared understanding of the available data preparation options between data and clinical domain experts through a software package developed in the R programming language. Descriptors used to create structured modeling variables are transparent, and this framework supports reproducibility of methods across projects and institutions.

Additionally, the software package supports both small- and large-scale hardware. When computers have multiple cores and sufficient memory to hold multiple copies of that data, parallel processing can be used to speed up data analysis. In the presence of computational constraints, particularly limited memory, our software package supports chunking to partition the raw data into smaller chunks that can be processed serially. Additionally, chunking can be combined with parallel processing when the number of computational cores outpace the requisite memory required to hold multiple copies of the data.

1.4 Contribution to Science

The key contribution is the development of a design framework, referred to as wizard and implemented in an R software package, that facilitates time-series data preparation using a standardized grammar. This makes predictor descriptors more easily transportable across institutions, allows for rapid prototyping and reproducibility, and can reduce seemingly complex predictor descriptions to a simple set of options, easily understood by collaborating domain experts.

I demonstrate the use of this framework on a de-identified dataset to show that it can achieve similar performance to existing frameworks using very little code and clear predictor definitions while handling a large number of predictors. I then apply this to inpatient maternal care to evaluate how much state-of-the-art models can be improved upon. The resulting model outperforms contemporary early warning systems using only EHR data after data preparation performed with input from clinical domain experts using the wizard software package.

1.5 Overview

In Chapter 2, I describe the relevant background for this project. Complexities in clinical decision-making contribute to the cognitive burden faced by clinicians. Health information systems have emerged to reduce the cognitive burden by distilling large volumes of data into clinically actionable information. Maternal care at a hospital is a unique experience because it represents one of the rare times that healthy adults are admitted for around-the-clock care. Because constant monitoring is required, maternal early warning systems have proliferated as a safety net to alert clinicians of potential adverse events. These alert systems largely rely on rule-based thresholds that only consider the most recent data elements, which can cause them to generate alerts in the presence of a single abnormal value. These systems are often highly sensitive, meaning that they can detect patients who will go on to experience a bleed, but not highly specific, leading to false alarms and consequently alert fatigue. Clinical prediction modeling is a subfield of machine learning that focuses on developing models that synthesize multiple data elements to predict clinical outcomes. Following expert-informed data preparation, the application of clinical prediction modeling to maternal care may yield better models.

Chapters 3, 4, and 5 are standalone scientific manuscripts that describe different aspects of this project. Thus, each contains an introduction that expands upon the information in this introduction (Chapter 1) and the background (Chapter 2). Each also includes a list of co-authors who assisted with revisions and will be listed as co-authors on resulting manuscripts.

In Chapter 3, I evaluate state-of-the-art models from the literature for predicting postpartum hemorrhage to serve as a benchmark for subsequent work. These models are intended to predict postpartum hemorrhage at a single point in time, when a patient is first admitted to the labor and delivery ward. We find that these state-of-the-art models, which performed well in clinical data from a decade ago, do not work well in our contemporary patient population at Michigan Medicine.

In Chapter 4, I describe the motivation and development of a novel software package written in the R programming language and known as wizard, short for windowed summarization for autoregressive data preparation. Retrospective clinical data is often collected at irregular intervals, contains implicit missingness, and different types of

variables. Transforming these data into features while capturing changes over time can be time-consuming and complex. We introduce a framework to parameterize a shared understanding between clinical experts and analysts. This framework uses easily understandable verbs to transform database data into a data structure ready for algorithm application such as statistical and machine learning methods. We apply this framework to a deidentified public clinical dataset, MIMIC-III, to evaluate performance.

In Chapter 5, I build on recent work in the area of predicting postpartum hemorrhage to develop early warning system models. I use the wizard data grammar framework from Chapter 4 to build a dataset where each row captures a patient's state in twenty-minute time steps continuously over each patient encounter. Informed by domain experts, we apply a gradient boosting machine algorithm to predict postpartum hemorrhage throughout each patient encounter, finding that postpartum hemorrhage can be predicted more accurately when the model can incorporate information from after delivery as it becomes available.

The discussion (Chapter 6) summarizes our overall findings, the contribution to science, the strengths and limitations of our approach, as well as potential future directions for the development and application of the wizard software package.

Chapter 2
Background

2.1 Clinical Early Warning Systems

Early warning systems (EWS) are used to predict the risk of an adverse outcome for an individual at multiple time points during an at-risk period. When early warning systems use clinical prediction models to generate these predictions, we refer to these as EWS models. EWS models stand to be beneficial in clinical settings where the outcome is missed by clinicians due to either having a rapid onset, such as when patients are not under direct evaluation by a clinician when the outcome occurs, or slow onset where clinicians may not recognize a gradual decline in clinical status. EWS models have shown good model performance in predicting sepsis [12]–[21], acute kidney injury [22]–[26], and other acute conditions that are challenging to identify in a timely manner [27], [28].

Despite these advances in EWS modeling approaches for identifying clinical deterioration and organ failure, most have not been translated to maternal care. Though many early warning systems may be expected to be collinear to some degree—failure of one bodily organ (e.g., liver) often is a risk factor for failure of others (e.g., heart and kidneys)—pregnancy is a unique physiological state in which the usual clinical cues do not apply. Outside of elective surgery, delivery of a baby is one of the few times a generally healthy adult arrives at the hospital for inpatient care. Pregnancy-related complications include hemorrhage, sepsis, pre-eclampsia, and eclampsia, which are serious though rarely lead to death. Because of physiological changes related to pregnancy, pregnant women on average have a higher resting heart rate, lower blood pressure, and a higher respiratory rate as compared to non-pregnant adults [29]. Thus, vital signs and laboratory results that may be considered abnormal in non-pregnant patients may represent normal physiology in pregnancy.

The primary reason that advances in other clinical areas have not been translated to maternal care is that the development of EWS models requires complex, time-consuming data preparation work that is not easy to transport between clinical domains. Unlike computer code used to train predictive models, which is often published alongside scientific manuscripts, code used to prepare data is rarely shared. The lack of widely available data preparation code has been identified as a threat to the reproducibility of artificial intelligence-driven diagnostic systems more generally [30]. Perhaps more importantly, lack of this code means that each EWS modeling project must rely on custom code, which means that the amount of work required by the programmer to represent a given variable may substantially impact the way in which variables are ultimately represented.

2.2 Early Warning Systems in Maternal Care

Because of the high worldwide incidence of postpartum (after delivery) complications attributed to delayed recognition, maternal care has been a major focus area for early warning systems over the past decade. The National Partnership for Maternal Safety (NPMS, based in the United States) and the Confidential Enquiry into Maternal and Child Health (CEMACH, based in the United Kingdom) have helped facilitate the early detection of maternal morbidity through their endorsement of early warning systems. In 2012, CEMACH endorsed a modified early obstetric warning system (MEOWS) [31] and in 2014, the NPMS recommended the adoption of the Maternal Early Warning Criteria (MEWC) [32]. MEOWS and MEWC both consist of simple rules that flag abnormalities in the underlying clinical variables. These variables include vital signs, pain assessment, physical exam findings, and laboratory values (Table 2.1). Abnormalities in any of these components are considered an early warning of postpartum complications and thus represent a stopping point at which a patient must be assessed by a clinician (Figure 2.1).

The simplicity of these EWS criteria has advantages. The criteria are readily interpretable by clinicians and relatively easy to implement within the electronic health record (EHR). In 2018, the University of Michigan implemented an automated notification system based on the MEWC in response to the national guidelines [9]. Based on the

finding that the original MEWC criteria resulted in an alert every 9 minutes [8], which is untenable in terms of the workload imposed on clinical care, the original criteria were modified and implemented using a technology named AlertWatch Obstetrics (AWOB). However, even in the AWOB criteria, the underlying notion remains; an alert is triggered by a single sustained abnormal value. As a result, the simplicity of criteria-based early warning systems comes at the cost of a low positive predictive value, where many of the alerts represent false positives. Despite threshold adjustments to reduce frequent alerts, these models still demonstrate low positive predictive values. Excessive alarms are known to cause fatigue, increased workload, and can contribute to ignoring alarms altogether [33]. This added workload without the added benefit can cause mistrust in automated systems.

2.3 A Comparison of Existing Rule-Based Maternal Early Warning Systems

Maternal early warning systems were originally implemented using clinical rules to identify patients in need of further clinical evaluation. A modified early obstetric warning system (MEOWS) was one of the first implementations of a maternal early warning system [31], which was adopted by institutions in a form requiring scores to be calculated using pen and paper. This was followed by the maternal early warning criteria (MEWC) [34], which was introduced as a simpler successor to MEOWS as a bedside calculation. Instead of requiring a composite calculation, MEWC simplified the criteria to trigger evaluation by meeting only one of several criteria. Recognizing computerized tools with more complex criteria could be implemented without the need for clinicians to initiate them, in 2018 AlertWatch Obstetrics (AWOB) was designed and implemented at the University of Michigan to notify clinicians immediately if the predefined criteria were met [9]. Considered one of the first of its kind, it coupled the afferent ability of detection with the efferent ability for notification using in-house, pre-existing pager systems to notify clinicians of high-risk patients in real-time. While data are limited to support the widespread implementation of electronic early warning systems, prior works suggests they may play a role in maternal care [35].

Differences between input variables across maternal EWSs are shown in Table 2.1. MEOWS used measures readily available in a nursing assessment including respiratory rate, blood oxygen level, temperature, blood pressure, heart rate, pain assessment, and neurological status. MEWC then introduced oliguria (abnormal urine output volume) and removed temperature and pain score. AWOB used a combination of the measurements used in the previous two systems but introduced lab results and shock index.

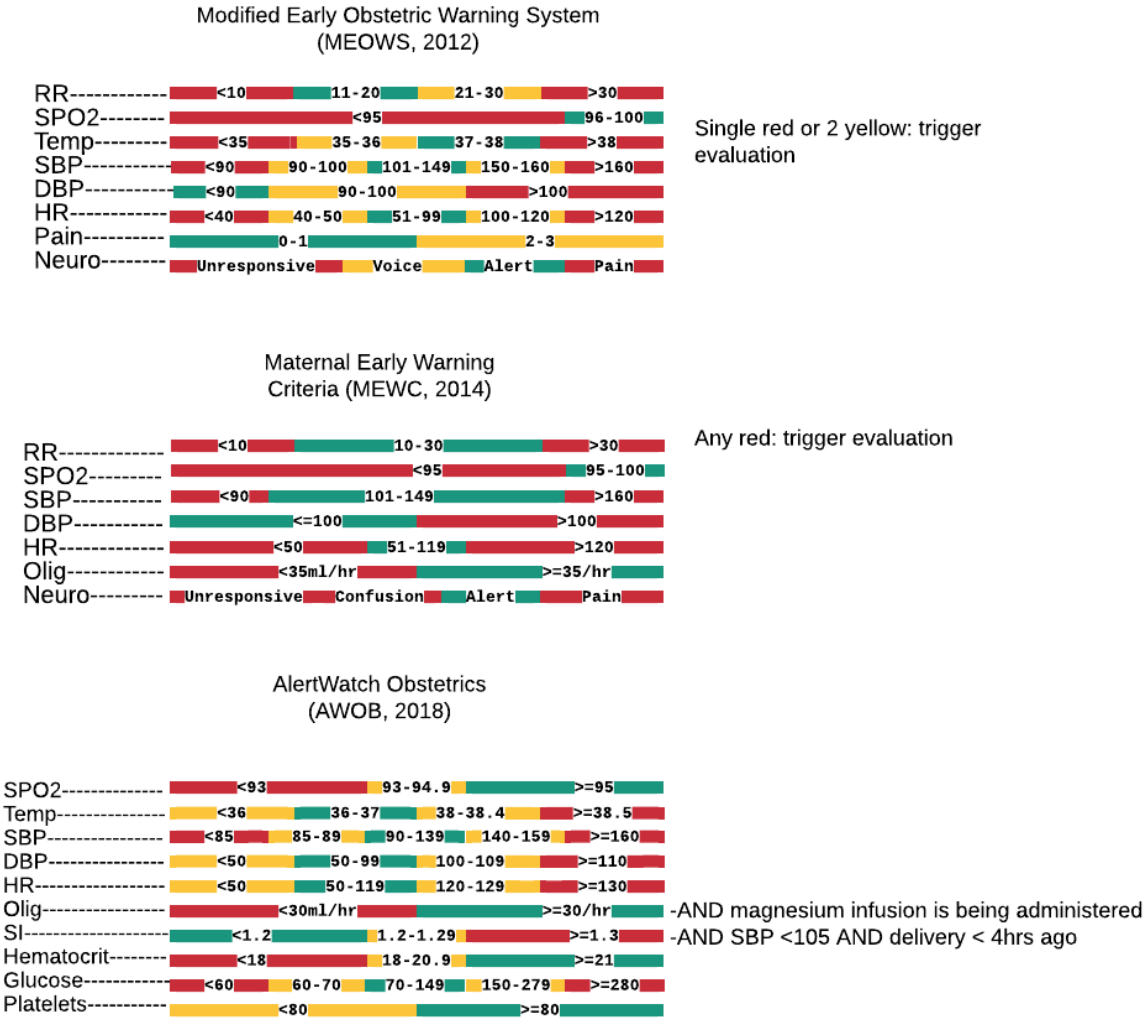
Table 2.1: Comparison of input variables

	MEOWS, 2012	MEWC, 2014	AWOB, 2018
Respiratory Rate	X	X	X
SPO2	X	X	X
Temperature	X		X
Systolic BP	X	X	X
Diastolic BP	X	X	X
Heart Rate	X	X	X
Pain Score	X		
Neurologic Status	X	X	
Oliguria		X	X
Shock Index			X
Hematocrit			X
Glucose			X
Platelets			X

Figure 2.1 shows the ranges used for each of the systems and how they identify increased risk. MEOWS uses a three tier color-coded system for most of the variables. A single red marker or two yellow markers trigger evaluation. MEWC thresholds were simplified to using only a two-color coding so any individual measurement could trigger evaluation if found outside the ideal range. AWOB, because it was automated, was able to reintroduce additional complexities, with a three-color system but adjusting the individual thresholds due to an increase in triggers [9]. Conditional logic was introduced

to add requirements of magnesium infusion for oliguria triggers. Additionally, a shock index of ≥ 1.3 accompanied by low systolic blood pressure would trigger an alert. These systems show a progression from complex to simple, and then complex again as the infrastructure matured and enabled automation. While these systems have been generally well-accepted, they leave room for improvement when compared against advances in early warning systems found in other inpatient areas due to their low specificity.

Figure 2.1: Maternal EWS variable thresholds



Note: RR = respiratory rate, SPO2 = pulse oxygen saturation, Temp = body temperature, SBP = systolic blood pressure, DBP = diastolic blood pressure, HR = heart rate, pain = pain score, Neuro = neuropathy, Olig = oliguria, SI = shock index

A drawback of criteria-based alerting systems is the lack of consideration for differences in physiology between patients. Differences in baseline vital signs and laboratory values means that a heart rate that would be considered abnormal for one patient may be within the expected range for another patient. These thresholds, selected through expert consensus in the field, also suffer from low positive predictive values. In comparison, more advanced mathematical models may perform better if the training data can be constructed effectively and with input from clinical domain experts.

2.4 Data Preparation Methods

In recent years, researchers developing maternal early warning systems have focused their attention on identifying better predictors of maternal morbidity with mathematical modeling techniques [36]–[38] while incorporating the automated notification of risk to care providers [9]. Postpartum hemorrhage represents one of the most preventable forms of maternal morbidity and mortality. An increase in data collection has made more complex approaches to prediction modeling feasible in maternal care. Preparing data to develop prediction models is a time-consuming but necessary step in prediction modeling [39], [40], but well-prepared data can often lead to better results [41], [42]. Despite this, standardized implementations of data preparation methods are not widely available [43].

Clinical data commonly contain unevenly spaced data, repeated measurements, and sparse data. These characteristics require special attention to data preparation when applied to early warning systems. While the transformed representation of predictors largely depends on domain expertise, outcome relevance, and algorithm requirements, authors who publish their transformation methods describe a broad range of methods [13], [14], [22], [23], [25]–[28].

Several factors need to be considered when preparing clinical data. In EWS models, domain experts must determine when to start and stop making predictions relative to the at-risk period. They must also determine at what interval to make predictions based on the available infrastructure (e.g., how frequently does the infrastructure allow scores to be recalculated on eligible patients) as well as the lead time required for timely intervention.

The relevant amount of time in the past, or lookback period, must be considered as well as how to break this lookback period into smaller sections using windows to describe smaller periods of time. For example, Tomasev et al. captured data up to five years in the past but split it up into 48 hours, 6 months, and > 6 months [22], while Lee et al. identified 8 significant points throughout the hospitalization to serve as markers for splitting windows of time [26].

Window transformations, when reported, can also differ greatly from model to model and are commonly used when there are repeated measures within one or more windows. For example, Delahanty et al. chose to capture the first, last, mean, minimum, max and trajectory for each window [14], while Koyner et al. used maximum values for some variables and number of occurrences within each window for others [25].

When preparing data, implicit missingness often becomes explicit, which depending on the algorithm or outcome, may require application of imputation methods. For example, Delahanty et al. imputed all missing numeric measures with -9999 to indicate they were missing [14]; this approach, while problematic in regression models, can work in tree-based models because it introduces the possibility of splitting criteria that introduce missing values on one side and non-missing values on the other. Harutyunyan et al. on the other hand chose “normal” baseline values [27], which makes the assumption that if a value is not recorded it must be normal. Koyner, Deist, and Lee chose to impute medians for continuous measures and modes for categorical variables [25], [26], [28]. Additionally, Lee et al. specifically designated variables that were 1-5% missing to be hot-deck imputed, a process that selects random values from existing values [26].

Based on the outcome, domain experts must determine a lookahead period, or amount of time into the future to evaluate for a specific event. Tomasev chose 48 hours [22], while Koyner chose 12 hours ahead [25], Mohamadlou chose 12-72 hours [23], all to detect acute kidney injury (AKI) while Downing chose 48 hours [12], Nemati chose between 4 and 12 hours [13], and Delahanty chose 1 to 24 hours [14] for suspected infection. This broad range of lookahead values is indicative of the imprecise nature of such a determination as well as differences across clinical outcomes.

There are many data design decisions involved in preparing data for modeling. While there are similarities in approaches, many methods used remain unpublished or highly specific to variables used in model development.

2.5 Maternal Care Machine Learning Models

There have been two large-scale studies published recently to predict adverse outcomes in maternal care. Venkatesh et al. developed multiple models to predict postpartum hemorrhage with EHR data available at the time of admission [36]. This model was developed using a de-identified dataset with high discriminative performance of up to 0.93. The other large-scale study authored by Escobar et al. focused not only on postpartum hemorrhage but multiple maternal adverse events [37]. However, it failed to identify patients with postpartum hemorrhage. These studies lend to the notion that training more sophisticated models should be possible.

2.6 Data Preparation Tools

Some major challenges in developing machine learning models are the inability to readily prepare data for modern early warning systems. As described in Section 2.4, authors developing machine learning models must reinvent the wheel for each institution to prepare EHR data. As described further in Section 4.5, existing time-based data preparation tools in R fall along a spectrum of two extremes; low-level feature extraction tools that focus on representing time-series or lagged variables and forecasting packages that primarily aim to model trends in a single variable over time (e.g. stock market) rather than separate predictions for separate patients.

2.7 Algorithms

There are many algorithms to choose from when developing a prediction model using binary outcomes. An algorithm is a set of instructions which generate a set of rules based on data processed through it. They can vary from simple statistical algorithms like logistic regression to more complex algorithms like neural networks, each having their own advantages and disadvantages.

2.7.1 Linear Models

Logistic regression [44], which is the most commonly used statistical algorithm is an extension of the generalized linear model (GLM). Like linear regression, an outcome is fitted to a linear combination of predictors, except with logistic regression a logit function restricts predictions to between zero and one. Coefficients for the predictors can then be interpreted as the difference in predicted values--either log-odds for raw coefficients or odds ratios for exponentiated coefficients--for each unit change in the predictor variable. The model predictions are calculated in terms of the probability of the outcome occurring.

Variations of logistic regression include ridge and lasso regression. They both use penalization methods to adjust coefficients of the model. Ridge, also known as ridge regression, reduces collinearity of variables and variables that do not have an association with the outcome by “shrinking” coefficients. Coefficients for collinear variables share the weight as they are penalized, and coefficients approach zero as more penalty is applied. Lasso (least absolute shrinkage and selection operator) regression takes a different approach to this problem. While coefficients in ridge regression can approach zero, lasso can actually set them to zero as more penalty is applied. By zeroing out coefficients and thus eliminating certain variables, lasso regression serves as a variable selection method on top of being a prediction method. Lasso commonly handles collinearity of variables by selecting a single variable within collinear group of variables and setting the others’ coefficients close to or at zero. Thus, a major difference between these methods with respect to collinear variables is that ridge regression places weight more equally on collinear variables and less weight on any given variable, while in lasso regression, coefficients are disproportionately assigned to a subset of collinear variables while eliminating others.

Support vector machines, when applied to binary outcomes, are closely related to logistic regression and classify outcomes by dividing a dataset into two classes using a hyperplane [45]. A hyperplane is a multidimensional line that maximizes the margin between training points in each class and the hyperplane. While SVM is related to regression, it is often extended through the use of kernel transformations (a set of mathematical functions) to form non-linear boundaries between classes.

2.7.2 Non-linear Models

K-nearest neighbors is a clustering algorithm designed to classify observations into groups [46]. It works by assigning a class to unclassified points based on the proximity (defined in a few different ways) to points which are already classified. Based on a number of neighboring points, K, a majority vote based on already classified points are calculated which determine the class of an unclassified points. This can be computationally expensive to calculate as the numbers of observations and dimensionality increases.

The rule-based maternal early warning systems in Section 2.3 are in fact decision trees applied on only a few predictors and determined by expert consensus, however, tree-based algorithms can be used to build a model through a process where optimal variables and splits are selected using computational criteria (e.g., Gini gain). Tree-based algorithms are essentially a chain of if-else statements strung together as rules. Each rule split is chosen based on cut-offs which produce maximum separation between subgroups and minimum variability with consideration of the outcome. Variables with the largest separation are at the top followed by others in order of the most separation. The algorithm stops when subgroups reach a minimum size or there is no further improvement in performance following additional splits. Each terminal node uses the distribution of outcomes in the training set to determine predictions. Trees are prone to overfitting so a process known as cost-complexity tuning prunes branches of the tree by using tree complexity as a penalization measure [47].

A random forest algorithm is an extension of decision trees in that it uses an ensemble of decision trees to build more robust models less prone to overfitting compared to simple decision trees [48], [49]. The random forest algorithm works by created many decision trees using a random subset of predictors on bootstraps (i.e., random samples with replacement) of the data. Predictions are then averaged to provide a final summative prediction.

While random forest builds a series of independent trees, boosting machines build trees that are trained in an interdependent manner. Trees are built in a successive order, applying more weight on observations which were difficult to classify and reducing weight on observations which were easy to classify forcing the algorithm to focus on them in subsequent trees [50]–[52]**.** Boosting trees do not require imputation; instead they select

decision splits based on the branch that most closely resembles an existing value using other variable splits. Hyperparameters provide flexibility for model tuning but can heavily influence performance, which often makes a large grid search necessary for optimal performance.

Neural networks produce a nonlinear model using intermediary hidden layers expressed as linear combinations of the original predictors (or the previous layers if there are multiple hidden layers). Neural network without hidden layers are simply logistic models [53]; thus, the hidden layers are responsible for representing non-linearity through modeling of interactions between variables. Similar to gradient boosting machines, neural networks can progressively learn from poor predictions. However, instead of using boosting to accomplish this, neural networks rely on backpropagation to progressively learn from errors.

Which modeling approach to choose can depend on the outcome prevalence, the underlying data, and the level of complexity required. Linear methods may be preferred to prevent overfitting when the number of observations is inadequate to use a non-linear approach. Non-linear models may be preferred when variables interact with one another or may not be expected to have a linear relationship. For example, as age increases, patients may tend to have more comorbidities, and this can be accounted using methods that implicitly handle interactions or by explicitly including the interactions in a linear model. Similarly, considering human anatomy as a system of systems, one organ system may impact another in a crescendo of events. For example, when blood pressure drops, heart rate often rises to compensate. A common practice is to use multiple modeling approaches to determine what the optimal performance would be and choosing the model with the right mix of performance and transparency. In a broad application of algorithms on a large number of datasets, boosted trees, random forests, bagged trees, support vector machines, and neural nets outperformed many other algorithms used [54]. When specifically looking at clinical data, support vector machines were the most commonly used algorithm, while random forest was found to be best performing across the studies which published the results using more than one algorithm [55].

Chapter 3
External Validation of Postpartum Hemorrhage Prediction Models Using Electronic Health Record Data

This chapter was co-authored with Alissa Carver, Hyeon Joo, Kartik K. Venkatesh, J. Eric Jelovsek, Thomas Klumpner, and Karandeep Singh.

3.1 Introduction

Postpartum hemorrhage (PPH) is the leading cause of severe maternal morbidity and among the most preventable causes of maternal death in the United States [10]. Delayed recognition and inadequate response to clinical warning signs contributes to a majority of these deaths [11]. Although many risk factors for PPH have been well-characterized [56], contemporary maternal early warning systems have generally relied upon simple rules (e.g., elevated heart rate) to identify women at risk of PPH. Such alerting systems have been shown to be highly nonspecific because alerts flagging patients with elevated heart rate, for example, can occur for reasons unrelated to PPH such as uncontrolled pain [8], [9]. Thus, future alerting systems will need to rely on prediction models that consider several contextual factors beyond individual vital signs. Recently, Venkatesh and colleagues developed models to predict PPH using the National Institute of Child Health and Human Development (NICHD) U.S. Consortium for Safe Labor (CSL) dataset [36]. Using a combination of statistical and machine learning methods, four models were trained to predict PPH upon admission, defined as estimated blood loss ≥1000 mL within 24 hours of delivery, using data from 152,279 deliveries between 2002 and 2008. The CSL models had high discriminatory ability, with C-statistics between 0.87 and 0.93 across the consortium. If these models prove to be effective in

contemporary settings when implemented within the electronic health record (EHR), they have the potential to reduce maternal morbidity when linked to targeted interventions.

Since the CSL models were developed, several changes related to the measurement of PPH have emerged that may affect their implementation. The prevalence of PPH has increased in the U.S. from 2010 to 2014, even after adjusting for known risk factors [10]. This mirrors similar increases in PPH prevalence from the 1990s [56], and taken together, the rising prevalence suggests improved recognition of PPH as well as a higher risk population of pregnant women. Spurred by recent recommendations by the American College of Obstetrics and Gynecology that recognize quantitative blood loss (QBL) as a more accurate method for measuring blood loss, the rising adoption of QBL has led to much higher estimates of PPH [11], [57], [58].

QBL methods may classify milder cases as PPH that would previously not have met criteria, and these differences in measurement may adversely affect the performance of models that were originally developed to predict EBL when applied to a setting using QBL. The change in model performance due to changes in case-mix and outcome definitions has been broadly termed as “dataset shift” [59]. Dataset shift can occur for a multitude of reasons [60], and its occurrence can lead to lower performance during external validation [61]. This has been recognized as a threat to the deployment of machine learning models in a recent Food and Drug Administration action plan [62]. Our objective was to externally validate the CSL models in a contemporary setting using predictors derived from the EHR and PPH measured by QBL methods. Due to known differences between the CSL data and our study cohort, we further compared the original CSL models against models that were refit using our EHR data.

3.2 Methods

3.2.1 Study Cohort

The University of Michigan Von Voigtlander Women’s Hospital is a tertiary care academic women’s hospital with approximately 4,600 deliveries per year. Our study cohort included women aged 18 or older who delivered an infant between February 1, 2019 and May 11, 2020 (Figure 3.1). Deliveries were excluded if the estimated gestational

age at birth was less than 22 weeks, if quantitative blood loss data was not documented, or if no prenatal records were available. Women without prenatal records were excluded because the model would have insufficient information to make accurate predictions.

EHR data collected for this study included delivery information, maternal characteristics, medication administrations, vital signs data, and laboratory results. Data were collected for the time period between the estimated date of conception and the first collection of vital signs on the labor and delivery unit. This was to ensure sufficient time for conditions present on admission to be observable in the EHR. Diagnostic data, including comorbidity data, was identified using International Classification of Diseases-10 (ICD-10) codes, which replaced ICD-9 codes at the University of Michigan in 2015.

Postpartum hemorrhage was defined as the documentation of QBL of ≥ 1000 mL in the 24 hours following delivery [63]. In February 2019, the University of Michigan implemented a protocol to quantify blood loss routinely for all deliveries. This study was approved by the University of Michigan Institutional Review Board (IRB), which waived the requirement for informed consent.

3.2.2 Model Validation

We used a three-step process to externally validate the original CSL model in our dataset [61]. We first evaluated the extent to which the original development cohort was related to our study cohort. Second, we evaluated the performance of the original CSL models in our cohort. And third, we refit models using the CSL variables in our study cohort to evaluate the extent to which the models could be improved in our cohort.

Step 1. Assessing the relatedness of the CSL cohort and our study cohort

We evaluated the relatedness of the cohorts qualitatively by comparing the distribution of predictors and outcomes. Although quantitative methods have been proposed to compare cohorts, these rely on having access to both cohorts [61]. Because we did not have access to the development cohort, we opted to compare cohorts qualitatively.

Step 2. External validation of the CSL models in our study cohort

Details regarding the original CSL prediction models and the CSL cohort have been published previously [36], [64]. We mapped the 55 predictors from Appendix 5 used in the original CSL prediction models to our EHR data [36]. This mapping process was performed in collaboration with the original study authors to ensure high fidelity of our predictors with the original CSL model definitions. To examine the plausibility of our mappings, we further compared the observed risk factor prevalence derived from our EHR data with published national estimates. Our final variables and how they were identified within our data set can be found in Supplemental Table 3.8.

We assessed discrimination using the C-statistic, which estimates the probability with which a model correctly distinguishes between higher and lower risk patients [65]. We qualitatively evaluated model calibration by comparing a loess curve comparing continuous predicted probabilities to the observed risks. A well calibrated model is one that does not over- or underestimate risk across the full range of predictions [66], [67].

Step 3. Refitting CSL models using our study cohort

To evaluate the extent to which the differences in the CSL development cohort and our study cohort could affect model performance, we compared the original CSL models against refit models using our study data. We used the same algorithms used in the CSL models: logistic regression with and without lasso penalty, random forests (RF), and gradient boosting machines (GBM). To evaluate the out-of-sample performance of the refit models, we used 10-fold cross-validation. The dataset was randomly divided into 10 folds. A model was developed using 90% (9 out of 10 folds) of the data and evaluated in the remaining 10% (1 out of 10 “evaluation folds”). This process was carried out 10 times until each of the folds was used for evaluation. In aggregate, the performance in the evaluation folds was used to estimate out-of-sample performance for the refit models. We assessed both model discrimination and calibration for the refit models.

3.2.3 Missing Data

Bagged tree models were used to non-parametrically impute missing values for continuous variables [68]. Education, which was a required predictor in the CSL models

but missing in a large proportion of our patients, was imputed with the median value. Binary variables, such as the presence or absence of diagnoses during pregnancy, were considered absent if not documented in the EHR.

3.2.4 Sensitivity Analyses

Our academic hospital's incidence of PPH, assessed by QBL, is higher than that reported in other cohorts, which could affect the performance of the CSL models. To evaluate the impact of differing thresholds on the prediction of PPH, we carried out sensitivity analyses that considered alternate definitions of PPH based on thresholds of 1500 mL, 2000 mL, and 2500 mL of QBL.

3.2.5 Statistical Software

R 4.0 was used to conduct all analyses [69]. We used the h2o package (v3.30) for the refit prediction models [70]. Discrimination, calibration, and threshold performance were visualized using the runway package [71]. C-statistic confidence intervals and comparisons were calculated with bootstrapping (1,000 replicates for 95% CI and 2,000 replicates for comparison) using the pROC package [72]. Our code is publicly available on GitHub [73].

3.3 Rationale for Study Design

Electronic health record data collected at unevenly intervals often contains repeated measurements, duplicate sources, and vague predictor descriptions (e.g., "baseline"). This can lead to ambiguity in descriptions of the data preparation techniques and requires clarification when replicating findings. This section describes study design decisions, in a detailed narrative form, from data acquisition through modeling intended to provide a thorough understanding of why specific choices were made around study design when a subjective lens was required for task completion.

3.3.1 Data Acquisition

Data was sourced from the Research Data Warehouse (RDW) hosted by Michigan Medicine using Structured Query Language (SQL) through both Microsoft Database

Management Studio and R packages DBI and dbplyr, which provide support for Open Database Connectivity (ODBC) connections within R and thus allow most of the data preparation code to remain within the R code repository.

The cohort date range was selected based on the implementation of quantitative blood loss (February 2019) and ended based on the current data available at the time minus thirty days to account for billing delays (May 2020). Hospitalizations were filtered based on the requirement that a patient's age was 18 or older and at least one measurement of QBL was captured. This decision was made on an assumption drawing from knowledge from clinical domain experts that even a normal delivery is accompanied by some degree of blood loss. Thus, if no QBL was recorded in the chart, this was most likely because the person involved in the delivery may not have been trained to calculate and record QBL, such as a midwife, rather than being a marker of insignificant blood loss. Hospitalizations were also removed if there were no data recorded between admission and one year prior because if there were no prenatal records for a patient, the model would have had insufficient information to make accurate predictions at the time of admission. Finally, hospitalizations were excluded if the gestational age of the fetus was less than 22 weeks, the minimum gestational age at which neonatal resuscitation is offered at our institution.

Based on input from a database analyst and clinical experts on the research team, candidate variables from database tables were identified. In some cases, there were duplicate sources for some variables (e.g. Group B Streptococcus [GBS] colonization was present in both diagnoses and lab results). Some variables of interest were extracted from more than one source (e.g. GBS colonization from both diagnoses and labs, vitals from both flowsheets and Physiobank). These variables were selected, and in some cases combined, based on the guidance of clinical experts. For example, vital signs are first collected in Physiobank and only become accessible from within electronic health record flowsheets after verified by a nurse. In a review, prior cases of women experiencing severe postpartum hemorrhage at our institution, abnormal vital signs were often not recorded in flowsheets because adverse events occurred shortly after the collection of vital signs (in Physiobank) but before they could be verified (in flowsheets). Thus, we opted to use vital signs from Physiobank to give models timely data, even if it carried the

potential to be unreliable because it had not been verified by a nurse. Database tables that were identified to contain variables needed for predictor development were extracted for the cohort specified above, including all data from one year prior to admission up until discharge. Tables retrieved included demographics, encounters, insurance, BMI, perioperative case times, diagnoses, clarity social history, lab results, medication administrations, procedures, clarity flowsheets, Physiobank, Charlson comorbidities, and Elixhauser comorbidities.

3.3.2 Preparation

Patients can either arrive at the hospital experiencing signs of spontaneous labor or they can be scheduled for induction, usually in the early evening. In our research database, arrival time, which is the variable closest to the beginning of the hospitalization, could be prior to the admission time if patients go through triage (e.g., in spontaneous labor) or are checked in early for induction, or it could be following the admission time if patients are late for their induction. To be sure the model would capture sufficient information to make accurate predictions, we opted to capture the first vitals which occurred after the earlier of two time points (i.e., either arrival to the hospital or admission).

The dichotomous outcome was calculated by summing QBL from delivery to 24 hours after delivery, also known as primary postpartum hemorrhage. In addition to the primary outcome ≥ 1000mL, bleeding of ≥ 1500, 2000, and 2500mL were calculated as part of a sensitivity analysis (Section 3.2.4). Because the proportion of patients experiencing postpartum hemorrhage at our institution is substantially higher than other public reports of PPH, one likely explanation is that our use of QBL overestimates PPH (or other reports underestimate PPH). To account for potential systematic overreporting of PPH in our data (or systematic underreporting in other sources), we evaluated the PPH models using a higher threshold of PPH to see if the models could reliably identify more extreme cases of PPH.

After the initial predictor mappings were established with our core team, we contacted the authors of the Venkatesh *et al.* study to obtain access to the original models and to check our variable mappings. Given the use of clinical descriptions to describe the temporal aspects variables (e.g., “baseline”), we additionally required clarification of the

timespan for variables included in the model. Based on conversations with the original study team, we learned that variables in the original cohort from which the models were constructed had a flag indicating whether a value was “available upon admission.” Any value containing this flag was considered eligible for inclusion in the model. However, many of these values were manually abstracted using information from the clinical documentation, up to and including the “History and Physical,” an intake note recorded by the obstetrician. Because the documentation may occur several hours after a laboring patient is admitted to the hospital, information may be considered as “available upon admission” even if it actually became available *after* admission. This is in contrast to the mechanism we used to determine which values were eligible for inclusion, which was based on a comparison of timestamps from when information was recorded with the time of arrival to the hospital. Additionally, disclosure text available for the original study in which the models were developed mentions that there was no precise definition for any of the predictors used, suggesting that the predictor design methods were left up to the participating institutions across all 10 sites.

The core research team met every week for two years to discuss topics including cohort selection and exclusion criteria, variable to predictor mappings, and appropriate descriptions for the predictors used in the original study based on a combination of perceived clinical relevance to the outcome (PPH) at the point of admission and literature or organizations describing generally accepted definitions of variables such as reVITALize [63]. In cases of subjective or ill-defined terms (e.g., pre-eclampsia and pre-eclampsia without severe features), definitions were selected based on a consensus of clinical experts on the team. Cases in which there was no source data present was documented for each data table collected. We opted to exclude patients with missing prenatal records because models would not be expected to make accurate predictions when key data elements were absent.

The diagnoses table was split into two temporal groups: pre-pregnancy and from the time of pregnancy to admission. This was determined by the estimated date of gestation recorded in the database. In cases where the estimated gestational age was unknown, it was imputed to 280 days (40 weeks). This table was further divided into sources: active problem summary, resolved problem summary, deleted problem

summary, present-on-admission billing, billing, medical history, and visit diagnosis. Based on reliability factors determined by clinical expert consensus, criteria were determined for what was considered past medical history and present conditions. History consists of `Resolved Problem Summary`, `Medical History`, `Visit Diagnosis` prior to admission. Present (to pregnancy) conditions include `Active Problem Summary`, `Billing` between 22 weeks after the estimated date of conception (EDC) and admission or `Present on Admission` and recorded between admission and discharge. The latter was to match the original study as closely as possible despite a lag likely present between condition onset and observation. The diagnoses were then filtered based on the criteria present in Supplemental Table 3.7 for both ICD-10 codes, regular expression, or a combination of both, which were the result of clinical expert guidance to match labels present in the original study as closely as possible.

The BMI table was used to engineer both weight and BMI at both the pre-pregnancy an admission. Pre-pregnancy predictors were defined by estimated gestational age of less than 20 weeks and admission weight was defined as any weight collected between 7 days prior to admission and the delivery date. The median value of all variables was calculated when repeated measures were present.

The only variables used for this study from the labs table was the results of the Group B Streptococcus (GBS) test and hemoglobin. The last value prior to admission was used as a predictor for GBS colonization. A significant complication in capturing this predictor was the way it was collected in the database. Instead of a simple positive or negative value indicating its presence, it is a categorical portion followed by text within the same field. To accurately measure whether GBS colonization was present, regular expression was used to capture several different parts of the column value to confirm positive matches for the lab test. Hemoglobin was captured if present between arrival and delivery to ensure inclusion of any hospital intake values only relevant to the delivery admission. Both results from internal and external labs were used to capture this predictor. A hemoglobin of < 10 was later used in combination with diagnoses to capture anemia of a patient using two different data sources.

Antepartum hospitalizations, described as admissions prior to the delivery admission, were captured by looking at the encounter table and filtering to inpatient visits between the estimated gestation start and arrival time for delivery admission.

Medications used in the model were antenatal steroids and magnesium sulfate prior to when the first vitals were taken on a patient, which was a time point used to identify admission intake. Use of antenatal steroids was defined by capturing any administration of betamethasone or dexamethasone using regular expression of the medication name. Magnesium sulfate was captured in a similar way, however, the only text needed to describe this administration is ‘magnesium sulfate’.

Insurance was captured at the encounter level by collecting the primary insurance plan present prior to the first vital signs being collected.

Education was captured by looking at the last temporal value per patient prior to delivery. However, this variable was both highly missing, with only ~10% of patients having entries and was likely collected as a text field or a combination of sources given that some entries were ordinal numeric while others were ordinal text (e.g. sophomore). As this was a required variable in the original models, we needed to supply a value to validate the model. Only the numeric values were captured, and any missing values were imputed to the median under the assumption that anyone who hadn’t provided their education would be reflected as a normal value instead of an outlier.

Cases which were missing any values of the outcome (QBL), social history, diagnoses data, labs, medications, prior appointments, or all demographic data were excluded from the model. The rationale for this is that the model would have insufficient data to make accurate predictions.

Maternal GBS colonization was captured by combining both sources, using the ICD-10 diagnosis codes from the diagnoses table as well as a positive lab for GBS from the labs table within the timeframe described in Section 3.3.2.

Multiple gestation was captured using both the ICD-10 code for the encounter and gestational period found in the diagnoses table as well as data collected from Stork, the database used to capture maternal care data.

The recipes R package was used to impute values. An important concern of model-based imputation in this study was cross-contamination of information between models

because each algorithm was provided a different subset of predictors (Supplemental Table 3.8). To avoid predictors *not used* inadvertently leaking information into predictors which *were used*, each subset of predictors was handled independently. All numeric predictors were imputed using bagged trees [68] separately for all four data subsets.

There are generally two categories for imputation which can depend on the intended representation. Single value imputation has computational advantages (e.g. median, mode) but assumes that any values that are missing are normal, identical, and uncorrelated with other variables. Model-based imputation assumes that the missingness of each variable is partially systematic and depends on available information from other variables. Examples of model-based imputation methods are linear regression (e.g., as used in multivariate imputation with chained equations), K-nearest neighbors, decision trees, and ensembles of trees (using bagging or boosting).

Bootstrap aggregated trees, or bagged trees, are one type of tree ensemble used commonly for model-based imputation. Bootstrapping refers to the random selection with replacement of an observations equal to the number of total samples in a dataset. In an ensemble of bagged trees, each bootstrapped dataset is used fit a decision tree to impute missing values for each variable using other variables. Each tree is used to generate a prediction on new observations, and the final probability is averaged. Bagged trees generally only need about 25-50 trees per predictor to converge, which is why they are preferred over random forests [74]. Related to bagging, boosting is used to improve algorithm performance by increasing weights of observations that were difficult to classify and reducing weights for observations that were easy to classify [50], [75].

The original study did not dummy code categorical predictors (marital status, insurance, race, and site id). While this would not have a large impact on flexible models like random forest and gradient boosting machines because they can handle misordered ordinal variables, statistical models would interpret them as numeric values and fit a regression model inappropriately. For this reason, all categorical predictors were converted to numeric to match the original model but dummy coded in refit models to follow analytical standards of practice. Missing categorical values were assigned a new category of “unknown or other.”

Race differed slightly in our data from those described in Appendix 2 of the original study [36]. Value comparisons can be reviewed in Table 3.1. This data was taken from the demographics table, only the primary most recent entry prior to delivery was captured. In cases where race was missing, *Other* was imputed in its place. Seizure disorder and fetal macrosomia were based solely from ICD-10 diagnosis codes so were assumed to be negative if they were missing.

Table 3.1: Categorical data mappings

Variable	**Our Validation Study Data**	**Original Study**
Seizure Disorder	*Missing*	No
	Yes	Yes
	Not Assigned	Unknown
Fetal Macrosomia	*Missing*	No
	Yes	Yes
	Not Assigned	Unknown
Race	Caucasian	White
	African American	Black
	Not Assigned	Hispanic
	Asian, Other Pacific Islander	Asian/Pacific Islander
	Other, American Indian, Alaska native	Multi-racial
Insurance	Private	Private
	Medicaid, Medicare, Other Governmental Insurance	Public
	Workers Compensation	Self pay
	Other	Other
	Unknown	Unknown
Marital Status	Married	Married
	Not Assigned	Not Married; Divorced/Widowed
	Unmarried	Not Married; Single
	Missing	Unknown

Since the original models we were evaluating chose two statistical models (logistic and lasso regression) and two machine learning models (random forest and gradient boosting machines), our study used the same algorithms in our evaluation.

3.4 Results

We identified 6,153 deliveries during the study period, of which 5,261 deliveries met the inclusion criteria (Figure 3.1). In comparing the CSL model development cohort

with our validation cohort, we identified several differences. The rate of postpartum hemorrhage was much higher in our validation cohort: 25% (1,321/5,261) in the validation cohort as compared to 4.8% (7,279/152,279) in the original CSL cohort. Compared to the CSL population, our validation cohort was older, had more comorbidities including chronic hypertension and pregestational diabetes, and greater maternal weight differences both pre-pregnancy and on admission (Table 3.2). There were also similarities among the cohorts. In both, patients who experienced PPH were more likely to have used assisted reproductive technology, have placenta previa, and have multiple gestation (Supplemental Table 3.6).

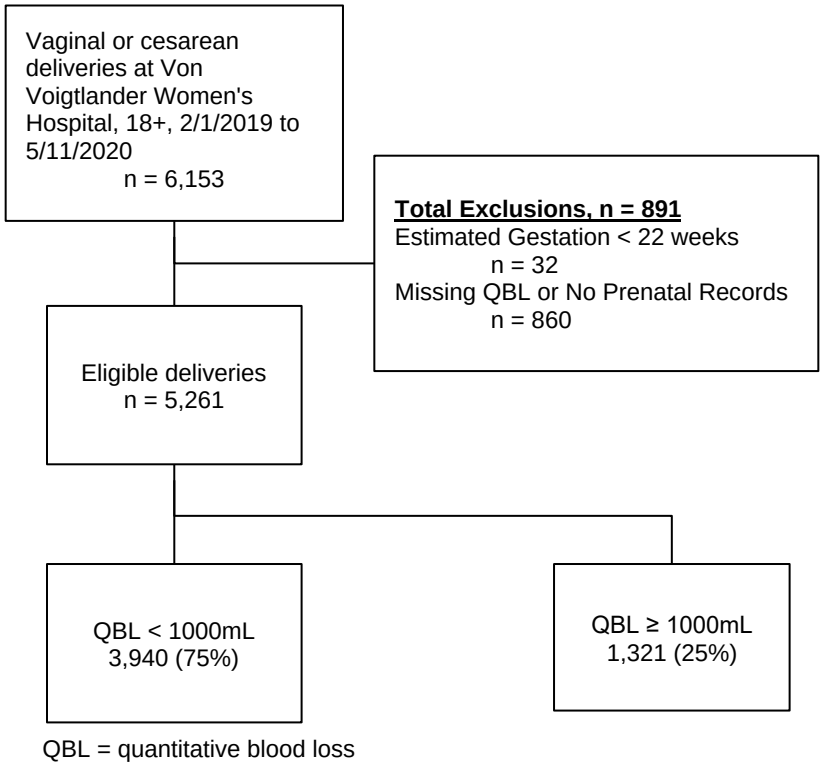


Figure 3.1: Cohort inclusion/exclusion criteria
Flow diagram of the inclusion and exclusion criteria applied to the study cohort. The electronic health record was queried to identify women aged 18 or older who delivered an infant. Exclusion criteria were then applied.

Table 3.2: Population cohort, stratified by data source

Characteristic	Consortium for Safe Labor Cohort, N = 228,438[a]	University of Michigan Validation Cohort, N = 5,261[a]
Age	29 (24, 30)	31 (27, 34)

Pre-pregnancy weight (kg[b])	63.5 (56.2, 75.8)	71 (61, 84)
Missing	67,294 (29.4%)	899 (17%)
Admission weight (kg[b])	79.3 (70.0, 91.1)	83 (73, 96)
Missing	35,984 (15.7%)	158 (3.0%)
Gravidity	Not Reported	2 (1, 3)
Missing	0 (0.0%)	1 (0.0%)
Parity	1 (0, 2)	1 (0, 1)
Missing	0 (0.0%)	1 (0.0%)
Gestational age	39w (38w, 40w)[c]	39w2d (38w2d, 40w3d)[c]
Missing	0 (0.0%)	19
Anemia	23,057 (10.4%)	812 (15%)
Missing	7,877 (3.4%)	0 (0.0%)
Assisted reproductive technology	1,101 (0.9%)	236 (4.5%)
Missing	107,479 (47.0%)	0 (0.0%)
Temperature	98.0 ± 0.81[d]	98.10 ± 0.50[d]
Missing	47,977 (21.0%)	1 (0.0%)
Cesarean delivery	65,990 (28.8%)	1,677 (32%)
Fetal macrosomia	1,858 (1.5%)	187 (3.6%)
Missing	102,247 (44.7%)	0 (0.0%)
Gestational diabetes	11,999 (5.2%)	423 (8.0%)
Gestational hypertension	6,286 (2.7)	547 (10%)
Multiple gestation	5053 (2.2%)	162 (3.1%)
Placenta previa	1,647 (0.7%)	226 (4.3%)
Prior cesarean delivery	31,321 (14.5%)	960 (18%)
Missing	13,219 (5.7%)	0 (0.0%)
Spontaneous labor	122,673 (53.7%)	107 (2.0%)
Systolic pressure	124.1 (14.89%)	125 (116, 134)
Missing	52,766 (23%)	0 (0.0%)
Trial of labor	192,074 (84%)	155 (2.9%)
Outcome: postpartum hemorrhage	7,279 (4.7%)	1,321 (25%)

[a]Statistics presented: median (IQR); n (%)
[b]"kg" kilograms
[c]"w" weeks; "d" days
[d]Statistics presented: mean ± SD

Table 3.3: Population characteristics stratified by outcome

Characteristic	Overall, N = 5,261	QBL < 1000mL, N = 3940 (75%)[a]	QBL ≥ 1000mL, N = 1321 (25%)[a]	p-value[b]
Age	31 (27, 34)	31 (27, 34)	31 (28, 35)	<0.001
Pre-pregnancy weight (kg[c])	71 (61, 84)	70 (61, 84)	73 (63, 87)	<0.001
Missing	899	675	224	
Admission weight (kg[c])	83 (73, 96)	82 (73, 95)	86 (75, 100)	<0.001
Missing	158	132	26	
Gestational diabetes	423 (8.0%)	286 (7.3%)	137 (10%)	<0.001
Gestational hypertension	547 (10%)	383 (9.7%)	164 (12%)	0.006
Gravidity[d]	2 (1, 3)	2 (1, 3)	2 (1, 3)	<0.001
Missing	1	1	0	
Parity	1 (0, 1)	1 (0, 1)	0 (0, 1)	<0.001
Missing	1	1	0	
Gestational Age	39w 2d (38w 2d, 40w 3d)[e]	39w 2d (38.43, 40w 3d)[e]	39w 1d (38, 40w 3d)[e]	0.046
Missing	19	19	0	
Anemia	812 (15%)	582 (15%)	230 (17%)	0.024
Assisted reproductive technology	236 (4.5%)	122 (3.1%)	114 (8.6%)	<0.001
Temperature	98.1 (97.9, 98.4)	98.1 (97.9, 98.3)	98.1 (97.9, 98.4)	0.058
Missing	1	1	0	
Cesarean Delivery[d]	1,677 (32%)	1,023 (26%)	654 (50%)	<0.001
Fetal macrosomia	187 (3.6%)	122 (3.1%)	65 (4.9%)	0.003
Multiple gestation	162 (3.1%)	80 (2.0%)	82 (6.2%)	<0.001
Placenta previa	226 (4.3%)	138 (3.5%)	88 (6.7%)	<0.001
Prior cesarean delivery	960 (18%)	663 (17%)	297 (22%)	<0.001
Spontaneous labor	107 (2.0%)	81 (2.1%)	26 (2.0%)	>0.9
Systolic pressure	125 (116, 134)	124 (116, 133)	127 (117, 136)	<0.001
Trial of labor	155 (2.9%)	114 (2.9%)	41 (3.1%)	0.8

[a]Statistics presented: median (IQR); n (%)
[b]Statistical tests performed: Wilcoxon rank-sum test; chi-square test of independence. Values <0.05 are in bold.
[c]"kg" kilograms
[d]Not included in models
[e]"w" weeks; "d" days

Table 3.4: Model performance, with C-statistics and 95% confidence intervals

Model	Original study	Validation set using original models[a]	Validation set using refit models[a]	p-value[b]
Logistic Regression	0.87; 95% CI[c] 0.86-0.87	0.54; 95% CI[c]: 0.52-0.55	0.63; 95% CI[c]: 0.61-0.65	<0.001
Lasso Regression	0.87; 95% CI[c] 0.86-0.88	0.55; 95% CI[c]: 0.53-0.57	0.63; 95% CI[c]: 0.61-0.65	<0.001
Random Forest	0.92; 95% CI[c]: 0.91-0.92	0.53 95% CI[c]: 0.51-0.55	0.64 95% CI[c]: 0.62-0.65	<0.001
Gradient Boosting	0.93; 95% CI[c]: 0.92-0.93	0.57; 95% CI[c]: 0.55-0.59	0.62; 95% CI[c]: 0.61-0.64	<0.001

[a]Validation set confidence interval was calculated using 2000 stratified bootstrap replicates (n = 5,261).
[b]Significance test comparing original models and refit models in the validation set using 2000 stratified bootstrap replicates (n = 5,261).
[c]"CI" confidence interval

3.4.1 External Validation of the CSL Models

The CSL models had poor discrimination in our validation cohort, with C-statistics of 0.54, 0.55, 0.53, and 0.57 for logistic regression (LR), lasso regression, random forests (RF), and gradient boosting machines (GBM), respectively (Table 3.4). The models were poorly calibrated, with the LR, lasso, and GBM models underestimating risk in the lower range of predicted risk and overestimating risk in the upper range of predicted risk. The RF model's predicted probabilities most closely matched the observed risk, though the predictions were tightly clustered together.

3.4.2 Refitting CSL Models Using Our Study Cohort

The refit models achieved better discrimination, with cross-validated C-statistics of 0.63 (LR), 0.63 (lasso), 0.64 (RF), and 0.62 (GBM), which were all statistically significant with $p < 0.001$ when compared to the original models (Table 3.4). Calibration was also improved for the refit models (Figure 3.3). A summary of the sensitivities, specificities, positive predictive values, and negative predictive values across the entire range of thresholds is shown in Supplemental Figure 3.4.

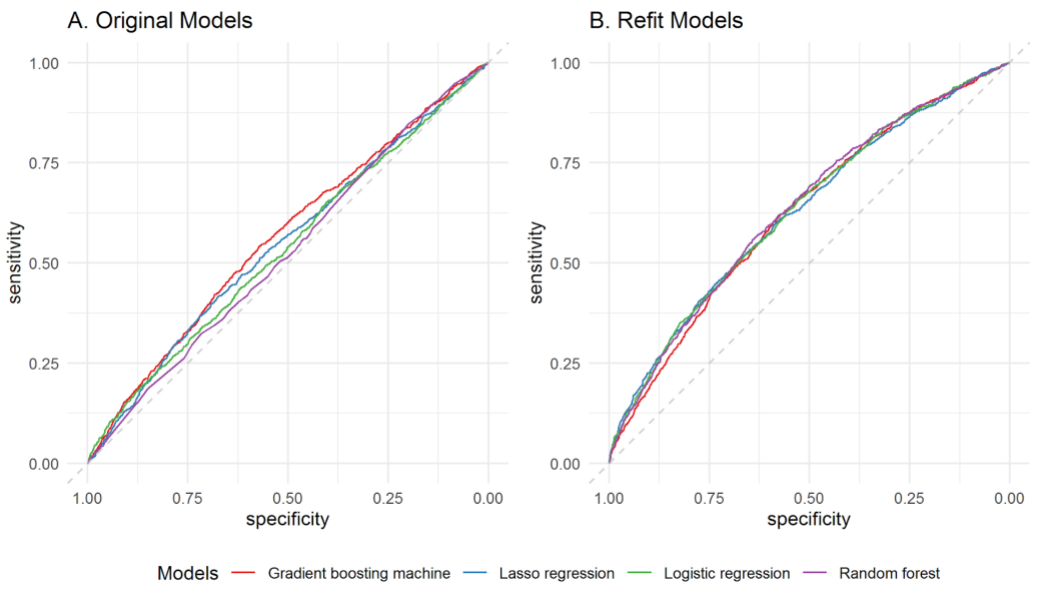


Figure 3.2: Comparison of the receiver operating characteristic curves
Comparison of the receiver operating characteristic curves for the original models (A) and refit models (B) in our study cohort. The refit models have a higher area under the receiver operating characteristic curve (i.e., C-statistic) as compared to the original models.

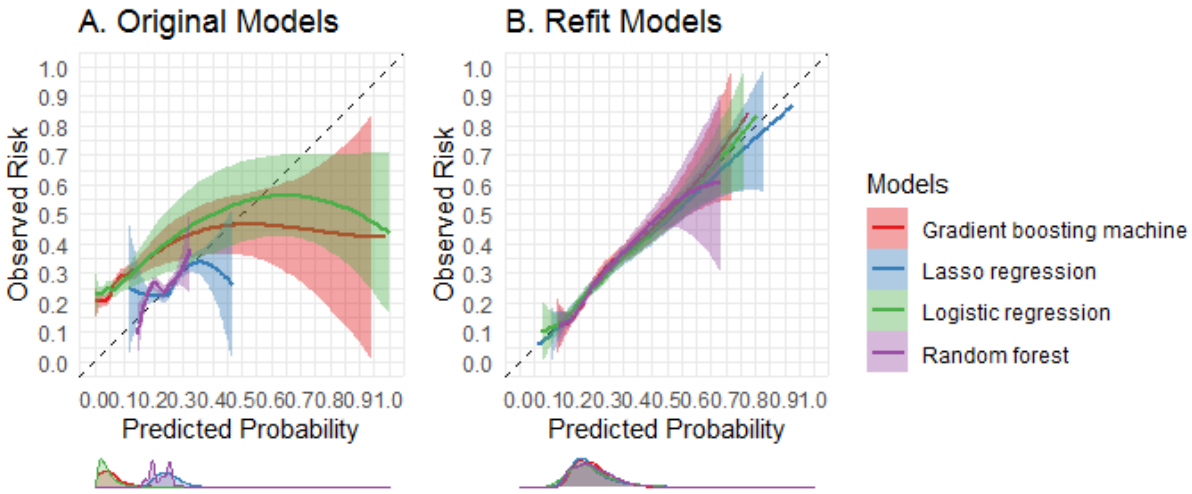


Figure 3.3: Calibration plots comparing predicted versus observed risk
Calibration plots comparing predicted versus observed risk for the original models (A) and refit models (B) in our study cohort. The dotted line demonstrates ideal calibration, and the calibration curves and 95% confidence intervals are calculated using loess smoothing [66], [67].

3.4.3 Sensitivity Analyses

Of the 5,261 deliveries, 569 (11%) met a QBL of ≥1500 mL, 249 (4.7%) met a QBL of ≥ 2000 mL, and 116 (2.2%) met a QBL of ≥2500 mL. The model performance did not

substantially improve (Table 3.5). The best-performing models achieved C-statistics of 0.59 (QBL ≥ 1500mL), 0.60 (QBL ≥ 2000 mL), and 0.61 (QBL ≥ 2500 mL).

Table 3.5: Sensitivity analysis

Model	≥1000mL (1321/5261; 25.1%)	≥1500mL (569/5261; 10.8%)	≥2000mL (249/5261; 4.7%)	≥2500mL (116/5261; 2.2%)
Logistic Regression	0.54; 95% CI: 0.52-0.55	0.54; 95% CI: 0.51-0.56	0.57; 95% CI: 0.53-0.61	0.59; 95% CI: 0.53-0.65
Lasso Regression	0.55; 95% CI: 0.53-0.57	0.55; 95% CI: 0.52-0.57	0.57; 95% CI: 0.53-0.61	0.61; 95% CI: 0.55-0.66
Random Forest	0.53; 95% CI: 0.51-0.55	0.53; 95% CI: 0.51-0.56	0.48; 95% CI: 0.44-0.51	0.52; 95% CI: 0.47-0.57
Gradient Boosting	0.57; 95% CI: 0.55-0.59	0.59; 95% CI: 0.57-0.62	0.60; 95% CI: 0.56-0.63	0.61; 95% CI: 0.56-0.67

3.5 Discussion

Overall, we found that the CSL models, which were highly accurate in the prediction of PPH in a geographically diverse development cohort across the U.S. involving 152,279 deliveries between 2002 and 2008, did not perform as well in our contemporary cohort of 5,261 deliveries when the outcome was defined using QBL.

The difference in model performance was quite substantial, with the best-performing model's C-statistic deteriorating from 0.93 in the original study to 0.57 in our cohort. Compared to the model development cohort, our validation cohort had a nearly 5-fold higher incidence of PPH (25% vs. 4.8%), was older, and had more chronic comorbidities. Refitting the models in our validation cohort mildly improved the model performance resulting in a cross-validated C-statistic of 0.64, which is lower than that observed in the original study.

The most likely explanation for the difference in model performance is due to how the outcomes were measured in both cohorts. Whereas the CSL study cohort defined PPH based on EBL, our study cohort relies on QBL, which is more accurate [58]. The incidence of PPH when assessed using the same ≥1000 mL threshold of blood loss is several-fold higher when measured by QBL as compared to EBL [57], suggesting that EBL may only be identifying severe hemorrhage whereas QBL may also be capturing

less severe bleeding. Unsurprisingly, severe hemorrhage may be easier for a model to predict and result in a higher C-statistic.

The difference in PPH incidence as well as measurement alone does not explain the deterioration in model performance. While models transported from a lower risk to a higher risk setting are commonly found to be miscalibrated, changes in baseline risk do not typically affect model discrimination (C-statistic). This has been demonstrated in multinational validation studies, where high-performing models often need to be recalibrated despite maintaining a high C-statistic [76]. The most extreme form of recalibration involves refitting the model on the new dataset [77]. When we refit models in our cohort, the improvement in model performance was modest, suggesting that the deterioration model performance cannot be attributed solely to higher PPH incidence, suggesting that there may be some differences in predicting EBL versus QBL outcomes.

It is also possible that the original CSL models overestimated model performance because of the way that the underlying data were collected. The CSL dataset was constructed two decades ago, when electronic health record (EHR) systems were in their infancy. In this context, there may have been under ascertainment of both predictors and PPH due to incomplete data capture. Additionally, in the CSL dataset, comorbidities were flagged as present on admission based on billing codes from the labor and delivery encounter. Since the CSL data were deidentified, it was not possible to confirm whether these comorbidities would have been accessible to a prediction model within the electronic health record (EHR) at the time of admission. Because of this [78], we limited billing codes in our cohort to those available during the prenatal period.

Strengths of our study include the use of a contemporary cohort, the use of QBL to measure PPH, and close coordination with the authors of the original CSL modeling study to ensure consistent variable definitions between the studies. The primary limitation of our study is that the data are drawn from a single tertiary care center while the original CSL models were developed from a multi-center U.S. cohort involving academic and community-based hospitals. Given known issues with EHR data quality, including missing data, some degree of model deterioration is expected simply due to the data source and differences in case-mix. We attempted to mitigate these concerns by comparing our EHR-derived predictors with national estimates, finding our predictors to generally be within the

expected range. Our center's PPH incidence is higher than other cohorts, which is a limitation of our study. Although this could be due to systematic overreporting of QBL, alternate definitions of PPH did not improve the performance of the CSL models, and thus the CSL models' lower performance in our cohort is not explained only by the choice of PPH definition.

Despite these limitations, our study has national implications for the implementation of prediction models for postpartum hemorrhage. Although predicting PPH by EBL appears highly feasible based on prior work [36], [37], predicting QBL appears to be quite difficult, likely because QBL identifies milder bleeds. As hospitals shift towards quantitative methods to assess bleeding severity, prediction models may be severely impacted. Whether bleeding identified by QBL confers the same risk of maternal morbidity and mortality as bleeding identified by EBL remains unknown. Improved understanding of the relationship between QBL and EBL is needed to better define the utility of prediction models predicting PPH. Our findings underscore the importance of external validation, particularly when data collection methods and outcome measurements evolve due to changes in clinical practice.

3.6 Supplemental Tables

Supplemental Table 3.6: Population characteristics by outcome (expanded)

Characteristic	**Overall, N = 5,261**	**QBL < 1000mL, N = 3940 (75%)[a]**	**QBL ≥ 1000mL, N = 1321 (25%)[a]**	**p-value[b]**
Admission height	164 (160, 168)	163 (160, 168)	165 (160, 170)	0.089
Missing	158	132	26	
Admission weight (Kg)	83 (73, 96)	82 (73, 95)	86 (75, 100)	<0.001
Missing	158	132	26	
Age at Admission	31 (27, 34)	31 (27, 34)	31 (28, 35)	<0.001
Anemia	812 (15%)	582 (15%)	230 (17%)	0.024
Antenatal steroids	619 (12%)	443 (11%)	176 (13%)	0.048
Antepartum hospitalizations	877 (17%)	607 (15%)	270 (20%)	<0.001
Antepartum vaginal bleeding	150 (2.9)	109 (2.8)	41 (3.1)	0.59
Assisted reproductive technology	236 (4.5%)	122 (3.1%)	114 (8.6%)	<0.001

Asthma/active airway disease	417 (7.9%)	299 (7.6%)	118 (8.9%)	0.13
BMI on Admission	31 (27, 36)	31 (27, 35)	32 (28, 37)	<0.001
Missing	158	132	26	
BMI Pre-pregnancy	26 (23, 31)	26 (23, 31)	27 (23, 32)	<0.001
Missing	899	675	224	
Body Temperature	98.10 (97.90, 98.40)	98.10 (97.90, 98.30)	98.10 (97.90, 98.40)	0.058
Missing	1	1	0	
Breech/abnormal lie	522 (9.9%)	365 (9.3%)	157 (12%)	0.007
Cesarean delivery	1,677 (32%)	1,023 (26%)	654 (50%)	<0.001
Chorioamnionitis on admission	61 (1.2%)	39 (1.0%)	22 (1.7%)	0.066
Chronic hypertension	579 (11%)	401 (10%)	178 (13%)	0.001
Chronic renal disease	65 (1.2%)	45 (1.1%)	20 (1.5%)	0.4
dataset				>0.9
train	2,750 (52%)	2,064 (52%)	686 (52%)	
tune	687 (13%)	510 (13%)	177 (13%)	
test	1,824 (35%)	1,366 (35%)	458 (35%)	
Depression	541 (10%)	398 (10%)	143 (11%)	0.5
Diastolic Pressure	77 (70, 84)	77 (70, 84)	78 (71, 86)	<0.001
Eclampsia	2 (<0.1%)	0 (0%)	2 (0.2%)	0.063
QBL	631 (374, 1,002)	494 (305, 706)	1,419 (1,176, 1,850)	<0.001
Ethnicity				0.8
Hispanic or Latino	85 (4.4%)	57 (4.1%)	28 (4.9%)	
Non-Hispanic or Latino	1,834 (94%)	1,303 (94%)	531 (94%)	
Patient Refused	6 (0.3%)	4 (0.3%)	2 (0.4%)	
Unknown	25 (1.3%)	19 (1.4%)	6 (1.1%)	
Missing	3,311	2,557	754	
Fetal demise	46 (0.9%)	32 (0.8%)	14 (1.1%)	0.5
Fetal macrosomia	187 (3.6%)	122 (3.1%)	65 (4.9%)	0.003
Gastrointestinal disease	665 (13%)	499 (13%)	166 (13%)	>0.9
Gestational age (weeks)	39w 2d (38w 2d, 40w 3d)[c]	39w 2d (38w 3d, 40w 3d)[c]	39w 1d (38w, 40w 3d)[c]	0.046
Missing	19	19	0	
Gestational diabetes	423 (8.0%)	286 (7.3%)	137 (10%)	<0.001
Gestational hypertension	547 (10%)	383 (9.7%)	164 (12%)	0.006

Gestational age (days)	275 (268, 283)	275 (269, 283)	274 (266, 283)	0.046
Missing	19	19	0	
Gravidity	2 (1, 3)	2 (1, 3)	2 (1, 3)	<0.001
Missing	1	1	0	
Heart disease	224 (4.3%)	171 (4.3%)	53 (4.0%)	0.7
Hemoglobin	345 (6.6%)	253 (6.4%)	92 (7.0%)	0.5
History of preterm labor	140 (2.7%)	106 (2.7%)	34 (2.6%)	0.9
History of seizures	177 (3.4%)	129 (3.3%)	48 (3.6%)	0.6
Illicit drug use during pregnancy	617 (12%)	482 (12%)	135 (10%)	0.053
Missing	14	12	2	
Insurance				0.041
Medicaid	790 (24%)	610 (25%)	180 (22%)	
Medicare	16 (0.5%)	8 (0.3%)	8 (1.0%)	
Other	9 (0.3%)	6 (0.2%)	3 (0.4%)	
Other Governmental Insurance	16 (0.5%)	9 (0.4%)	7 (0.8%)	
Private Insurance	2,465 (74%)	1,832 (74%)	633 (76%)	
Workers Compensation	15 (0.5%)	11 (0.4%)	4 (0.5%)	
Missing	1,950	1,464	486	
Intrauterine growth restriction	707 (13%)	517 (13%)	190 (14%)	0.3
Large for gestational age	15 (0.3%)	10 (0.3%)	5 (0.4%)	0.5
Length of stay	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	3.00 (2.00, 4.00)	<0.001
Magnesium sulfate	300 (5.7%)	206 (5.2%)	94 (7.1%)	0.013
Marital status				0.039
Married	2,315 (60%)	1,698 (59%)	617 (62%)	
Unmarried	1,572 (40%)	1,200 (41%)	372 (38%)	
Missing	1,374	1,042	332	
Maternal GBS colonization	1,163 (22%)	871 (22%)	292 (22%)	>0.9
Multiple gestation	153 (2.9%)	73 (1.9%)	80 (6.1%)	<0.001
Multiple gestation	162 (3.1%)	80 (2.0%)	82 (6.2%)	<0.001
Non-gestational diabetes	137 (2.6%)	90 (2.3%)	47 (3.6%)	0.016
Parity	1 (0, 1)	1 (0, 1)	0 (0, 1)	<0.001
Missing	1	1	0	

Placenta accreta spectrum	19 (0.4%)	8 (0.2%)	11 (0.8%)	0.002
Placenta previa	226 (4.3%)	138 (3.5%)	88 (6.7%)	<0.001
Placental abruption	59 (1.1%)	33 (0.8%)	26 (2.0%)	0.001
Polyhydramnios	177 (3.4%)	110 (2.8%)	67 (5.1%)	<0.001
Pre-pregnancy height	164 (160, 168)	164 (160, 168)	165 (160, 170)	0.020
Missing	899	675	224	
Pre-pregnancy weight (Kg)	71 (61, 84)	70 (61, 84)	73 (63, 87)	<0.001
Missing	899	675	224	
Preeclampsia with severe features	147 (2.8%)	93 (2.4%)	54 (4.1%)	0.001
Preeclampsia without severe features	147 (2.8%)	85 (2.2%)	62 (4.7%)	<0.001
Premature rupture of membranes	599 (11%)	474 (12%)	125 (9.5%)	0.013
Preterm labor	81 (1.5%)	60 (1.5%)	21 (1.6%)	>0.9
Prior cesarean delivery	960 (18%)	663 (17%)	297 (22%)	<0.001
Race				0.6
African American	639 (12%)	482 (12%)	157 (12%)	
Asian	448 (8.5%)	328 (8.3%)	120 (9.1%)	
Caucasian	3,679 (70%)	2,768 (70%)	911 (69%)	
Other	495 (9.4%)	362 (9.2%)	133 (10%)	
Seizure disorder	51 (1.0%)	33 (0.8%)	18 (1.4%)	0.13
Spontaneous labor	107 (2.0%)	81 (2.1%)	26 (2.0%)	>0.9
Superimposed preeclampsia	44 (0.8%)	29 (0.7%)	15 (1.1%)	0.2
Systolic Pressure	125 (116, 134)	124 (116, 133)	127 (117, 136)	<0.001
Thyroid disease	497 (9.4%)	348 (8.8%)	149 (11%)	0.010
Tobacco use during pregnancy	432 (8.2%)	339 (8.6%)	93 (7.1%)	0.080
Missing	14	12	2	
Trial of labor	155 (2.9%)	114 (2.9%)	41 (3.1%)	0.8
Years of education	16.0 (14.0, 18.0)	16.0 (14.0, 18.0)	16.0 (15.0, 18.0)	>0.9
Missing	4,756	3,565	1,191	

[a]Statistics presented: median (IQR); n (%)

[b]Statistical tests performed: Wilcoxon rank-sum test; chi-square test of independence; Fisher's exact test

[c]"w" weeks; "d" days

Supplemental Table 3.7: Feature definitions

Label	ICD-10 definition	Timepoint(s) Included	Expected Incidence	Reference
Admission diastolic blood pressure		Admission intake		
Admission systolic blood pressure		Admission intake		
Admission temperature		Admission intake		
Admission weight		Admission intake		
Age		Admission intake		
Anemia	O99.0|D50.|D51.|D52.|D53.|D55.|D56.|D57.|D58.|D59.|D60.|D61.|D62.|D63.|D64.; excludes text matches to "anemia of pregnancy" and "sickle cell trait"	Intra-pregnancy	2%	ACOG
Antenatal steroids	text matching "betamethasone|dexamethasone"	Intra-pregnancy		
Antepartum hospitalization		Intra-pregnancy		
Antepartum vaginal bleeding	O20.|O46.|O44.31; excludes text matching "subchorionic hemorrhage", "subchorionic hematoma", "bloody show and cramping" prior to admission	Intra-pregnancy	15-25%	ACOG
Assisted reproductive technology	O09.81; excludes O30.009, "twin pregnancy"	Intra-pregnancy	1.50%	ACOG
Asthma/active airway disease	J44.|J45.	Intra-pregnancy	4-8%	ACOG
Breech presentation/abnormal lie	O32.1|O32.2|O32.8|O64.1|O80.1|O83.0|O83.1	Intra-pregnancy	3-4%	ACOG
Chorioamnionitis on admission	O41.1	Intra-pregnancy	2-5%	ACOG
Chronic hypertension	O10., O16.	Pre-pregnancy, Intra-pregnancy	.9-1.5%	ACOG
Chronic Renal disease	N02.2|N03.|N04.|N05.|N08.|N17.1|N17.2|N18.|N25|O26.82|O26.83; excluded nephrolithiasis, kidney stone, calculus	Intra-pregnancy	.02-.12%	Edipidis 2011 (DOI: 10.1111/aogs.13751)
Depression	099.34|F32.9; excludes anxiety unless "anxiety and depression"	Intra-pregnancy	9%	ACOG
Drug use		Intra-pregnancy		
Eclampsia	O15.	Intra-pregnancy		
Education status	Years of education	Intra-pregnancy		
Fetal demise	O36.4	Intra-pregnancy	0.60%	ACOG
Fetal macrosomia	O36.6	Intra-pregnancy	7.80%	ACOG

Gastrointestinal disease	K20.|K21.|K22.|K23.|K24.|K25.|K26.|K27.|K28.|K29.|K30.|K31.|K35.|K36.|K37.|K38.|K50.|K51.|K52.|K55.|K56.|K57.|K58.|K59.|K60.|K61.|K62.|K63.|K64.|K70.|K71.|K72.|K73.|K74.|K75.|K76.|K77.|K80.|K81.|K82.|K83.|K84.|K85.|K86.|K87.|K90.|K91.|K92.|K93.|K94.|K95.	Intra-pregnancy		
Gestational age		Admission intake		
Gestational diabetes	O24.4|O24.9	Intra-pregnancy	7%	ACOG
Gestational hypertension	O13.	Intra-pregnancy	2-8%	ACOG
Heart disease	I05.|I06.|I07.|I08.|I09.|I34.|I35.|I36.|I37.|I38.|I39.|I50.0|I20.|I25.|Q20.|Q21.|Q22.|Q23.|Q24.|Q25.|Q26.|O99.4	Intra-pregnancy	1-4%	ACOG (DOI: 10.1111/aogs.13749)
History of preterm birth	Z87.51	Pre-pregnancy		
History of seizures	G40.	Pre-pregnancy		
Illegal Drug Use during Pregnancy		Intra-pregnancy		
Insurance status		Admission intake		
Intrauterine growth restriction	O36.59	Intra-pregnancy	3-5%	Romo 2008
Large for gestational age		Intra-pregnancy		
Magnesium sulfate	text matching "magnesium sulfate"	Intra-pregnancy		
Marital status		Admission intake		
Maternal GBS colonization (dx)	O99.820|B95.1 or A49.1 and "GBS" or R82.71|O23.40|O09.89|O09.29 and "GBS|GROUP B STREPTOCOCCAL"	Intra-pregnancy	10-30%	Shabayek 2018
Maternal GBS colonization (lab)	positive strep group b results	Intra-pregnancy		
Multiple gestation (dx)	O30.|O31.8|O32.5|O84.8	Admission intake	.1-3%	ACOG
Multiple gestation (stork)		Admission intake		
Non-gestational diabetes	O24.0|O24.1|O24.2|O24.3	Intra-pregnancy	1-2%	ACOG
Parity		Admission intake		
Placenta accreta spectrum	O43.2	Intra-pregnancy	.29-.38%	
Placenta previa	O44.	Intra-pregnancy	0.50%	Durst 2018
Placental abruption	O45.	Intra-pregnancy	7.4-11.9%	Ananth 2015
Polyhydramnios	O40.	Intra-pregnancy		
Pre-pregnancy BMI		Pre-pregnancy		

Pre-pregnancy weight		Pre-pregnancy		
Preeclampsia with severe features	O14.1|O14.2	Intra-pregnancy		
Preeclampsia without severe features	O14.; anything not captured under "preeclampsia with severe features"	Intra-pregnancy		
Premature rupture of membranes	O42.	Intra-pregnancy	8%	
Preterm labor	O60.	Intra-pregnancy		
Prior cesarean delivery	O34.21|O34.29	Pre-pregnancy		
Race		Admission intake		
Seizure disorder	G40. on admission	Intra-pregnancy	.3-.7%	Ruth 2013
Spontaneous labor	O60.1|O80.|O84.0	Intra-pregnancy		
Superimposed preeclampsia	O11.	Intra-pregnancy		
Thyroid disease	E00.|E01.|E02.|E03.|E04.|E05.|E06.|E07.; excluded E08. (Diabetes mellitus)	Intra-pregnancy		
Tobacco use		Intra-pregnancy		
Trial of labor	O75.7|O66. or text matching "trial of labor|TOLAC"	Intra-pregnancy		
Outcome: Primary Postpartum Hemorrhage		Delivery encounter		

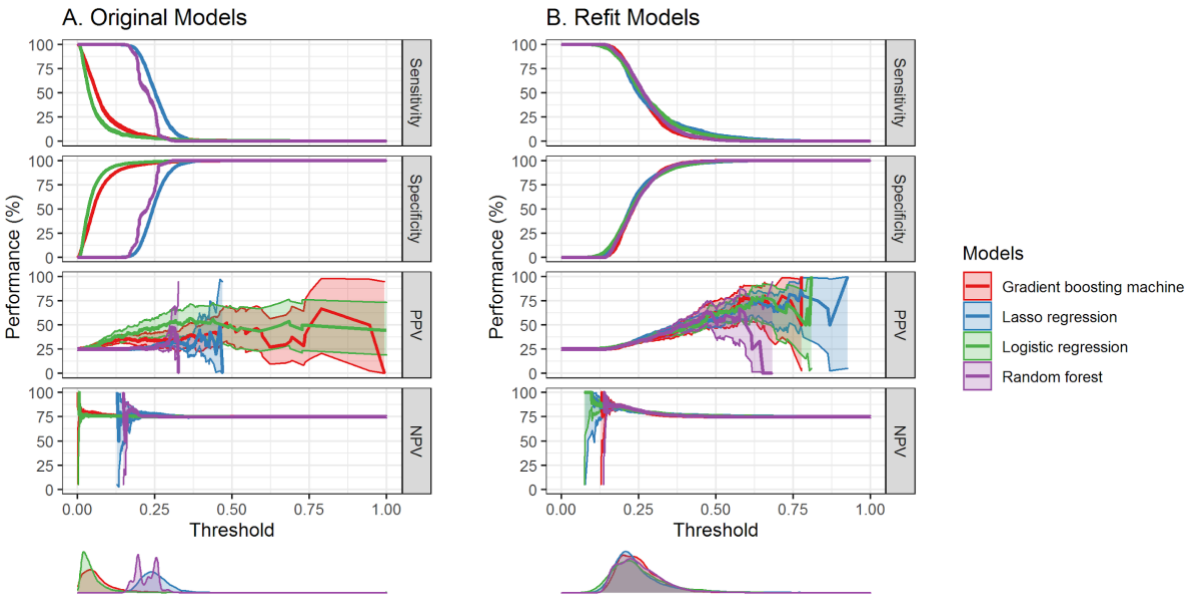
Supplemental Table 3.8: Predictor inputs

	Statistical Models		**Machine Learning Models**	
Variable	**Lasso Regression**	**Logistic Regression**	**Extreme Gradient Boosted**	**Random Forest**
Admission diastolic blood pressure	X	X	X	X
Admission systolic blood pressure	X	X	X	X
Admission temperature	X	X	X	
Admission weight	X	X	X	
Age	X	X	X	X
Anemia	X	X	X	
Antenatal steroids	X	X	X	X
Antepartum vaginal bleeding	X	X	X	X
Assisted reproductive technology	X	X	X	X
Asthma	X	X		
Breech presentation	X	X	X	X
Chorioamnionitis on admission	X	X	X	X
Chronic hypertension	X	X	X	
Depression	X	X		
Drug use	X	X		
Ecclampsia	X	X		

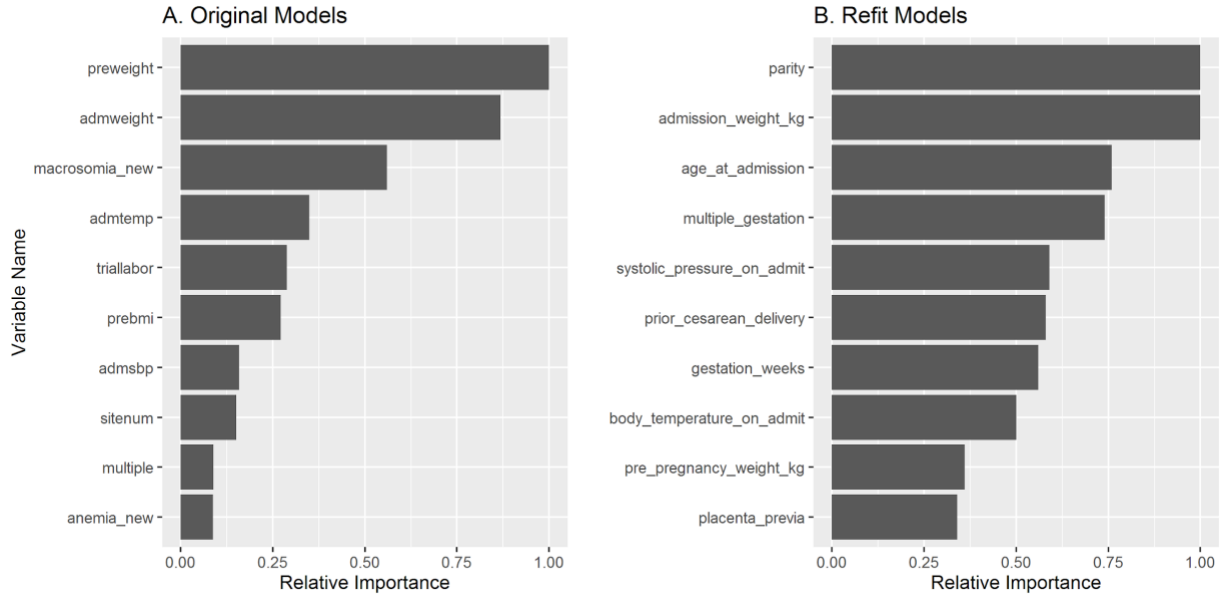
Education status	X	X	X	
Fetal demise	X	X		
Fetal macrosomia	X	X	X	X
Gastrointestinal disease	X	X		
Gestational age at delivery	X	X	X	X
Gestational diabetes	X	X	X	
Gestational hypertension	X	X		
Heart disease	X	X		
History of prior cesarean delivery	X	X	X	
History of prior preterm birth	X	X		
History of seizures	X	X		
Insurance status	X	X	X	
Large for gestational age, antenatal diagnosis	X	X	X	X
Magnesium sulfate	X	X		
Marital status	X	X	X	
Maternal GBS colonization	X	X		
Maternal race	X	X	X	X
Multiple gestation	X	X	X	
Non-gestational diabetes	X	X		
Parity	X	X	X	
Placenta accreta	X	X	X	X
Placenta previa	X	X	X	X
Placental abruption	X	X	X	X
Polyhydramnios	X	X	X	
Pre-pregnancy BMI	X	X	X	
Pre-pregnancy weight	X	X	X	
Preeclampsia with severe features	X	X	X	
Preeclampsia without severe features	X	X	X	X
Premature rupture of membranes	X	X	X	
Prior antepartum hospitalization	X	X	X	
Renal disease	X	X		
Seizure disorder	X	X	X	X
Site Number			X	
Small for gestational age, antenatal diagnosis	X	X	X	X
Spontaneous labor	X	X	X	X
Superimposed preeclampsia	X			
Threatened preterm labor	X	X	X	X
Thyroid disease	X	X		

Tobacco use	X	X		
Trial of labor	X	X	X	X

3.7 Supplemental Figures



Supplemental Figure 3.4: Threshold performance plot
This plot compares (A) original model performance to (B) refit model performance.



Supplemental Figure 3.5: Variable importance
A comparison of relative variable importance between (A) original models and (B) refit models.

Chapter 4
wizard for R: Windowing and Summarization for Autoregressive Data Preparation

This chapter was co-authored with Karandeep Singh.

4.1 Introduction

Preparing data for the purpose of developing prediction models is a time-consuming but necessary step [39], [40], but well-prepared data can often lead to better results [41]. Data preparation requires many design decisions to be made with respect to which variables to include and how to represent them. Effective decision-making ensures that prediction models perform as well as possible while the underlying predictors remain interpretable to clinicians. These design decisions typically evolve through repeated conversations between domain experts (clinicians) and data experts. They often occur implicitly because clinicians are imprecise when discussing clinical concepts (e.g., “baseline value”), and data analysts do not have a framework capturing the different ways in which clinical knowledge can be represented for modeling.

While data preparation is challenging for all clinical prediction models, this is especially true of early warning system (EWS) models, which refer to a class of prediction models that estimate an individual’s risk of an adverse outcome at multiple time points during an at-risk period such as a hospitalization. When a patient’s risk exceeds a specified threshold, an alert is sent to a clinician to institute a specific action. Such models are widely used in the detection of sepsis (a severe inflammatory response to infection) and clinical deterioration, such as the need for intensive care unit-level care.

In EWS models, each patient’s at-risk period begins at a specific time point (e.g., admission to the hospital and ends at a specific time point (e.g., discharge from the

hospital). For the duration of the at-risk period, risk is reassessed at regular intervals (e.g., every hour). For the purpose of prediction modeling, each row of data prepared for use in an EWS model represents a patient's clinical state at a specific point in time based on characteristics that are fixed through the hospitalization (e.g., biological sex) and those that are time-varying (e.g, vital signs or laboratory values).

When prior values are used to predict future values at regular intervals, this is referred to as "discrete time survival modeling" in the statistical literature, as "auto-regression" in the economics literature, and as "sequence modeling" in the machine learning literature. Because clinical data elements are collected at irregular intervals, including clinical elements as predictors in models requires summarization over multiple values (e.g., mean or exponentially weighted moving average), and in some cases, windowing (e.g., separate sets of mean vital signs for each of the last 3 nursing shifts). Clinical domain experts have access to explicit and tacit knowledge that can inform how clinical information is represented, a practice commonly referred to as "feature engineering," but communicating and incorporating this knowledge into the data preparation process requires substantial analytical effort and custom code development.

In this paper, we present a case study of maternal early warning systems to highlight how implicit clinical knowledge can be made explicit through engagement with domain experts (clinicians). We consider the implications in more detail as they pertain to the bounds and frequency of predictions, baseline predictors, time-varying predictors, and the conversion of vague clinical concepts into precise mathematical ones. We discuss the challenges of converting unevenly spaced data into evenly spaced data. We then provide an overview of commonly used tools for preparing time-series data for modeling in the R ecosystem, with a focus on why existing tools lack the expressivity to handle common use cases in clinical domain areas. Finally, we present the wizard R package, a software program that encodes a grammar of data preparation for EWS models through windowing and summarization for auto-regressive data preparation. Using wizard, analysts can effectively and efficiently prepare electronic health record (EHR) data for EWS modeling through a framework that is expressive and can directly leverage domain expertise.

4.2 Methods

Wizard was created out of a need to partially automate the data transformation process in a structured way while maintaining design controls over the result, which in the context of this body of work, we refer to as domain-expert-informed predictors. This section describes, in a qualitative manner, how components of wizard were developed and iterated over.

Over the course of two years, I met weekly with a team of domain experts to discuss the use of a time-series approach to predict postpartum hemorrhage using electronic health record data with the intent to design a model for early warning systems. The team consisted of two clinical domain experts, of which one was an expert in obstetric anesthesia and another in obstetrics and gynecology, and a physician with expertise in clinical prediction modeling. I led the discussion, conducted the analysis, and worked with a database analyst to identify the correct data elements and extract the data. Topics of discussion were predictors to be used in our modeling approach, the timespan or timespans that would be clinically relevant, and how to summarize each timespan of data. Based on feedback received from clinicians, we iteratively developed features of this package to accommodate a multitude of use cases for the transformation of electronic health record data guided by domain expertise.

Early in the work, I found that the domain experts and I lacked a common and consistent vocabulary to describe predictors with a time aspect, and that this hindered productivity. Over the course of our discussions, we started establishing design elements which we could then map to the conversations (Figure 4.1). We further realized that although initially we had only planned to use information from a fixed number of hours in the past (Section 4.6.10), recalculating these predictors at each new step was computationally expensive. Some of this complexity was likely unnecessary because it involved information collected prior to the at-risk period where no predictions are made. This warranted the need to differentiate timespans relative to the at-risk period, We determined that the first at-risk prediction was the fulcrum around which we chose to differentiate these timespans.

Complementary to baseline predictors, we found that there are variables that are only collected during the at-risk period. For example, blood loss is likely only collected

during a hospitalization and therefore would introduce sparsity and complexity to the model by creating additional predictors from timespans that didn't exist. Our solution was to introduce growing predictors (Section 4.6.9), which start at the first prediction and are recalculated at each new time step.

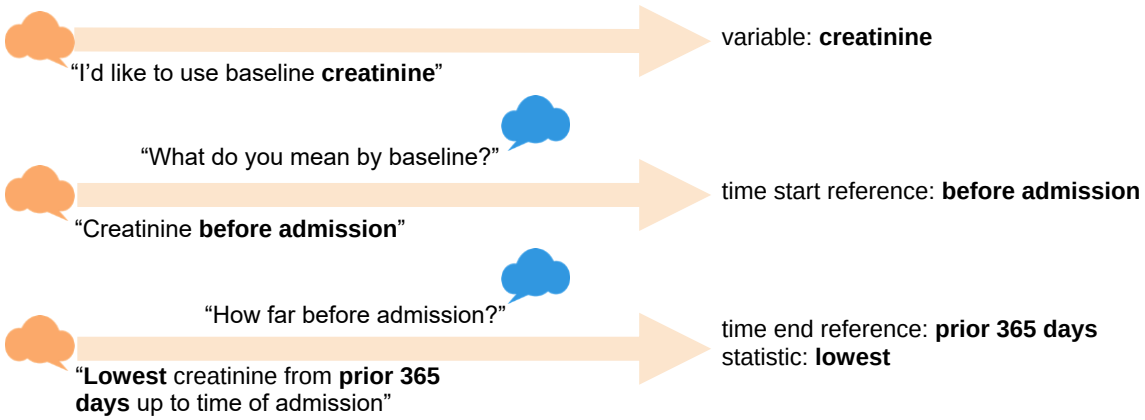


Figure 4.1: Clinical/data conversation

With respect to outcomes, we started by looking at a rolling outcome use case (Section 4.6.12) in which a model makes predictions for a certain amount of time in the future. We found that training a model with a fixed outcome (Section 4.6.11) that reflect whether a patient ever experiences the outcome during the at-risk period. We found a slight performance boost in using a fixed outcome, likely due to the way the algorithm trains the model. By having a persistent outcome for all predictions, the algorithm can focus on differences which occur between patients who do and do not experience the outcome rather than changes that occur when the rolling outcome changes within individuals.

We found other efficiencies in reducing the complexity of the data by splitting temporal from non-temporal data, the latter were termed fixed data. Fixed data is not expected to change throughout the course of an at-risk period, and the benefits of maintaining persistence between predictions outweighs the complexity of using a temporal approach (e.g., an age of 30 years 1 day old, 30 years 2 days old, etc.).

We started to look at summarization methods used in similar approaches. The initial implementation supports dplyr verbs but can be extended to accept custom functions. This allows basic statistics like the minimum, maximum, median, mode, but

also introduces the first and last value. Custom functions allow the ability to introduce more sophisticated calculations like slope or net change.

While we attempted to incorporate a broad use case for this work, there are still functions that have not been introduced. For example, in some cases when temporal data does not vary substantially (e.g., a type and screen lab test collected throughout a hospitalization), the information may be better served as fixed data rather than in the temporal data. Similarly, we have not yet implemented a function to calculate the duration from an event, such as the amount of time since surgery began.

4.3 Case Study: Maternal Early Warning Systems for Detecting Postpartum Hemorrhage

Maternal early warning systems are commonly used to detect adverse outcomes experienced by women during hospitalizations for labor and delivery. When the adverse events occur after delivery, they are referred to as being “postpartum.” One of the most common sources of material morbidity and mortality during the postpartum period is excessive bleeding (or “hemorrhage”). Thus, timely prediction of postpartum hemorrhage (PPH) is an ideal use case for an EWS model. We developed a model to predict postpartum hemorrhage among females (i.e., inclusive of women and trans men) admitted to the labor and delivery ward of a large academic medical center with direct input from an obstetrician and an obstetric anesthesiologist. In Table 4.1, we describe examples of the questions we asked them to elicit domain knowledge as well as the answers they provided and resulting implications.

Table 4.1: Examples of questions to elicit domain knowledge for data preparation

Questions	Answers	Implications
When does the risk of PPH begin?	Immediately after delivery, either through vaginal birth or a cesarean section (C-section).	The first row of data for each patient should be generated at the time of delivery.
When is a person no longer at risk of PPH?	After discharge from the hospital or 7 days after delivery has occurred if still hospitalized.	The final row of data for each patient should be generated at the time of hospital discharge or 7 days after delivery, whichever happens first.
How often should the risk of PPH be	Vital signs are checked every 20	A new row is needed for each

reassessed?	minutes, so it would be important to reassess the risk every 20 minutes.	patient every 20 minutes that captures the patient's state at that point in time.
What are predictors of PPH we should include in our model?	Whether the delivery was a C-section or vaginal delivery makes a big difference. C-section generally puts women at higher risk. Pre-pregnancy body mass index is important. But we should also include the body mass index from each of the trimesters. We should also include systolic blood pressure in the model because a low value can be an early indicator for PPH.	C-section is a "fixed" predictor for each patient after delivery has taken place (but would be considered time-varying if the at-risk period included the prepartum period) "Pre-pregnancy" and "during each of the trimesters" is not precise because each can refer to multiple values. The date of conception is not always accurately known and is often estimated. Systolic blood pressure is checked frequently. This statement is not precise enough to know which time period matters (e.g., last 48 hours) and what matters (e.g., mean value, minimum value, or slope).

4.3.1 The Bounds and Frequency of Predictions

By asking a clinician when the at-risk period begins (at delivery), how often predictions should be made (every 20 minutes), and when the at-risk period ends (at discharge or 7 days after delivery), the analyst can derive a grid of time points for each patient that will represent rows in the modeling data. Because the delivery time is different for each patient, the grid will need to be anchored to the respective patient's delivery time. Because each patient's at-risk period may be of a different length (e.g., 2 days versus 7 days), patients will not be equally represented in the data in terms of the number of rows. This complexity, which is a common source of errors, is known to the analyst but not easy to communicate to clinicians.

4.3.2 Baseline Predictors

As expressed in Table 4.1, the risk of PPH begins at the time of delivery. This has important implications because data elements available prior to (and up to the time of) delivery can be considered as fixed or baseline information, whereas subsequent data could be represented in a variety of ways. Whether a data element is definitively fixed depends on the format in which it has been made available to an analyst. When a data element is expressed at the level of the hospitalization in a "wide" dataset (Table 4.2), it

must be considered as fixed because it is only available at the hospitalization-level and thus cannot change during the course of the hospitalization. For example, each hospitalization can have at most one value of Delivery Type and Delivery Time in Table 4.2 because each row represents a single hospitalization.

Table 4.2: Example of fixed data expressed in "wide" format

Hospitalization ID	Age	Delivery Type	Admission Time	Delivery Time	Discharge Time
1	29	C-section	June 2, 2021 7:00 PM	June 3, 2021 4:10 PM	June 7, 2021 11:50 AM
2	35	Vaginal	June 6, 2021 6:30 AM	June 6, 2021 3:02 PM	June 8, 2021 1:05 PM

Baseline information can also be present in a temporal or time-varying format as expressed in Table 4.3. In this "long" format where each row represents a single observation, Delivery Type is still only represented once per patient. However, this constraint is not enforceable by the data structure itself. If Delivery Type were being extracted from multiple data sources, it is possible that an individual hospitalization could have two or more rows describing this information, potentially in conflicting ways.

Table 4.3: Example of temporal (or time-varying) data expressed in "long format"

Hospitalization ID	Timestamp	Variable	Value
1	June 2, 2021 9:19 PM	Systolic blood pressure	130
1	June 3, 2021 5:30 AM	Systolic blood pressure	150
1	June 3, 2021 4:10 PM	Delivery Type	C-section
2	June 6, 2021 8:05 AM	Systolic blood pressure	100
2	June 6, 2021 8:20 AM	Systolic blood pressure	110
2	June 6, 2021 3:02 PM	Delivery Type	Vaginal

Thus, an analyst faced with these two different datasets may readily pick up on the fact that Delivery Type is a fixed variable in Table 4.2 but not in Table 4.3 even though a clinician seeing data in either format will have a mental model for which predictors are actually fixed or time-varying, which will inform how they are represented.

4.3.3 Time-varying Predictors

Leveraging longitudinal EHR data in a meaningful way has long been a strategy absent in clinical prediction modeling. A systematic review revealed only around 9% of clinical models used time-varying repeated measures [79]. Recent clinical prediction models have shown that capturing predictors over time can improve performance [22], [37]. How a given time-varying predictor is best incorporated into a modeling dataset depends on domain knowledge about the potential causal mechanism or pace at which a predictor affects an outcome. While variation in body mass index is clinically relevant at the granularity of trimesters, blood pressure changes occurring in the span of minutes may signal imminent or early hemorrhage. If a medication is known to increase or decrease the risk of an adverse outcome, information about its half-life and dose-response may be biologically important factors that determine how far back in time the analyst needs to look (from the current time) to see if the medication was administered, or how the dose should represented (e.g., total dose, mean dose). Thus, the way in which time-varying predictors are represented will differ between clinical variables. Importantly, the time-related components (e.g., a trimester is 13 weeks) may be entirely unrelated to the frequency at which predictions will be made (e.g., every 20 minutes).

4.3.4 Conversion of Vague Clinical Concepts into Precise Mathematical Ones

Clinical concepts are often expressed in terms that, while generally understood by clinician practitioners, are typically too imprecise for analysts. As in the example in Table 4.1, clinicians may desire pre-pregnancy body mass index (BMI) to be included as a predictor. There is no variable in the electronic health record (EHR) that directly captures the pre-pregnancy BMI—only BMI in general. Even the date of conception may not be recorded in the EHR, so which values of BMI to consider pre-pregnancy may not be obvious to analysts. However, knowing trimesters are 13 weeks long (based on domain knowledge), the date of conception is commonly estimated to occur 39 weeks (13 x 3 = 39) prior to the date of delivery. Thus, pre-pregnancy BMI can be thought of as referring to a patient's weight in the period preceding 39 weeks prior to the date of delivery.

In practice, even this definition is not precise enough to implement. Patients may have multiple values of BMI recorded during the pre-pregnancy period. Thus, one may

need to condense several eligible values of BMI into a single value by applying a "summary function." Reasonable options to apply here could be the mean BMI value, the median BMI value, or the most recent BMI value among the series of eligible values.

4.4 The Challenge of Unevenly Spaced Data

4.4.1 Unevenly Spaced Data

A patient's state of health and subsequently the data collected are often observed at unevenly spaced time intervals (Figure 4.2) and different patients observed at different points in time. Without transformation, sequences at the multivariable or multi-subject level become out of phase in relation to each other which would exclude the application of many time-series modeling techniques. This often leaves two methods to convert unevenly spaced data into evenly spaced data: direct value interpolation of clinical state at regular time periods, and window-based segmentation.

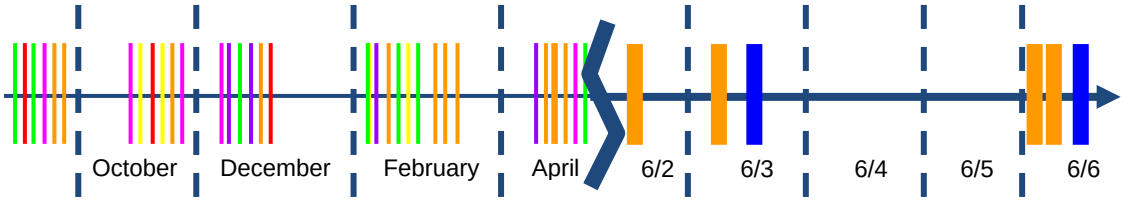


Figure 4.2: Unevenly spaced and sparse data
Note: Single-patient example with tick marks representing variable presence and color representing different variables.

In direct value interpolation (DVI), values are estimated as the mean of the value before and after a given time point at a designated interval. If no such value is present at a prediction interval, the interpolation may be based solely on information after the prediction, and thus cause leakage. While interpolation of the current value could be performed using only previous values, interpolation is commonly performed using linear regression, whose assumptions will need to be checked and may commonly be violated.

4.4.2 Aggregating Repeated Measures

Aggregation solves the problem of unevenly spaced data but as data are downsampled (multiple observations summarized as one value), information is lost and

approaches white noise [80]. Consider a scenario where all prior values of a variable (e.g., body mass index [BMI]) are summarized by a single value, expressed as a mean. If there are important trends in the BMI, such as those that occur during pregnancy, this solution will result in loss of important information. This can be overcome by dividing the lookback period (e.g., 12 months) into smaller intervals (e.g., 3 months) to capture state change, where each of the prior three 3-month periods represents a trimester of pregnancy, and the prior value represents pre-pregnancy values. Data can be further summarized for each interval by calculating multiple summary statistics (e.g., minimum, maximum, first, last) within each window.

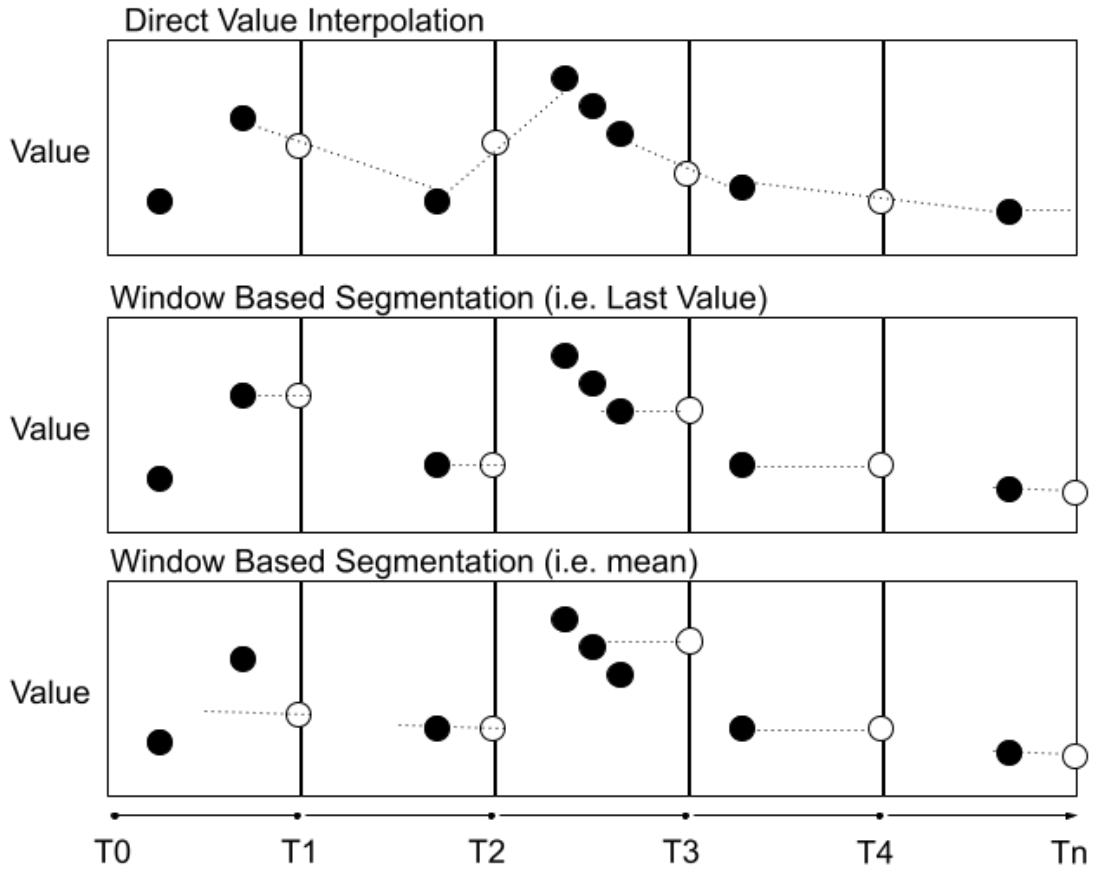


Figure 4.3: Methods for converting unevenly spaced data into evenly-spaced data
Let T_n be the prediction interval and let the window be the period between each interval inclusive of T_n. Real values are denoted as black circles and derived values are denoted by white circles. On the other hand, window-based segmentation converts unevenly spaced data into evenly-spaced data by dividing time into fixed-length windows, and then calculating summary statistics (e.g., mean, median, minimum, and/or maximum) for each window, as depicted in Figure 4.3.

4.4.3 Aggregation of Statistics as a Hierarchy

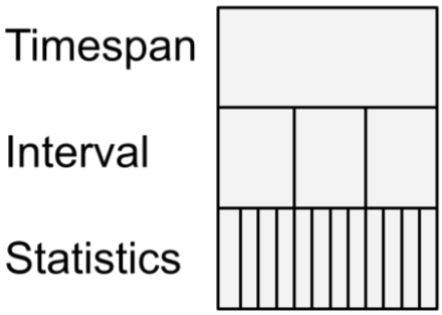


Figure 4.4: Temporal hierarchy

Temporal data can be thought of in terms of a time-based hierarchy (Figure 4.4). Assuming a prediction occurs at the right-most point, a timespan is the amount of time prior to the prediction to include information. Within that timespan, information can then be split into intervals assuming there is reason to do so (i.e. repeated measurements). From there, each interval can then be summarized using standard statistical methods like minimum, maximum as well as data order based methods like first or last within an interval to better capture trajectory (Figure 4.5). Since there is a single spike which trails off, it would be difficult to identify the relevant timespan if the entire block was statistically summarized given an acute outcome event. This would be even more pronounced for dichotomous variables since diagnoses in the distant past contribute albeit to a lesser extent than more recent diagnoses that may be more relevant to the outcome.

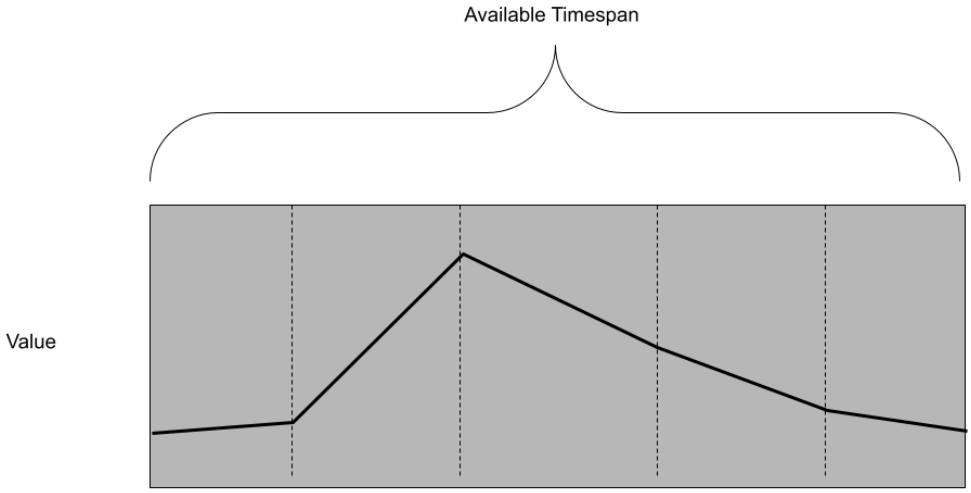


Figure 4.5: Hypothetical predictor highlighting variability
This figure highlights the advantage of breaking a time-series into chunks in order to summarize.

4.5 Existing Time-Series Preparation Tools

Existing time-based data preparation tools in R fall along a spectrum of two extremes; low-level feature extraction tools that focus on representing time-series or lagged variables and forecasting packages that primarily aim to model trends in a single variable over time (e.g. stock market) rather than separate predictions for separate patients. The tsibble package in R is intended to represent time-series data in a “tidy” format through time-based indexing. The slider, and runner R packages provide low-level functionality for creating different types of lagged predictors, with some functionality for summarizing multiple values. Only runner includes the concept of creating a “grid” of time points for each set of observations within a group (e.g., separately for each hospitalization) with the “at” argument, but this does not account for varying length of hospitalizations.

Several forecasting packages exist in R, including fable, forecast, and modeltime, but these are generally designed to model time-varying components by decomposing data into different components (e.g., overall trends, seasonality, etc.). While these may be useful for modeling signals with a high degree of regularity and/or seasonal components (e.g., stock market).

The closest software package for preparing time-series patient data to our intended use case is FIDDLE (in Python) [4], which aims to automate data preparation with minimal parameters, approaching parameterless functionality [81]. A parameterless, data-driven approach has its own advantages with respect to ease of use. The resulting dataset, however, may not have face validity to domain experts due to not taking into account important domain-specific information (such as medication half-life as described above). Inpatient care is inherently complex, and while a data-driven approach alleviates human-decision making, ignoring attributes in the underlying data can lead to loss of important information.

Despite the importance of data preparation on the results, it is uncommon for code used to generate features to be publicly shared as this is written specifically for a given dataset, and so not viewed as generalizable, and because it is often deemed as less important than the underlying modeling itself. Thus, there is a need for reproducible, domain expert-driven data preparation methods that can be readily reproduced and shared.

4.6 The wizard Package

In this paper we present wizard, an R package intended to transform clinical data into a modeling-ready dataset. Our primary design goal to simplify the data preparation process was to embrace clinical and data expertise through a shared grammar language. While this can be contrasted to data driven approaches in frameworks of similar purpose, our approach sought to preserve interpretability in data preparation and modeling decisions. The grammar is conceptualized using verb-like functions with easily interpretable parameters for both predictors and outcomes. It then builds a data structure in time sequence relative to an anchor point common in every observation (i.e. admission). It supports both single occurrence outcomes (i.e. mortality) as well as rolling outcomes which can occur multiple times throughout the observable period. Wizard builds on concepts related to discrete-time survival analysis commonly used to make multiple predictions based on updated information, autoregression focusing on a sequence of measurements, and statistical summaries used to represent a span of time.

Clinical prediction modeling often starts as a conversation between domain and data experts with a common goal of turning raw data into predictors known to have a relationship with an outcome. Approaching data structure as a design problem can alleviate this problem of translating time-based features derived from expert clinical knowledge to human-readable features. Wizard aims to capture that conversation between clinicians and modeling designers, distilling components into interpretable arguments at a precision level not achievable with current autonomous pipeline tools.

In wizard, we identify common feature engineering techniques for both predictors and outcomes and distill them to a consistent and human-readable set of components that can be used as parameters through a shared language. We first identify timespan based techniques with examples on how they can be useful. We then identify concepts related to left-censoring data we call lookback which informs the model how far into the past to include measurements. We then identify the interval with which predictors are to be aggregated to allow multiple sequential summarizations which can hypothetically capture a state-based trajectory. And finally, we support multiple summarization techniques.

Wizard provides a suite of functions intended to transform clinical data. Data collected at different intervals and periods may have distinct relevance to the current patient state. To account for variabilities, predictors can be timespan defined as 1) baseline, preceding any predictions and can be described as historical data; 2) rolling, which only includes a certain period in the past moving forward with updated predictions; or 3) growing, which starts at the first prediction and continues to expand when new information is received. Each timespan can then optionally be subdivided into intervals, focusing on a state-based representation of individual predictors. For example, a baseline predictor which goes one year in the past can be subdivided into three-month intervals which can preserve the progression of a condition. Finally, each timespan can then be summarized by using standard statistical methods such as min, max, mean and order-based methods such as first and last to capture trajectory. Embracing principles which support reusable and interpretable data structures has proven successful in other frameworks and in this work we aim to bridge tidy data principles with time-series data structure [82].

4.6.1 Wiz Frame Object: A Data Structure for Fixed and Time-Series Data

The wiz_frame() function

The wiz_frame function creates a wiz_frame object. The object contains a copy of fixed data (in “wide” format as in Table 4.2) which contain time-invariant data that are used persistently for every prediction. These are typically data which do not change over the series of individual observations (i.e. age, sex). It also contains the temporal data (in “long” format as in Table 4.3) using a subject identifier, timestamp (as a date-time or a number), and a variable name and value (either categorical or numeric). Additional columns such as category can be used to capture related groups of variables (e.g., all vital signs).

```
wiz_frame = function(fixed_data,
                     temporal_data,
                     fixed_id = 'id',
                     fixed_start = NULL,
                     fixed_end = NULL,
                     temporal_id = 'id',
                     temporal_time = 'time',
                     temporal_variable = 'variable',
                     temporal_value = 'value',
                     temporal_category = temporal_variable,
                     step = NULL,
                     max_length = NULL,
                     output_folder = NULL,
                     create_folder = FALSE,
                     save_wiz_frame = TRUE,
                     chunk_size = NULL,
                     numeric_threshold = 0.5)
```

4.6.2 fixed_start: Anchor Point

The wiz_frame keeps track of when the first prediction takes place, T_0 (Figure 4.6). This can be a specific event timestamp or index and the default is the early data available. It is used as the reference point for the rest of the variables.

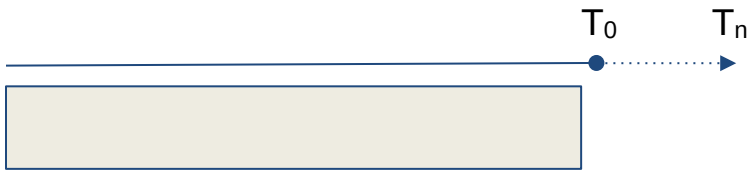


Figure 4.6: Fixed start

4.6.3 fixed_end: The Final Prediction

End_time identifies when the final prediction should be T_n (Figure 4.7). For survival analysis, it should be the early of either time of event or when the last prediction desired should be (i.e. discharge).

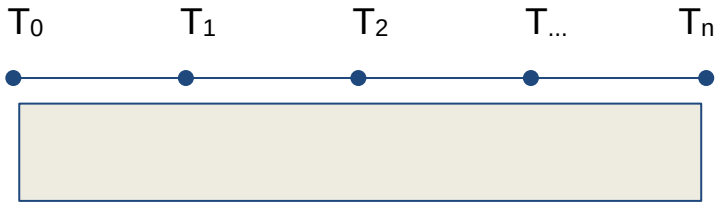


Figure 4.7: Fixed end

4.6.4 Step: Prediction Time Interval

The step is the time step expressed in either time (i.e. hours) or absolute numbers describing how often new predictions should be made (Figure 4.8).

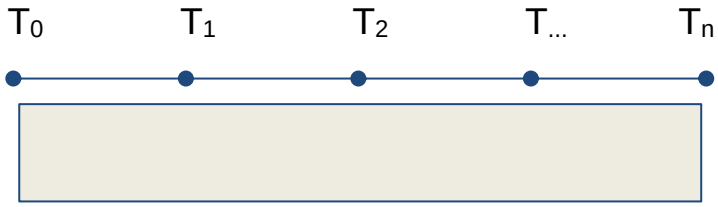


Figure 4.8: Prediction time step

4.6.5 Initial Processing of Data by wizard Upon Creation of a wiz_frame

Wizard converts timestamps to an index anchoring at either the first value per subject or a specific timestamped event found in fixed_data and based on the prediction interval (step). Wizard then summarizes measurements grouped by index into either statistical or order representations (Figure 4.9). Wizard then borrows methods from discrete-time analysis in which data are temporally separated into person-period steps. Each step can be considered a separate prediction so there can be multiple predictions

per subject [83]. However, Figure 4.10 shows when transforming the data from long (database format) to wide (person-period) we lose information in previous prediction observations or “states” of the subject. A solution to this is incorporating autoregressive techniques which group column descriptors both by variable and sequence. This method allows multiple states to be observed per prediction observation (each row).

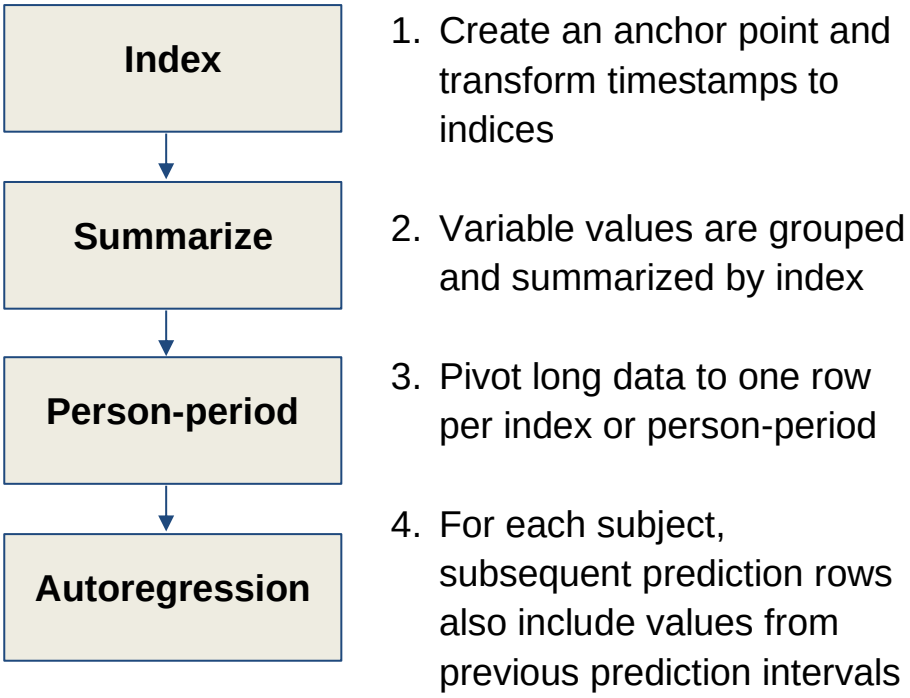


Figure 4.9: Raw data transformation

Subject	Time	Label	Value
2	0	Systolic BP	120
2	0	Diastolic BP	80
2	0	Respiratory Rate	12
2	1	Systolic BP	130
2	1	Diastolic BP	95
2	1	Respiratory Rate	20
2	2	Systolic BP	129
2	2	Diastolic BP	100
2	2	Respiratory Rate	21
2	3	Systolic BP	131
2	3	Diastolic BP	92
2	3	Respiratory Rate	20

Subject	Time	SBP	DBP	RR
2	0	120	80	12
2	1	130	95	20
2	2	129	100	21
2	3	131	92	20

Figure 4.10: Long to wide format data transformation

Subject	Time	SBP_1	DBP_1	RR_1	SBP_2	DBP_2	RR_2	SBP_3	DBP_3	RR_3
2	1	130	95	20						
2	2	130	95	20	129	100	21			
2	3	130	95	20	129	100	21	131	92	20

Figure 4.11: Autoregressive transformation

4.6.6 Categorical Dummy Coding

```
wiz_dummy_code = function(wiz_frame = NULL,
                          numeric_threshold = 0.5,
                          variables = NULL,
                          save_wiz_frame = TRUE)
```

This function automatically converts categorical variables of type character to dummy coding. In effect, it creates a binary variable for every category in each variable. By default this function saves the results back to the wiz_frame RDS file stored on disk. Variables can either be defined or wizard will automatically dummy code all non-numeric variables as defined by the data dictionary.

4.6.7 Feature Types

Predictors can be defined by invoking add_predictor(), add_baseline_predictor(), or add_growing_predictor() with a “wiz_” prefix. The difference between the three relates to their use of left and/or right censoring which can be described as temporal data filtered prior to the desired timespan (left censoring) or temporal data filtered after the desired timespan (right censoring). Predictor spans can then be broken into smaller components (window) which capture different states of a patient.

4.6.8 Adding Baseline Predictors

```
wiz_add_baseline_predictors = function(wiz_frame = NULL,
                                       variables = NULL,
                                       category = NULL,
                                       lookback = lubridate::hours(48),
                                       window = lookback,
                                       offset = lubridate::hours(0),
                                       stats = c(mean = mean,
                                                 min = min,
                                                 max = max),
                                       impute = TRUE,
                                       output_file = TRUE,
                                       log_file = TRUE,
                                       check_size_only = FALSE,
                                       last_chunk_completed = NULL)
```

Baseline predictors are variable measurements which occur prior to the first prediction (Figure 4.12). This has a computational advantage over other predictor types because they are calculated once for the duration of predictions. Parameters include lookback period which is the timespan into the past from the first prediction (T_0) to include. Lookback can also be optionally divided into smaller pieces via window. Offset can be used to move the timespan to the left of the first prediction (T_0).

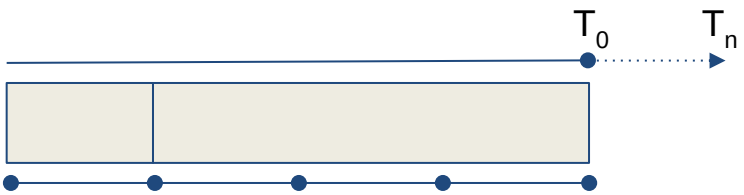


Figure 4.12: Baseline predictors

4.6.9 Adding Growing Predictors

```
wiz_add_growing_predictors = function(wiz_frame = NULL,
```

```
variables = NULL,
category = NULL,
stats = c(mean = mean,
          min = min,
          max = max),
output_file = TRUE,
log_file = TRUE,
check_size_only = FALSE,
last_chunk_completed = NULL)
```

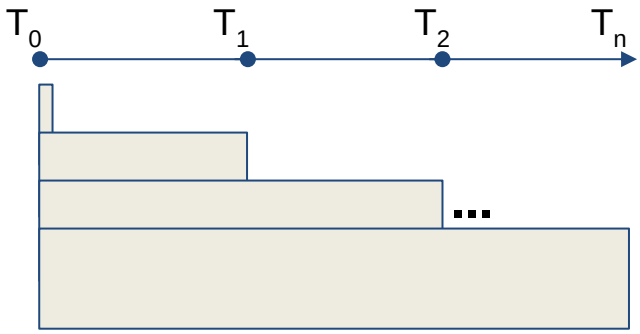


Figure 4.13: Growing predictors

Growing predictors can be seen as a complement to baseline predictors (Figure 4.13). They start at the first prediction and continue to expand with each prediction interval to include new data points.

4.6.10 Adding Rolling Predictors

```
wiz_add_predictors = function(wiz_frame = NULL,
                              variables = NULL,
                              category = NULL,
                              lookback = lubridate::hours(48),
                              window = lookback,
                              stats = c(mean = mean,
                                        min = min,
                                        max = max),
                              impute = TRUE,
                              output_file = TRUE,
                              log_file = TRUE,
                              check_size_only = FALSE,
                              last_chunk_completed = NULL)
```

Rolling predictors have a fixed width timespan while stepping forward with each new prediction (Figure 4.14). Lookback period is the timespan into the past of the first and all subsequent predictions. The lookback window can be further divided into smaller intervals using the window parameter. Carry forward imputation is used by default but can be disabled.

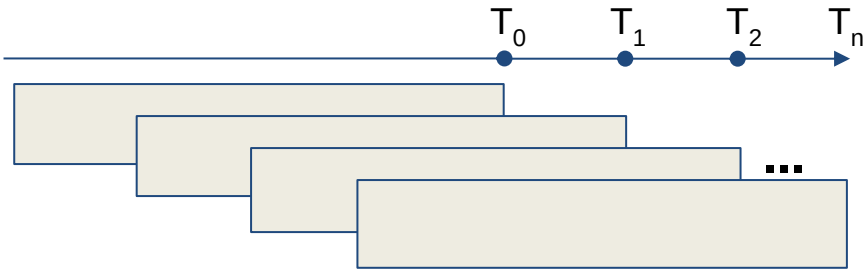


Figure 4.14: Rolling predictors

4.6.11 Adding Fixed Outcomes

An outcome that occurs at a time in the indefinite future either once (i.e. death) or only the first occurrence is of interest (i.e. shock). In these outcomes, data after the outcome has occurred must be censored so no subsequent predictions are made. While the event in question may be predicted more than once, a repeated event cannot be predicted like it can with a rolling outcome. Fixed outcomes can be included directly in the fixed_data, and if no predictions should be made after the outcome has occurred, the timestamp for the outcome can be taken into account when setting the fixed_end argument in wiz_frame().

4.6.12 Adding Rolling Outcomes

```
wiz_add_outcomes = function(wiz_frame = NULL,
                            variables = NULL,
                            category = NULL,
                            lookahead = lubridate::hours(48),
                            window = lookahead,
                            stats = c(mean = mean,
                                      min = min,
                                      max = max),
                            impute = FALSE,
                            output_file = TRUE,
                            log_file = TRUE,
                            check_size_only = FALSE,
                            last_chunk_completed = NULL)
```

Outcomes behave in a similar fashion to predictors except instead of looking prior to the prediction, it considers the future timespan defined by lookahead. Wizard supports multiple outcomes and can be used for survival analysis in which the outcome can occur at most once or a rolling outcome in which the lookahead slides forward along with subsequent predictions (Figure 4.15).

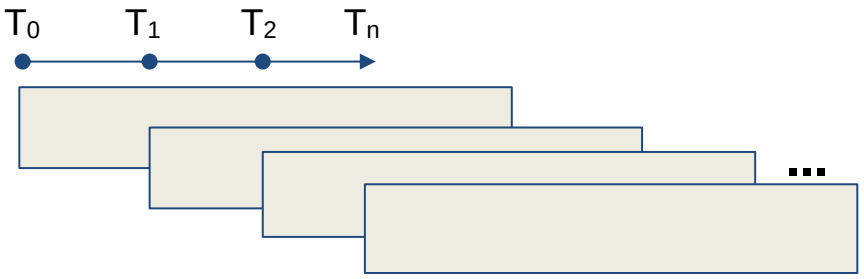


Figure 4.15: Rolling outcomes

4.6.13 Predictor Type Overview

Choosing the right predictor type can depends on the clinical relevance of the variable and timespan. Figure 4.16 illustrates each predictor type and how they are related to the initial at-risk or first prediction period, T_0.

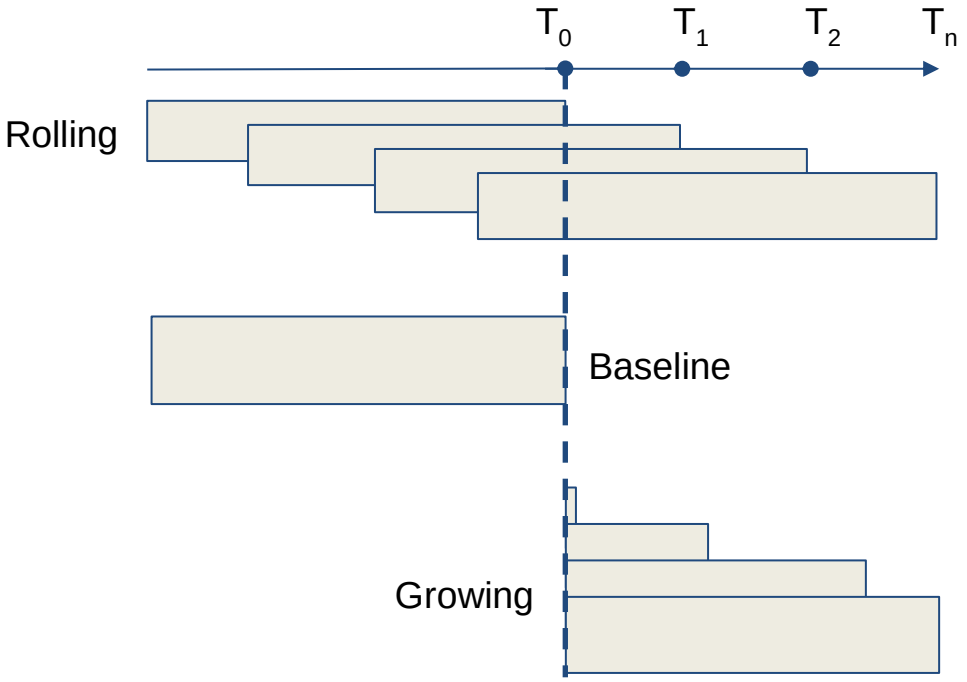


Figure 4.16: Predictor type comparison

4.6.14 Statistical Summaries

Statistical summaries for each of the previously mentioned functions can be used to override defaults of mean, min, and max. Additional options include median, length (count of occurrences), as well as first and last values found in the summarize function of the dplyr package (e.g. dplyr::first or dplyr::last) [84].

4.7 Evaluating wizard on Public Benchmarks

4.7.1 Data Source

To demonstrate performance of our framework we identified an existing data pipeline framework applied to a de-identified openly accessible clinical dataset. We found the FIDDLE framework created by Tang et al. to closely resemble our framework in purpose: to transform long form clinical data into a dataset ready for supervised machine learning [4]. Among the demonstrations Tang et al. published, they used the MIMIC-III dataset to predict a group of commonly tested outcomes [85]. We used a portion of the code published on GitHub by Tang et al. to build the cohort datasets which took place prior to FIDDLE transformation [86].

4.7.2 Study Cohort

We chose to exclusively use data from MIMIC collected using MetaVision (2008-2012), similarly to the FIDDLE implementation, because of its more recent collection than CareVue (2001-2008). We divided the cohort into train, tune, and test cohorts by patient first and then by ICU admission to preserve patient separation between cohorts. The resulting task cohorts were between 14,355 and 21,044 ICU encounters and XX and XX total patients. Encounters were excluded for subjects under 18, when the subject died prior to the prediction of interest, and if the outcome occurred prior to the prediction of interest (Figure 4.17).

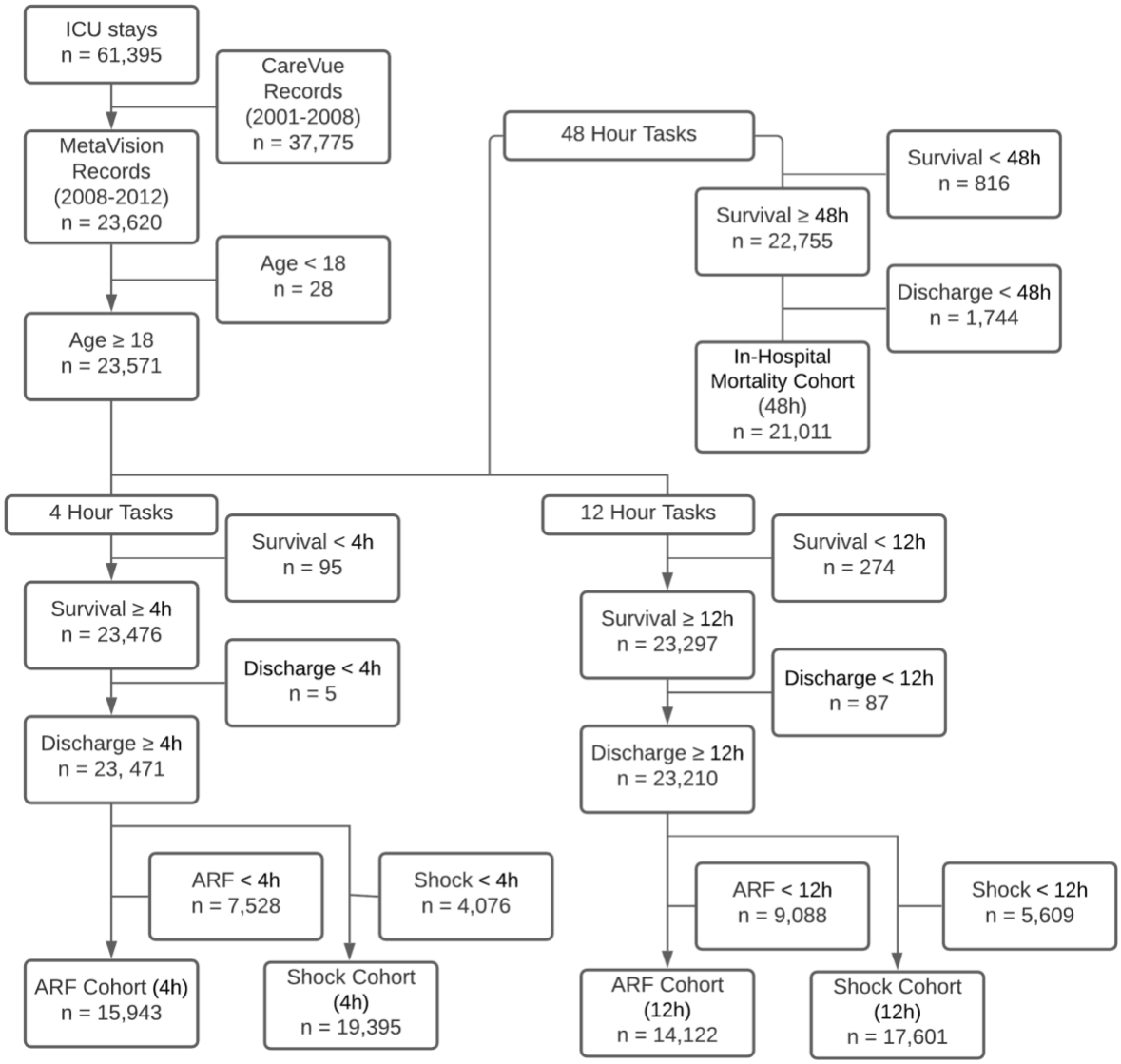


Figure 4.17: MIMIC-III case study cohort

4.7.3 Outcomes

A total of three outcomes and five outcome tasks were conducted. A single prediction was calculated for each outcome task at a fixed time following admission to the ICU using predictors collected between ICU admission and the time of prediction. Outcome tasks conducted were acute respiratory failure (ARF) for the next 4 and 12 hours (Invasive Ventilation, Non-invasive Ventilation, or PEEP set), shock described as the use of vasopressors (Norepinephrine, Epinephrine, Dopamine, Vasopressin, Phenylephrine), and in-hospital mortality at 48 hrs following ICU admission for the remainder of admission.

4.7.4 Predictors

While similar cohorts were used, RARS used a different transformation methodology to prepare predictors. Although only benchmark values are tested and reported, predictions were made every four hours and recursively used to train the model. Imputation was not conducted because h2o supports missingness by making decision splits based on the branch that most closely resembles an existing value using other variable splits. Predictors were excluded from analysis when there was less than 5% prevalence among patients. Numeric predictors were split into binary (two or less unique values) and non-binary (more than two unique values). Categorical predictors were dummy coded. Numeric non-binary variables were summarized at a four hour interval capturing minimum, maximum, median, and length (count of occurrences). Numeric binary and categorical variables were dummy coded when necessary, summarized at a four hour interval and summarized by length, capturing both the presence and number of occurrences (Table 4.4).

Table 4.4: wizard parameters

Variables	**timespan**	**Lookback**	**Stats**
Continuous	growing	NA	first, last, min, max, median
Continuous, Dichotomous, Categorical	growing	NA	length

4.7.5 Model Development

Gradient boosting machine (GBM) was used due to its ability to make corrections based on previous predictions made and its performance using imbalance outcomes. Since our transformation methods include recursion, a strength is including variable summaries from previous predictions, generally described as autoregression. We used the train and tune cohorts including all recursive predictors up to and including the time points of interest. This allowed previous predictions to be used in model corrections. We used the test cohort filtered to the prediction and outcome specific to each task. We used 1000 trees, a learning rate of 0.01, and AUC as the stopping metric.

4.7.6 Model Validation

Discrimination was calculated using a single prediction per hospitalization occurring at time T hours following ICU transfer or admission. Model calibration was assessed using a calibration curve comparing deciles of all predictions to the observed risk.

4.7.7 Statistical Software

R 4.0 was used to conduct all analyses. We used the h2o R package (v3.30) for the gradient-boosting decision tree and the wizard R package to transform the data. Discrimination, calibration, and threshold performance were visualized using the runway R package. In addition to statistical software used, Yottabyte research cloud, Great Lakes (HPC Cluster), and ARMIS (HIPAA-aligned Slurm Cluster) were used for computational needs [87]–[89].

4.7.8 Results

Three outcomes including five prediction tasks were conducted using wizard to prepare the data and GBM to build the models (Table 4.5). We found AUCs for each task: in-hospital mortality, 0.879, ARF at 4 hours, 0.786, ARF at 12 hours, 0.717, shock at 4 hours, 0.814, and shock at 12 hours, 0.781. AUC values for FIDDLE and MIMIC-Extract are ranges reported from Tang et al. for all algorithms used [4], [90].

Table 4.5: Outcome prevalence, cohort size, and discrimination comparison across outcomes

Outcome (Prevalence)	Cohort Size	Prediction Task: ICU Admit + T	AUC		
			wizard	FIDDLE	MIMIC-Extract
Mortality (8.7%)	21,011	48hr	0.879	0.814-0.886	0.837-0.859
ARF (18.2%)	15,943	4hr	0.786	0.817-0.827	0.777-0.821
ARF (9.5%)	14,122	12hr	0.717	0.757-0.771	0.700-0.747
Shock (14.9%)	19,395	4hr	0.814	0.809-0.831	0.796-0.824
Shock (7.7%)	17,601	12hr	0.781	0.773-0.792	0.741-0.778

Note: ARF: acute respiratory failure; AUROC: area under the receiver operating characteristics curve; RARS: Recursive Autoregressive Summarization; FIDDLE: Flexible Data-Driven Pipeline.

4.8 Discussion

Trust has long been a subject of contention in machine learning models. While machine learning models often outperform statistical or rule-based models, transparency often comes as the trade-off. However, a distinction must be made about where the focus is commonly placed in the predictive modeling process. There are in fact two separate trust issues: trust in the modeling input, often referred to as prepared data, and the output or model itself. As models get more complex, they get harder to understand and therefore harder to trust. However, our work does not apply to that as it is intended to be model agnostic, from simple linear regression to neural networks.

It is equally important to establish trust in the data used for modeling. Many approaches make data transformation decisions that make the predictors incomprehensible before even applying an algorithm to them (e.g., log transformation, normalization, tensors). We, on the other hand, are helping clinicians understand what a given predictor means. Our goal is that the results can be used in any modeling approach and an extension of this work would allow more transparency in the model as this is a prerequisite to establishing trust. Therefore, by focusing on a model's input, we believe we are addressing one of the two problems related to trust in predictive modeling.

Our framework has shown favorable results transforming data intended for early warning systems to predict postpartum hemorrhage. However, we expect that it would be helpful to apply it to more kinds of problems beyond just maternal care. Additionally, we intend to expand use cases beyond what is currently contained in the package to include more preprocessing steps with framework improvements. This includes visualization tools to help decide to represent the data as well as automated techniques when domain expertise isn't available using statistical analysis.

4.9 Conclusion

The wizard R package is a software program that encodes a grammar of data preparation for EWS models through windowing and summarization for auto-regressive data preparation. Using wizard, analysts can effectively and efficiently prepare electronic health record data for EWS modeling through a framework that directly leverages domain

expertise. Wizard provides a framework with interpretable parameters and human-readable output ready for use with any prediction modeling algorithm. Wizard supports parallel processing using the future package and can process data in chunks to reduce the need to keep the entire temporal dataset in-memory when used on systems with limited resources.

Application of wizard to MIMIC-III data was intended to be a proof of concept with the goal of making sure it worked on a large dataset without any prior extensive exploratory analysis often required in manual data preparation. Whether the results are statistically significant or not is immaterial to the fact that wizard was able to successfully transform the MIMIC-III data effectively with performance in close-range of those of previous attempts using automated data preparation tools.

Chapter 5
Continuous Prediction of Postpartum Hemorrhage Using Time-Series Machine Learning Models

This chapter was co-authored with Alissa Carver, Hyeon Joo, Thomas Klumpner, and Karandeep Singh.

5.1 Introduction

In the United States, postpartum hemorrhage is a leading cause of preventable maternal mortality, which is often due to delayed or inadequate response to clinical warning signs [10], [91], [92]. While postpartum hemorrhage is commonly diagnosed by methods to quantify blood loss, bleeding may be concealed [93], [94]. Therefore, vital sign surveillance is an important mechanism by which postpartum hemorrhage is identified [95], [96]. To facilitate early detection of maternal morbidity like postpartum hemorrhage, the National Partnership for Maternal Safety has recommended institutions adopt maternal early warning systems, which use a set of vital sign thresholds as triggers to escalate care [97]. Simplifications of these systems, like the Maternal Early Warning Criteria, have been endorsed to facilitate recognition of actionable conditions at the bedside [97]. Unfortunately, simplified systems such as these lack important patient-specific and contextual information that should probably alter notification thresholds. For example, some vital sign abnormalities may be the result of more benign conditions like maternal expulsive efforts [8], [9]. In addition, proper use of these systems requires a lack of hesitation to escalate care, good judgment, and is subject to other cognitive biases [35]. Even if a vital sign abnormality is registered by the bedside monitor, the bedside provider must recognize its severity and escalate care without reservation.

These limitations have led to efforts to both identify better predictors of maternal morbidity using mathematical modeling techniques [37], [98], and also to automate

notification of providers using risk scores [9]. A model was recently developed by Venkatesh et al to predict postpartum hemorrhage using data available at the time of admission [98]. However, this model does not appear to generalize to contemporary practice settings that quantify blood loss using QBL. More recently, Escobar et al. [37], described the development of a model to predict a wide range of adverse maternal and fetal outcomes using an early warning system. While the model performed well (C-statistic 0.79), virtually all alerts were generated in the antepartum period, highlighting its limitation as a system for capturing postpartum hemorrhage. Finally, we have previously described the development of an automated maternal early warning system which while widely accepted by providers [99], has a positive predictive value of only 5.1% for the detection of severe postpartum hemorrhage and does not account for patient-specific baseline risk factors, other than stage of labor, when issuing an alert [8].

Recognizing these shortcomings, and the unique challenges inherent in detecting postpartum hemorrhage, we sought to develop a continuous prediction model focused solely on the early identification of postpartum hemorrhage using a novel time-series machine learning approach.

5.2 Materials and Methods

5.2.1 Data Source

The University of Michigan Von Voigtlander Women's Hospital is a tertiary care academic women's hospital with approximately 4,600 deliveries per year.

5.2.2 Study Cohort

Our study included women aged 18 or older, hospitalized for delivery between February 1, 2019, and May 11, 2020, including deliveries for fetal loss. Deliveries were excluded if the estimated gestational age was less than 22 weeks, the minimum gestational age at which neonatal resuscitation is offered at our institution. This study was approved by the University of Michigan Institutional Review Board (IRB), which waived the requirement for informed consent.

Maternal demographics, comorbidities, and delivery data were collected from the electronic health record (EHR). Time-varying data were also collected from the EHR, including obstetric care (e.g., Cesarean section versus vaginal delivery), nursing flowsheets, telemetry data, lab results, active problem lists, and medication administration. Data collected spanned from one year prior to delivery admission to the time of discharge.

The primary outcome was defined as postpartum hemorrhage with a cumulative quantitative blood loss (QBL) of ≥ 1000 mL occurring between delivery and discharge. Blood loss prior to delivery was not included in this calculation. We selected 105 variables based on those with known relationships to postpartum hemorrhage led by clinical experts. Data was sourced from multiple clinical information systems and merged based on advice from clinical domain experts. For example, data from the anesthesia information management system were combined with inpatient data with input from authors T.K. (an anesthesiologist) and A.C. (an obstetrician). Data were divided into three temporal categories: time-invariant, described as variables which are unlikely to change over the course of admission (i.e. age, biological sex, marital status, gravidity parity), baseline variables, which are collected prior to the first prediction (i.e. diagnostic conditions, body max index (BMI), weight), and growing predictors, or events for whom all values were considered from the first prediction up until discharge or the event, whichever occurred first (i.e. medication administration, lab results, and vital signs). A full list of predictors and their temporal properties captured can be found in Supplemental Table 5.5.

5.2.3 Model Development

The wizard R package was used to transform row-level patient data into windowed summaries (Chapter 4). This method of data transformation builds indices from time stamped data around the first prediction of interest, in this case, the delivery time. The indexed data are then grouped and summarized in autoregressive windows of time. Recursive observations are calculated for each prediction so that each subsequent hospitalization prediction contains the data from the previous observations. Windows with

repeated numeric measures were then summarized using the first, last, minimum, maximum, mean, median, and number of occurrences within each window.

Deliveries between February 1st, 2019 and November 15th, 2019 consisting of 64% were used to train the model. Deliveries between November 16th, 2019 and February 1st, 2020 were used to tune the model, and the remaining 20% which occurred between February 2nd, 2020 and May 11th, 2020 were used to test the model.

Gradient boosting machine (GBM) was used due to its ability to make corrections based on previous predictions made and its performance using imbalanced outcomes [75]. Since our transformation methods recursively use information from prior time points to predict outcomes at future time points, this approach is a generalization of the discrete time survival analysis approach. We used the train and tune cohorts including all recursive predictors up to and including the time points of interest. This allowed previous predictions to be used in model corrections. We used the test cohort filtered to the prediction and outcome specific to each task. We used 1000 trees, a learning rate of 0.01, and AUC as the early stopping metric.

The model was trained on an outcome to predict whether postpartum hemorrhage would occur anytime later in the hospitalization, referred to as the hospitalization-level outcome. This allowed decision trees to focus on the differences in overall outcomes between patients (hospitalization-level outcome) rather than solely focusing on changes within individual patients (rolling outcome), which can lead to overemphasis on time-varying predictors and an underemphasis on baseline risk factors. We predicted postpartum hemorrhage starting at the delivery timestamp, T_0, and every 20 minutes ending at the time of discharge.

5.2.4 Model Validation

Area under the receiver operating characteristic curve (AUROC) and calibration were used to assess model performance. AUROC, also known as discrimination, is the ability for the model to classify risk groups across all thresholds [100]. Calibration is the comparison of predicted and actual risk probabilities [66], [67].

Model performance was first tested as the data were presented to the algorithm. This consisted of multiple predictions per patient with the last consisting of the entire

hospitalization right-censored to the outcome, also called the person-period prediction level described in discrete-time survival analysis. This measure cannot be compared across studies because it is dependent on the number of predictions made across hospitalizations (e.g. prediction interval, first prediction timepoint, and duration of hospitalizations). The second method consisted of the first high-risk alert per hospitalization in relation to the outcome. This is more comparable across studies as long as the outcome is the same. Other performance measures used to determine a threshold are sensitivity, specificity, positive predictive value (PPV), number needed to evaluate (1/PPV), and number willing to evaluate which can be used as a way to assess whether a model is a good fit over no model at all in a practical setting.

A lead time analysis was conducted to assess the distribution of hypothetical alert lead time prior to the outcome. In this assessment, the first prediction in the test set which exceeded the high-risk threshold (probability ≥ 10%) for each encounter was used to calculate the time difference between when the selected prediction occurred and when the patient experienced PPH.

5.2.5 Missing Data

Missing values were handled directly by the gradient-boosting decision tree algorithm. During model training, optimal binary splits were determined by minimizing the error using non-missing data. After a variable split was determined, missing values for that variable were assigned to the direction minimizing the error. When generating predictions, missing values followed the assigned direction [101].

5.2.6 Statistical Software

R 4.0 was used to conduct all analyses. We used the h2o R package (v3.30) for training the gradient-boosting decision tree models. Discrimination and threshold performance were visualized using the runway R package. C-statistic confidence intervals and comparisons were calculated with bootstrapping (2,000 replicates for 95% confidence intervals) using the pROC R package. Calibration was calculated using the rms R software package [102]. In addition to statistical software used, Yottabyte research

cloud and ARMIS (HIPAA-Aligned Slurm Cluster) were used for computational needs [87], [88].

5.3 Rationale for Study Design

5.3.1 Data Acquisition

Data acquired for this study is the same source data used in Chapter 3. See Section 3.3.1 for details on methodological decisions and the rationale supporting those decisions.

5.3.2 Preparation

The primary differences between our validation study in Chapter 3 is taking a time-series approach to clinical prediction modeling. This consisted of careful consideration around what constituted the at-risk period as well as the interval with which to make predictions. We determined the at-risk period to start at the point of delivery because postpartum hemorrhage can only occur after delivery. While another option was to start predictions at the time of admission, which was the single prediction point for our validation model, we found that performance was similar between using admission and delivery as the starting point, and the predictors between admission and delivery added complexity to the model. Moreover, the actionability of predictions made prior to delivery are not clinically intuitive and less likely to be useful as predictions issued during the postpartum period.

The highest frequency with which to make predictions could be every time a new measurement was collected. This would likely be infeasible because no mechanism exists in our current electronic health record infrastructure to run a model only when new information becomes available. We chose a twenty-minute time step because Klumpner et al. found in that the median time between an alert and intervention was for the most severe PPH cases was 48 minutes [8]. To ensure adequate lead time to act, we chose to divide this amount of time in half and round down to regularize the interval, for example, so alerts would be fired at the same time on the hour. If the selected time step was too long, there likely would not be enough time to take action for a severe PPH case. This

approach is fundamentally different from the approach taken by Venkatesh et al. where only one prediction was issued at admission.

Over the course of two years, I filled the role as lead data scientist throughout this project. Our research team included three clinical experts who met on a weekly basis to discuss potential predictors to be used in the model. We also discussed in an iterative manner as to what the different options and determined how to optimally represent each variable included in the model.

We discussed the parameters to be used with the wizard framework in terms which could be communicated and understood by the entire team (Supplemental Table 5.5). For example, we determined that billing data would be best considered as baseline predictors that would only be considered prior to the first prediction. We also wanted to separate recent from historical events, so we approximated the third trimester into a 90-day lookback and split it into 30-day increments. We then summarized each window of time by calculating the number of occurrences which served two purposes: measuring the presence of a diagnosis as well as the number of occurrences.

There were other variables that required a different approach. For example, vital signs data were split into baseline and growing predictors because vital signs can change rapidly during postpartum hemorrhage. Identifying these changes as they evolve in the context of their normal values is clinically relevant. This meant summarizing all baseline values to establish a normal range of measurements on each patient. This also required using other methods to capture the “high frequency” component of vital signs interpretation using statistical and order summarizations (i.e., minimum, maximum, median, last, and first value) for each prediction window.

5.3.3 Measuring Performance

Another important consideration was the method in which we represented the outcome which can affect model performance. Performance may be measured at the encounter level in which model performance is evaluated based on the highest probability during the encounter. For patients who do not experience the outcome, a single high prediction will bring down the encounter-level performance even if the majority of

predictions are accurate (i.e., low). Therefore, at the encounter level, model performance is penalized if patients who do not experience the outcome ever generate an alert.

An alternative is to evaluate every prediction independently, in this case in twenty-minute increments. Since each prediction is considered in the calculation of model performance, one inaccurate prediction will not substantially affect performance. Both methods are relevant but different ways to measure model performance, and the differences between them can indicate gaps in model performance. From a clinical standpoint, the first alert may be the most important as clinicians will perceive subsequent alerts on the same patient as a false positive and are less likely to comply. Therefore, the encounter-level prediction likely matters most in the context of clinical practice.

5.3.4 Algorithms

There are many options to take into consideration when choosing which algorithm(s) to develop a model (summarized in Section 2.7). To take advantage of the flexibility non-linear modeling approaches have to offer we opted to use gradient boosting machines (GBM). GBM performs well when benchmarked against other algorithms [54]. GBMs work by creating decision trees iteratively with each subsequent tree focused on observations that were poorly predicted. They can perform well with a full grid search where hyperparameters, or options affecting algorithm behavior, (i.e., learning rate, tree depth, variable sampling rate) are tuned to find optimal settings, but this can be computationally expensive. GBM models were also chosen because, compared to linear models, they reduce the number of underlying assumptions about model data.

5.3.5 Assessing Performance

The definition of the outcome, rolling or persistent, can also affect performance. Using the same dataset, we initially benchmarked performance based on how the model was trained and tested. Table 5.1 shows how cross-testing models can result in differing levels of performance based on whether the training set and test set had the same or different representations of the outcome (rolling or persistent). Although difficult to discern, it is likely the algorithm is increasing importance of events which happen throughout the hospitalization leading to the outcome when persistent. Alternatively, using

a rolling outcome could force the algorithm to put greater weights on events which happen in the short term or immediately following delivery since predictions would continue beyond when a patient first experienced PPH.

To evaluate the differences when comparing outcome methods, discrimination was calculated using two models. The first used a rolling outcome across the entire dataset (A), and the second used the hospitalization outcome which was censored when the outcome occurred, or the patient was discharged (B). Each test set was then trained on both models on a rolling basis (2,3), a single prediction at the time of delivery indicating detection for postpartum hemorrhage on delivery (1), and the maximum prediction for each hospitalization which occurred prior to outcome or discharge (4).

Table 5.1: Model training benchmark

Dataset		A) Rolling outcome trained GBM	B) Hospitalization outcome trained GBM
Rolling outcome test set (24 hour lookahead)	Primary PPH[1] (n = 2,030)	0.6302 95% CI: 0.6007-0.6594	0.6371 95% CI: 0.6083-0.6662
	Rolling prediction[2] (n = 276,604)	0.9325 95% CI: 0.9241-0.9415	0.7506 95% CI: 0.7434-0.7577
Hospitalization outcome test set (right censored at outcome/discharge)	Rolling prediction[3] (n = 195,207)	0.7205 95% CI: 0.7107-0.7303	0.8551 95% CI: 0.8499-0.8602
	Maximum probability[4] (n = 2,030)	0.649 95% CI: 0.6217-0.6761	0.693 95% CI: 0.6664-0.7214

[1]single prediction at the time of delivery for 24 hours following (n = 2,030)
[2]each prediction is the probability of PPH in the next 24 hours (n = 276,604)
[3]each prediction is the probability of PPH anytime during hospitalization (n = 195,207)
[4]highest probability before outcome or discharge (n = 2,030)

The amount of time between when the first alert fired and when a patient experienced the outcome was calculated, also known as a lead-time analysis to assess hypothetical performance of the model. While this analysis is based on a specific high-risk threshold, its merit can only be validated using a prospective analysis.

Calibration was also evaluated. Calibration is a comparison of predicted probability of a set of points against the observed probability [66], [67]. This is commonly conducted by dividing probability distributions up into deciles. Each probability decile is considered

a sample of data. Assuming there are 10 points in the sample, let's assume that 70% of those samples experienced the outcome. Therefore, the fraction of positive points would be 0.7 (observed probability or y-axis). This means that samples within this range have a 70% probability of experiencing the outcome. Now plot the average of the probability estimates predicted by the model (predicted probability or x-axis) for each of those decile ranges in observed probability. In a well calibrated model, we would expect these values for each probability to be close.

5.4 Results

We identified 6,153 deliveries during the study period, of which 6,000 deliveries met the inclusion criteria (Figure 5.1). Out of the included deliveries 4,486 (75%) did not experience PPH and 1,514 (25%) experienced PPH. In comparing the temporally separated train, tune, and test cohorts, we found no significant differences between their baseline characteristics (Supplemental Table 5.3). In comparing across outcome groups, patients who experienced PPH had a higher prevalence of cesarean delivery, longer labor duration, higher rate of multiple gestation, and had more comorbidities including chronic hypertension and preeclampsia (Supplemental Table 5.4).

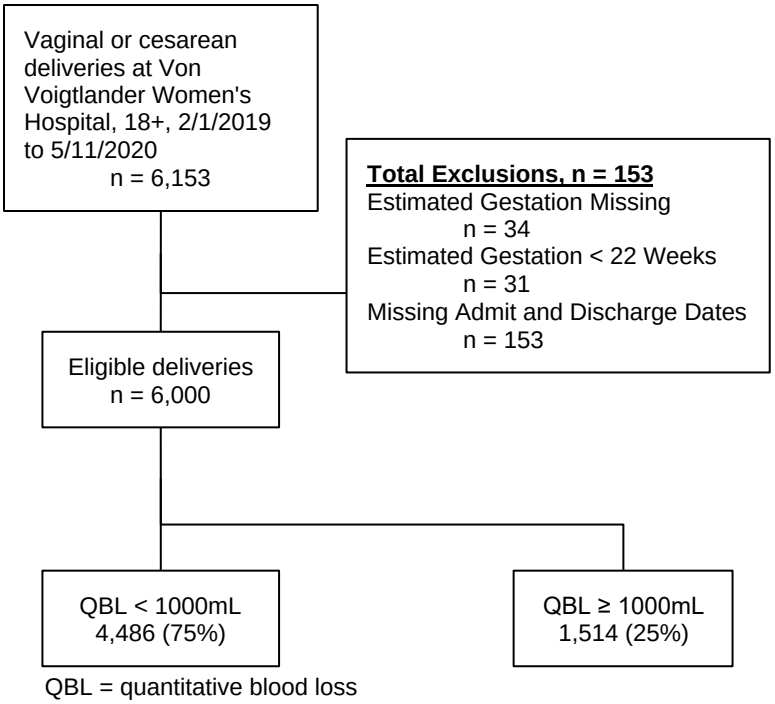


Figure 5.1: Cohort selection criteria
Out of 6,153 deliveries between 2/1/2019 and 5/11/2020, 6,000 contained the minimum amount of time-based data points to conduct a time-series analysis. From the 6,000 eligible deliveries, 1,514 (25%) experienced PPH while 4,486 did not experience PPH.

Area under the receiver operating curve (AUROC) was calculated using four methods but one training dataset/model (Table 5.2), (1) each prediction occurring at twenty-minute intervals was considered independently and (2) the patient's maximum risk was captured prior to the outcome or discharge. Method (1) yielded an AUC of the interval prediction in that predictions at each interval were considered until either when the outcome first occurs or the patient is discharged, 0.975 (95% CI: 0.972-0.977). The second method of evaluation was calculated by using the maximum risk predicted prior to when the hospitalization outcome first occurs or when the patient is discharged, 0.694 (95% CI: 0.659-0.727).

Two additional methods of discrimination calculations were conducted to evaluate (3) primary postpartum hemorrhage by definition occurring within 24 hours of delivery, 0.645 (95% CI: 0.607-0.682), and (4) a rolling outcome predicting postpartum hemorrhage for the next 24 hours at each prediction interval, 0.975 (95% CI: 0.972-0.977).

Table 5.2: Model discrimination

Evaluation Method	AUROC	95% CI
Persistent Outcome[1]	0.954	0.948-0.961
Maximum Probability[2]	0.694	0.659-0.727
Primary PPH[3]	0.645	0.607-0.682
Rolling Outcome[4]	0.975	0.972-0.977

[1]PPH occurring anytime following delivery to discharge
[2]Highest risk of PPH between delivery and either the outcome occurrence or discharge
[3]PPH occurring within the first 24 hours of delivery
[4]PPH occurring in the next 24 hours at each prediction interval

Note: Discrimination measured using the area under the receiver operating characteristic curve (AUROC) on the test set, accompanied by the 95% confidence interval (CI).

Calibration for the primary analysis was visualized using predictions at the interval level (Figure 5.2). Predictions at probabilities below 0.17 are being under predicted while probabilities above 0.17 are being over predicted.

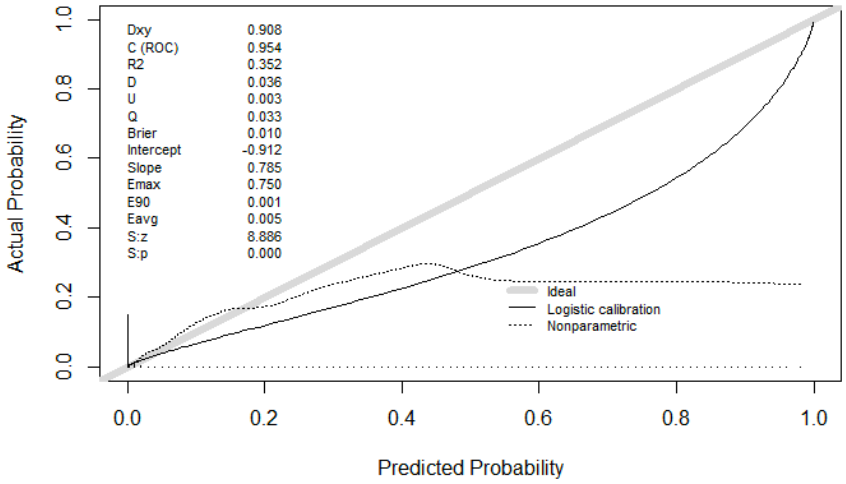


Figure 5.2: Interval level calibration plot
This plot compares predicted versus observed risk. The thick grey line indicates ideal calibration while the dotted line indicates observed nonparametric calibration.

A lead-time analysis based on the first alert was conducted on patients in the test set who exceeded a risk probability of 0.10 (Figure 5.3) and experienced postpartum hemorrhage. A threshold of 0.10 was chosen which is equal to approximately 20% specificity with consideration of the highest alert threshold during the encounter. However, determining a probability threshold cutoff is largely dependent on the context with which the model is used [103, Sec. 16.4]. The first alert which exceed the threshold of 0.10 was 60 minutes prior to postpartum hemorrhage on average (Figure 5.4), the median lead-time was 39 minutes, and interquartile range of 32 minutes.

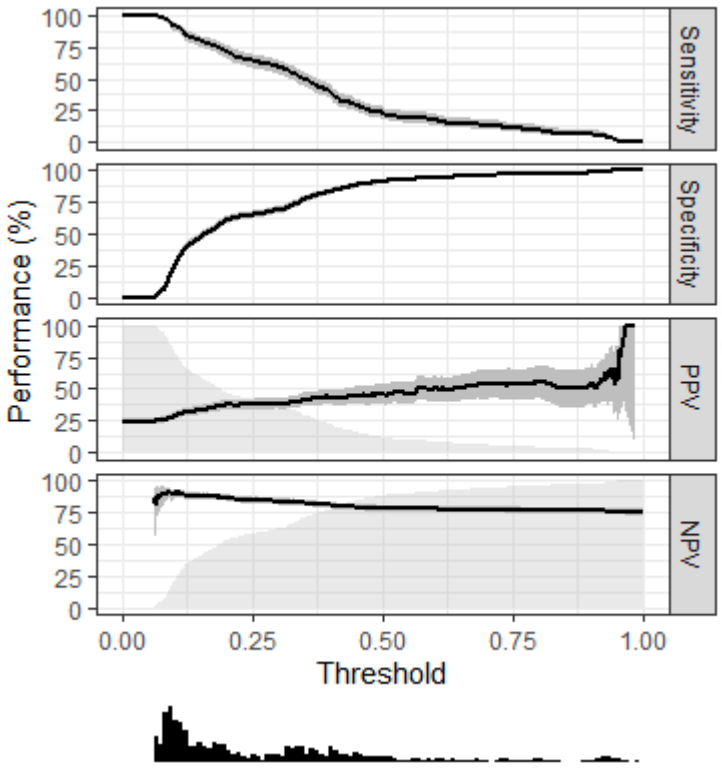


Figure 5.3: Threshold performance plot, maximum probability per hospitalization
Plot showing the relationship between the probability threshold and the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), with a histogram demonstrating the distribution of maximum probability per patient.

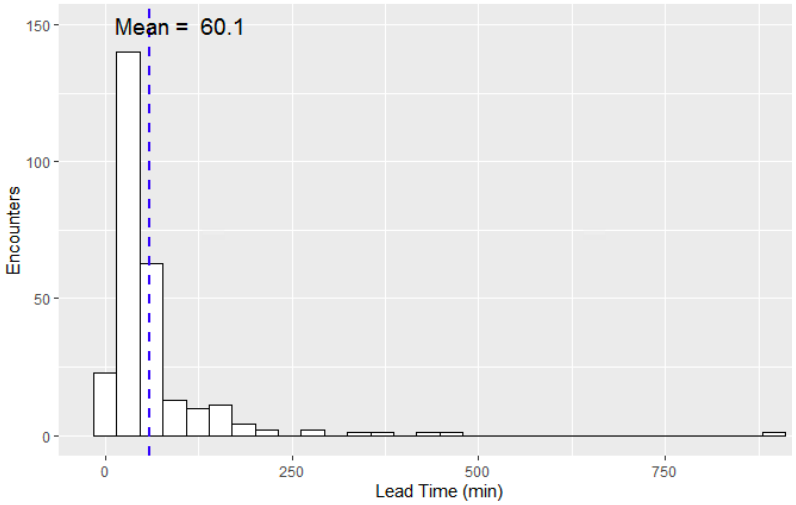


Figure 5.4: Distribution of alert times
Alert times are based on exceeding the high-risk threshold with a probability ≥ 10. Selected predictions from the test set accounted for 273 total encounters, with a mean high-risk alert time of 60.1 minutes, a median high-risk alert time of 39.2 minutes. Note: Each alert time represents a hypothetical alert; no actual alerts were generated.

Variable importance was calculated for the GBM using h2o [104], [105]. Relative variable importance shows that time, relative to the time of delivery, is the most important variable followed by cumulative QBL and then delivery method (Figure 5.5). Variable names are concatenated into the variable type, variable name, variable value, summary method, and lookback (if applicable). Baseline variables can be described as predictors which occur prior to or at the time of delivery while growing predictors occur at the time of delivery up to the current prediction interval. Summary methods are relative to either the lookback period for baseline variables, which in most cases is 2880 minutes (48 hours) or within the interval itself (20 minutes). More details on how the framework works can be found in Section 4.6.8.

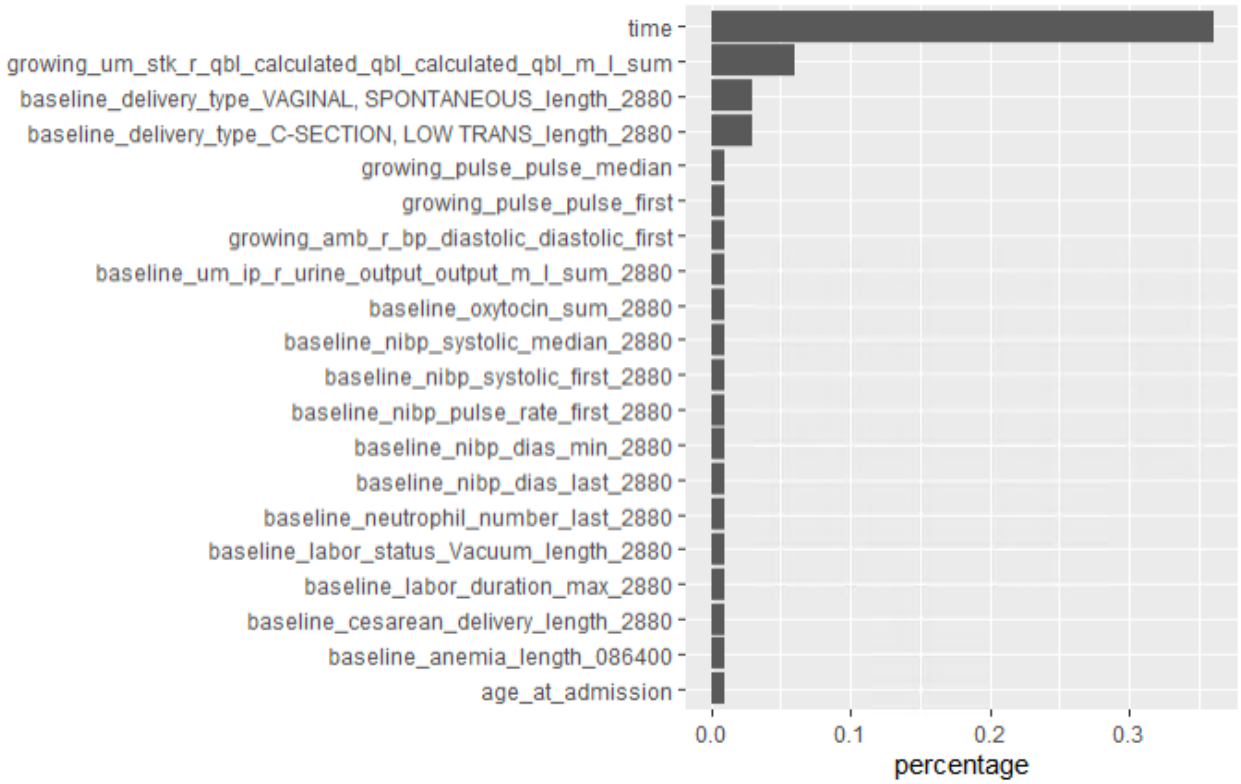


Figure 5.5: Relative variable importance

5.5 Discussion

In this study we developed a model to predict postpartum hemorrhage defined as ≥ 1000 mL quantitative blood loss using 6,000 delivery hospitalizations. Using 105 predictors we were able to predict PPH with strong time-horizon discrimination with an

AUC of 0.97 at the interval prediction level and a moderate AUC of 0.69 when considering the maximum probability per hospitalization.

Existing predictive models are not useful for predicting PPH in a real-time clinical setting. We believe this model to be useful in a clinical setting because previously published models either only made a single prediction at the time of admission or were not generalizable to institutions using QBL methods to estimate blood loss after delivery. One other model made multiple predictions per hospitalization using an outcome which included PPH, however, 99.9% of alerts were issued antepartum with many alerts soon after admission [37]. Our model uses parameterized methods for capturing information prior to delivery at a lower frequency. We chose the time of delivery as the first prediction for several reasons. An antepartum alert for postpartum hemorrhage may be difficult to interpret and act upon. Anchoring the first prediction to the time of delivery focuses the model on time varying predictors in the postpartum period, while still allowing incorporation of the patient's care up to the point of delivery into the model. This potentially increases the actionability of any alerts.

The frequency with which predictions are made is important in relation to the outcome event because PPH can occur rapidly. In a previous study, Klumpner et al. found the time to intervention for severe postpartum hemorrhage from delivery was less than an hour in 45.8% (55/120) of cases of severely morbid PPH [8]. Therefore, predictions must be issued at least more frequently than every hour. In order to capture the outcome prior to the lag, a prediction interval of less than half the median lead time is needed, or at least every 24 minutes. Escobar et al. included PPH in their composite outcome with a prediction interval every hour but only 0.1% of alerts were postpartum with no true positives eliminating any effectiveness in predicting postpartum hemorrhage after delivery [37]. Escobar et al. also discovered that the median time from admission to the first prediction was 1.6 hours (at threshold probability of 4.1% for logistic model) and median time from admission to delivery was 10.8 hours. Even if an intrapartum alert for PPH was a true positive, the clinical intervention of an alert that occurs 9 hours before an event as acute as PPH is not well-defined.

Given that PPH is commonly an acute event which occurs in a short amount of time with little warning, a more realistic lead time allows more reactionary measures to be

taken, like preparing an operating room or ordering blood products in preparation to treat hemorrhage. Our study found an average lead time of 60 minutes, a median lead time of 39 minutes, and an interquartile range of 32 minutes for patients who both met the outcome and exceeded the probability of 10% prior to that outcome. This lead time would hypothetically trigger an alert to the clinician for further evaluation within an actionable timeframe. However, further analysis must be conducted to understand how this would affect clinician workflow in terms of frequency of alerts as well as multiple alerts for each patient.

This study has several limitations. The most significant assumption we made is that QBL is calculated and entered in real-time. A lag in documentation could greatly affect the performance in a prospective study as relative time since delivery was deemed the most important variable. Our prevalence for PPH was very high (25%) in comparison with the Venkatesh model (4.7%) and the Escobar model (0.32%). Possible explanations for the difference in prevalence are that being a tertiary care center, our hospital is more likely to care for patients with a higher rate of comorbidity and difference in clinical practice patterns. Both studies we compared our model to define PPH based on EBL [36], [37], but our study cohort relies on QBL, which is more accurate [58]. The incidence of PPH when assessed using the same ≥1000 mL threshold of blood loss is higher when measured by QBL as compared to EBL [57], suggesting that EBL may only be identifying severe hemorrhage whereas QBL may also be capturing less severe bleeding. Unsurprisingly, severe hemorrhage may be easier for a model to predict and result in a higher AUROC.

Despite these limitations in our model, we are the first, to our knowledge, to introduce a model capable of solely predicting PPH with reasonable discrimination using a multiple prediction approach. We believe this is the first step to incorporating this into clinical practice. We anticipate future studies will focus on prospective validation of this model to assess its effectiveness in real-time. If prospective validation is successful, then a clinical trial would be the next step to evaluate its potential to make a clinical impact.

5.6 Supplemental Tables

Supplemental Table 5.3: Patient characteristics, stratified by development/validation sets (expanded)

Characteristic	Overall, N = 6,000	train, N = 3840 (64%)[1]	tune, N = 960 (16%)[1]	test, N = 1200 (20%)[1]	p-value[2]
age_at_admission	31.0 (27.0, 34.0)	31.0 (27.0, 34.0)	31.0 (27.0, 34.0)	31.0 (27.0, 34.0)	>0.9
anemia	1,674 (28%)	1,053 (27%)	281 (29%)	340 (28%)	0.5
antepartum_vaginal_bleeding	1,248 (21%)	830 (22%)	181 (19%)	237 (20%)	0.11
assisted_reproductive_technology	302 (5.0%)	198 (5.2%)	54 (5.6%)	50 (4.2%)	0.3
asthma_active_airway_disease	1,157 (19%)	742 (19%)	200 (21%)	215 (18%)	0.2
bmi	31 (27, 35)	31 (27, 35)	31 (27, 36)	31 (27, 35)	0.4
Missing	43	31	3	9	
breech_abnormal_lie	1,149 (19%)	755 (20%)	195 (20%)	199 (17%)	0.031
carboprost					0.4
0	5,986 (100%)	3,833 (100%)	956 (100%)	1,197 (100%)	
250	14 (0.2%)	7 (0.2%)	4 (0.4%)	3 (0.2%)	
cesarean_delivery	1,741 (29%)	1,072 (28%)	293 (31%)	376 (31%)	0.041
chorioamnionitis_on_admission	317 (5.3%)	206 (5.4%)	48 (5.0%)	63 (5.2%)	>0.9
chronic_hypertension	665 (11%)	403 (10%)	119 (12%)	143 (12%)	0.14
chronic_renal_disease	154 (2.6%)	110 (2.9%)	12 (1.2%)	32 (2.7%)	0.009
depression	1,351 (23%)	840 (22%)	227 (24%)	284 (24%)	0.3
eclampsia	15 (0.2%)	9 (0.2%)	4 (0.4%)	2 (0.2%)	0.5
epis_gravida_count	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	0.3
Missing	15	9	4	2	
epis_para_count	1.00 (0.00, 1.00)	1.00 (0.00, 1.00)	1.00 (0.00, 1.00)	1.00 (0.00, 2.00)	0.072
Missing	15	9	4	2	
fetal_demise	135 (2.2%)	88 (2.3%)	16 (1.7%)	31 (2.6%)	0.4
fetal_macrosomia	311 (5.2%)	201 (5.2%)	51 (5.3%)	59 (4.9%)	>0.9
gastrointestinal_disease	1,774 (30%)	1,124 (29%)	300 (31%)	350 (29%)	0.5
gbs_colonization	1,620 (27%)	1,069 (28%)	240 (25%)	311 (26%)	0.14
genital_tract_laceration	64 (1.1%)	35 (0.9%)	5 (0.5%)	24 (2.0%)	0.003

gestation_days_on_admit	275 (268, 283)	275 (268, 283)	275 (268, 283)	275 (267, 283)	0.7
Missing	27	15	4	8	
gestational_diabetes	600 (10%)	395 (10%)	92 (9.6%)	113 (9.4%)	0.6
gestational_hypertension	1,165 (19%)	731 (19%)	207 (22%)	227 (19%)	0.2
heart_disease	349 (5.8%)	222 (5.8%)	59 (6.1%)	68 (5.7%)	0.9
height_height	165 (160, 170)	163 (160, 169)	165 (160, 170)	165 (160, 168)	0.3
Missing	4,562	2,979	727	856	
illegal_drug_use	552 (9.2%)	396 (10%)	79 (8.2%)	77 (6.4%)	<0.001
initial_labor_status					0.005
Induction	2,081 (43%)	1,297 (42%)	377 (48%)	407 (43%)	
Labor	2,746 (57%)	1,811 (58%)	405 (52%)	530 (57%)	
Missing	1,173	732	178	263	
instrumental_vaginal_delivery	106 (1.8%)	66 (1.7%)	19 (2.0%)	21 (1.8%)	0.8
insurance					>0.9
Medicaid	1,656 (28%)	1,071 (28%)	255 (27%)	330 (28%)	
Medicare	25 (0.4%)	15 (0.4%)	4 (0.4%)	6 (0.5%)	
Other	13 (0.2%)	7 (0.2%)	2 (0.2%)	4 (0.3%)	
Other Governmental Insurance	43 (0.7%)	29 (0.8%)	8 (0.8%)	6 (0.5%)	
Private Insurance	4,185 (70%)	2,675 (70%)	680 (71%)	830 (70%)	
Workers Compensation	55 (0.9%)	36 (0.9%)	6 (0.6%)	13 (1.1%)	
Missing	23	7	5	11	
international_normalized_ratio	1.00 (0.90, 1.00)	1.00 (0.90, 1.00)	0.90 (0.90, 1.00)	1.00 (0.90, 1.00)	0.3
Missing	4,910	3,147	781	982	
intrauterine_growth_restriction	1,086 (18%)	645 (17%)	161 (17%)	280 (23%)	<0.001
labor_duration	15 (7, 24)	15 (8, 24)	16 (8, 27)	15 (7, 26)	0.027
Missing	1,173	732	178	263	
last_labor_status					0.001
Ante	85 (1.5%)	57 (1.6%)	9 (1.0%)	19 (1.7%)	
C/S	1,082 (19%)	666 (18%)	195 (21%)	221 (19%)	
Forcep	1 (<0.1%)	1 (<0.1%)	0 (0%)	0 (0%)	
Induction	1,850 (32%)	1,171 (32%)	319 (34%)	360 (32%)	

Labor	2,545 (44%)	1,674 (46%)	380 (41%)	491 (43%)	
Loss/Ind	33 (0.6%)	19 (0.5%)	1 (0.1%)	13 (1.1%)	
Loss/Labor	21 (0.4%)	11 (0.3%)	3 (0.3%)	7 (0.6%)	
Loss/PP	1 (<0.1%)	0 (0%)	0 (0%)	1 (<0.1%)	
NSVD	28 (0.5%)	11 (0.3%)	8 (0.9%)	9 (0.8%)	
Obs	11 (0.2%)	6 (0.2%)	2 (0.2%)	3 (0.3%)	
Tr-Fall/trauma	1 (<0.1%)	0 (0%)	0 (0%)	1 (<0.1%)	
Tr-r/o labor	26 (0.5%)	16 (0.4%)	4 (0.4%)	6 (0.5%)	
Tr-r/o ROM	5 (<0.1%)	2 (<0.1%)	3 (0.3%)	0 (0%)	
Triage	30 (0.5%)	26 (0.7%)	1 (0.1%)	3 (0.3%)	
Vbac	2 (<0.1%)	0 (0%)	1 (0.1%)	1 (<0.1%)	
Missing	279	180	34	65	
large_for_gestational_age	74 (1.2%)	53 (1.4%)	9 (0.9%)	12 (1.0%)	0.5
large_uterine_fibroids	214 (3.6%)	144 (3.8%)	31 (3.2%)	39 (3.2%)	0.7
magnesium_level	1.80 (1.60, 2.00)	1.80 (1.60, 2.00)	1.90 (1.60, 2.10)	1.80 (1.60, 2.00)	0.4
Missing	5,387	3,446	855	1,086	
magnesium_sulfate	0.00 (0.00, 2.00)	0.00 (0.00, 2.00)	0.00 (0.00, 2.00)	0.00 (0.00, 2.00)	0.3
Missing	5,479	3,513	874	1,092	
maternal_gbs_colonization	867 (14%)	570 (15%)	119 (12%)	178 (15%)	0.2
median_height_cm	164 (160, 168)	163 (160, 168)	164 (160, 168)	165 (160, 168)	0.3
Missing	43	31	3	9	
median_weight_kg	83 (73, 95)	82 (72, 95)	83 (73, 95)	83 (73, 95)	0.2
Missing	43	31	3	9	
multiple_gestation	224 (3.7%)	141 (3.7%)	42 (4.4%)	41 (3.4%)	0.5
nifedipine					0.01
0	50 (23%)	22 (16%)	14 (40%)	14 (32%)	
10	63 (29%)	47 (34%)	6 (17%)	10 (23%)	
20	23 (11%)	17 (12%)	2 (5.7%)	4 (9.1%)	
30	54 (25%)	33 (24%)	6 (17%)	15 (34%)	
60	23 (11%)	15 (11%)	7 (20%)	1 (2.3%)	
90	4 (1.8%)	4 (2.9%)	0 (0%)	0 (0%)	

Missing	5,783	3,702	925	1,156	
non_gestational_diabetes	169 (2.8%)	112 (2.9%)	29 (3.0%)	28 (2.3%)	0.5
number_of_fetuses					0.4
0	1 (<0.1%)	1 (<0.1%)	0 (0%)	0 (0%)	
1	5,731 (98%)	3,677 (97%)	899 (97%)	1,155 (98%)	
2	132 (2.2%)	91 (2.4%)	20 (2.2%)	21 (1.8%)	
3	11 (0.2%)	5 (0.1%)	4 (0.4%)	2 (0.2%)	
Missing	125	66	37	22	
other_puerperal_infection	4 (<0.1%)	4 (0.1%)	0 (0%)	0 (0%)	0.6
placenta_accreta_spectrum	41 (0.7%)	24 (0.6%)	7 (0.7%)	10 (0.8%)	0.6
placenta_previa	213 (3.5%)	137 (3.6%)	31 (3.2%)	45 (3.8%)	0.8
placental_abruption	143 (2.4%)	95 (2.5%)	18 (1.9%)	30 (2.5%)	0.6
polyhydramnios	242 (4.0%)	154 (4.0%)	36 (3.8%)	52 (4.3%)	0.8
pph	1,514 (25%)	952 (25%)	263 (27%)	299 (25%)	0.3
preeclampsia_with_severe_features	359 (6.0%)	214 (5.6%)	60 (6.2%)	85 (7.1%)	0.2
preeclampsia_without_severe_features	425 (7.1%)	270 (7.0%)	68 (7.1%)	87 (7.2%)	>0.9
premature_rupture_of_membranes	1,406 (23%)	912 (24%)	208 (22%)	286 (24%)	0.4
preterm_labor	197 (3.3%)	118 (3.1%)	36 (3.8%)	43 (3.6%)	0.4
prior_cesarean_delivery	1,072 (18%)	667 (17%)	161 (17%)	244 (20%)	0.043
prior_pph	1,852 (31%)	1,152 (30%)	288 (30%)	412 (34%)	0.014
r_bmi_bmi_calculated	32 (28, 38)	32 (28, 38)	32 (28, 37)	32 (28, 37)	>0.9
Missing	4,719	3,084	752	883	
respirations_resp	18.00 (16.00, 18.00)	18.00 (16.00, 18.00)	18.00 (16.00, 18.00)	18.00 (16.00, 18.00)	0.009
Missing	328	225	37	66	
seizure_disorder	74 (1.2%)	47 (1.2%)	12 (1.2%)	15 (1.2%)	>0.9
sp_o_2	99.00 (98.00, 100.00)	99.00 (98.00, 100.00)	99.00 (98.00, 100.00)	99.00 (98.00, 100.00)	0.026
Missing	402	257	57	88	
spontaneous_labor	312 (5.2%)	208 (5.4%)	47 (4.9%)	57 (4.8%)	0.6
superimposed_preeclampsia	99 (1.7%)	63 (1.6%)	14 (1.5%)	22 (1.8%)	0.8

temperature_temp	98.20 (97.90, 98.60)	98.20 (98.00, 98.60)	98.20 (97.90, 98.50)	98.20 (97.90, 98.50)	<0.001
Missing	327	227	39	61	
thyroid_disease	773 (13%)	486 (13%)	144 (15%)	143 (12%)	0.081
tobacco_use	311 (5.2%)	203 (5.3%)	45 (4.7%)	63 (5.2%)	0.8
trial_of_labor	459 (7.6%)	277 (7.2%)	81 (8.4%)	101 (8.4%)	0.2
um_ip_r_magnesium_sulfate_weight_magnesium_sulfate_dose_weight					0.2
0	5,922 (99%)	3,782 (98%)	950 (99%)	1,190 (99%)	
2000	78 (1.3%)	58 (1.5%)	10 (1.0%)	10 (0.8%)	
um_ip_r_pulse_pressure_pulse_pressure_systemic	49 (41, 57)	49 (42, 57)	49 (41, 57)	49 (41, 55)	0.088
Missing	241	170	22	49	
um_ip_r_urine_output_output_m_l	250 (125, 450)	250 (125, 450)	250 (138, 425)	200 (120, 400)	0.017
Missing	3,306	2,114	524	668	
um_r_oxygen_flow_rate_o_2_flow_rate_l_min	10.0 (10.0, 10.0)	10.0 (10.0, 10.0)	10.0 (10.0, 10.0)	10.0 (1.0, 10.0)	0.02
Missing	5,792	3,679	938	1,175	
um_r_oxytocin_volume_volume_m_l	18 (6, 43)	18 (6, 42)	19 (6, 46)	18 (7, 42)	>0.9
Missing	4,553	2,952	723	878	
um_r_pain_scale_numeric_1_pain_rating	5.0 (2.0, 7.0)	5.0 (2.0, 7.0)	6.0 (2.0, 8.0)	5.0 (2.0, 7.0)	0.003
Missing	3,497	2,220	551	726	
um_stk_r_qbl_calculated_qbl_calculated_qbl_m_l	90 (30, 201)	88 (30, 168)	112 (71, 133)	108 (46, 300)	0.4
Missing	5,887	3,765	950	1,172	
um_stk_r_qbl_running_total_qbl_running_total	128 (66, 292)	124 (58, 286)	118 (71, 243)	197 (74, 355)	0.7
Missing	5,885	3,763	950	1,172	
urine_output_urine	225 (150, 400)	238 (150, 400)	250 (150, 400)	200 (150, 394)	0.084
Missing	4,125	2,810	673	642	
urine_protein_creatinine_ratio	0.13 (0.09, 0.23)	0.13 (0.09, 0.22)	0.14 (0.09, 0.24)	0.13 (0.09, 0.24)	0.7
Missing	3,968	2,569	626	773	
weight_scale_weight	84 (74, 98)	84 (74, 98)	84 (75, 98)	84 (75, 97)	0.6
Missing	2,921	1,876	451	594	

white_blood_cell_count	5 (2, 14)	5 (2, 15)	5 (1, 13)	5 (2, 13)	0.4
Missing	4,557	2,942	700	915	

[1]Statistics presented: median (IQR); n (%)

[2]Statistical tests performed: Kruskal-Wallis test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

Supplemental Table 5.4: Patient characteristics, stratified by the outcome (expanded)

Characteristic	**Overall, N = 6,000**	**No PPH, N = 4486 (75%)[1]**	**PPH, N = 1514 (25%)[1]**	**p-value[2]**
age_at_admission	31.0 (27.0, 34.0)	31.0 (27.0, 34.0)	31.0 (28.0, 35.0)	<0.001
amb_r_bp_diastolic_diastolic	73 (65, 81)	72 (65, 80)	74 (66, 82)	<0.001
Missing	241	199	42	
amb_r_bp_systolic_systolic	122 (112, 131)	121 (111, 131)	124 (114, 133)	<0.001
Missing	241	199	42	
anemia	1,674 (28%)	1,206 (27%)	468 (31%)	0.003
antepartum_vaginal_bleeding	1,248 (21%)	910 (20%)	338 (22%)	0.092
assisted_reproductive_technology	302 (5.0%)	162 (3.6%)	140 (9.2%)	<0.001
bmi	31 (27, 35)	30 (27, 35)	32 (28, 37)	<0.001
Missing	43	35	8	
breech_abnormal_lie	1,149 (19%)	774 (17%)	375 (25%)	<0.001
cesarean_delivery	1,741 (29%)	943 (21%)	798 (53%)	<0.001
chorioamnionitis_on_admission	317 (5.3%)	197 (4.4%)	120 (7.9%)	<0.001
chronic_hypertension	665 (11%)	455 (10%)	210 (14%)	<0.001
delivery_type				<0.001
C-SECTION LOW VERT	13 (0.2%)	6 (0.1%)	7 (0.5%)	
C-SECTION, CLASS	60 (1.0%)	31 (0.7%)	29 (1.9%)	
C-SECTION, LOW TRANS	1,685 (28%)	889 (20%)	796 (53%)	
MEDICAL TERMINATION	4 (<0.1%)	3 (<0.1%)	1 (<0.1%)	
SPONTANEOUS LOSS	9 (0.2%)	9 (0.2%)	0 (0%)	
VAGINAL, FORCEPS	38 (0.6%)	23 (0.5%)	15 (1.0%)	
VAGINAL, SPONTANEOUS	3,857 (65%)	3,250 (73%)	607 (40%)	
VAGINAL, VACUUM (EXTRACTOR)	61 (1.0%)	53 (1.2%)	8 (0.5%)	

VBAC, FRCPS	1 (<0.1%)	1 (<0.1%)	0 (0%)	
VBAC, SPONTANEOUS	239 (4.0%)	191 (4.3%)	48 (3.2%)	
VBAC, VACUUM	6 (0.1%)	5 (0.1%)	1 (<0.1%)	
Missing	27	25	2	
eclampsia	15 (0.2%)	8 (0.2%)	7 (0.5%)	0.072
epis_gravida_count	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	<0.001
Missing	15	14	1	
epis_para_count	1.00 (0.00, 1.00)	1.00 (0.00, 2.00)	1.00 (0.00, 1.00)	<0.001
Missing	15	14	1	
fetal_demise	135 (2.2%)	104 (2.3%)	31 (2.0%)	0.6
fetal_macrosomia	311 (5.2%)	203 (4.5%)	108 (7.1%)	<0.001
gestation_days_on_admit	275 (268, 283)	275 (268, 283)	274 (266, 283)	0.012
Missing	27	27	0	
gestational_diabetes	600 (10%)	410 (9.1%)	190 (13%)	<0.001
gestational_hypertension	1,165 (19%)	831 (19%)	334 (22%)	0.003
hemoglobin	12.00 (11.10, 12.80)	12.00 (11.10, 12.80)	11.90 (11.00, 12.80)	0.074
Missing	17	14	3	
initial_labor_status				<0.001
Induction	2,081 (43%)	1,535 (41%)	546 (50%)	
Labor	2,746 (57%)	2,205 (59%)	541 (50%)	
Missing	1,173	746	427	
insurance				0.002
Medicaid	1,656 (28%)	1,292 (29%)	364 (24%)	
Medicare	25 (0.4%)	14 (0.3%)	11 (0.7%)	
Other	13 (0.2%)	7 (0.2%)	6 (0.4%)	
Other Governmental Insurance	43 (0.7%)	31 (0.7%)	12 (0.8%)	
Private Insurance	4,185 (70%)	3,080 (69%)	1,105 (73%)	
Workers Compensation	55 (0.9%)	41 (0.9%)	14 (0.9%)	
Missing	23	21	2	
international_normalized_ratio	1.00 (0.90, 1.00)	1.00 (0.90, 1.00)	1.00 (0.90, 1.00)	0.8

Missing	4,910	3,739	1,171	
labor_duration	15 (7, 24)	14 (7, 22)	21 (11, 32)	<0.001
Missing	1,173	746	427	
large_uterine_fibroids	214 (3.6%)	137 (3.1%)	77 (5.1%)	<0.001
median_weight_kg	83 (73, 95)	82 (72, 94)	86 (75, 100)	<0.001
Missing	43	35	8	
nibp_systolic	122 (111, 133)	122 (111, 132)	124 (112, 135)	<0.001
Missing	156	136	20	
non_gestational_diabetes	169 (2.8%)	110 (2.5%)	59 (3.9%)	0.005
number_of_fetuses				<0.001
0	1 (<0.1%)	0 (0%)	1 (<0.1%)	
1	5,731 (98%)	4,316 (99%)	1,415 (95%)	
2	132 (2.2%)	58 (1.3%)	74 (4.9%)	
3	11 (0.2%)	5 (0.1%)	6 (0.4%)	
Missing	125	107	18	
oxytocin	4 (0, 10)	6 (0, 10)	0 (0, 10)	<0.001
Missing	1,737	1,317	420	
placenta_accreta_spectrum	41 (0.7%)	22 (0.5%)	19 (1.3%)	0.003
placenta_previa	213 (3.5%)	149 (3.3%)	64 (4.2%)	0.11
placental_abruption	143 (2.4%)	91 (2.0%)	52 (3.4%)	0.003
platelet_count	220 (184, 261)	221 (186, 261)	217 (181, 260)	0.01
Missing	22	17	5	
polyhydramnios	242 (4.0%)	158 (3.5%)	84 (5.5%)	<0.001
preeclampsia_with_severe_features	359 (6.0%)	211 (4.7%)	148 (9.8%)	<0.001
preeclampsia_without_severe_features	425 (7.1%)	264 (5.9%)	161 (11%)	<0.001
premature_rupture_of_membranes	1,406 (23%)	1,103 (25%)	303 (20%)	<0.001
prior_cesarean_delivery	1,072 (18%)	692 (15%)	380 (25%)	<0.001
prior_pph	1,852 (31%)	1,373 (31%)	479 (32%)	0.5
pulse_pulse	89 (79, 101)	89 (78, 100)	90 (80, 102)	<0.001
Missing	220	186	34	
r_bmi_bmi_calculated	32 (28, 38)	32 (28, 37)	33 (29, 38)	0.003

Missing	4,719	3,593	1,126	
superimposed_preeclampsia	99 (1.7%)	65 (1.4%)	34 (2.2%)	0.046
temperature_temp	98.20 (97.90, 98.60)	98.20 (97.90, 98.60)	98.20 (98.00, 98.60)	<0.001
Missing	327	268	59	
terbutaline				0.005
0	1,177 (90%)	664 (88%)	513 (93%)	
0.25	122 (9.4%)	85 (11%)	37 (6.7%)	
1	3 (0.2%)	3 (0.4%)	0 (0%)	
250	2 (0.2%)	1 (0.1%)	1 (0.2%)	
Missing	4,696	3,733	963	
trial_of_labor	459 (7.6%)	319 (7.1%)	140 (9.2%)	0.009
um_ip_r_pulse_pressure_pulse_pressure_systemic	49 (41, 57)	49 (41, 56)	50 (42, 58)	<0.001
Missing	241	199	42	
um_r_oxytocin_volume_volume_m_l	18 (6, 43)	16 (6, 37)	24 (8, 55)	<0.001
Missing	4,553	3,487	1,066	
um_stk_r_qbl_calculated_qbl_calculated_qbl_m_l	90 (30, 201)	70 (18, 129)	130 (60, 266)	0.001
Missing	5,887	4,420	1,467	
um_stk_r_qbl_running_total_qbl_running_total	128 (66, 292)	100 (31, 251)	197 (100, 414)	0.004
Missing	5,885	4,419	1,466	
urine_protein_creatinine_ratio	0.13 (0.09, 0.23)	0.13 (0.09, 0.20)	0.15 (0.09, 0.29)	<0.001
Missing	3,968	3,071	897	
weight_scale_weight	84 (74, 98)	83 (73, 96)	88 (76, 102)	<0.001
Missing	2,921	2,274	647	

[1]Statistics presented: median (IQR); n (%)

[2]Statistical tests performed: Wilcoxon rank-sum test; Fisher's Exact Test for Count Data; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

Supplemental Table 5.5: Variable temporal definitions

Category	**Variable**	**Timespan**	**Type**	**Lookback**	**Window**	**Stats**
diagnoses	anemia	baseline	dichotomous	days(90)	days(30)	length

diagnoses	antepartum_vaginal_bleeding	baseline	dichotomous	days(90)	days(30)	length
diagnoses	assisted_reproductive_technology	baseline	dichotomous	days(90)	days(30)	length
diagnoses	asthma_active_airway_disease	baseline	dichotomous	days(90)	days(30)	length
diagnoses	breech_abnormal_lie	baseline	dichotomous	days(90)	days(30)	length
diagnoses	chorioamnionitis_on_admission	baseline	dichotomous	days(90)	days(30)	length
diagnoses	chronic_hypertension	baseline	dichotomous	days(90)	days(30)	length
diagnoses	chronic_renal_disease	baseline	dichotomous	days(90)	days(30)	length
diagnoses	depression	baseline	dichotomous	days(90)	days(30)	length
diagnoses	eclampsia	baseline	dichotomous	days(90)	days(30)	length
diagnoses	fetal_demise	baseline	dichotomous	days(90)	days(30)	length
diagnoses	fetal_macrosomia	baseline	dichotomous	days(90)	days(30)	length
diagnoses	gastrointestinal_disease	baseline	dichotomous	days(90)	days(30)	length
diagnoses	genital_tract_laceration	baseline	dichotomous	days(90)	days(30)	length
diagnoses	gestational_diabetes	baseline	dichotomous	days(90)	days(30)	length
diagnoses	gestational_hypertension	baseline	dichotomous	days(90)	days(30)	length
diagnoses	heart_disease	baseline	dichotomous	days(90)	days(30)	length
diagnoses	intrauterine_growth_restriction	baseline	dichotomous	days(90)	days(30)	length
diagnoses	large_for_gestational_age	baseline	dichotomous	days(90)	days(30)	length
diagnoses	large_uterine_fibroids	baseline	dichotomous	days(90)	days(30)	length
diagnoses	maternal_gbs_colonization	baseline	dichotomous	days(90)	days(30)	length
diagnoses	multiple_gestation	baseline	dichotomous	days(90)	days(30)	length
diagnoses	non_gestational_diabetes	baseline	dichotomous	days(90)	days(30)	length
diagnoses	other_puerperal_infection	baseline	dichotomous	days(90)	days(30)	length
diagnoses	placenta_accreta_spectrum	baseline	dichotomous	days(90)	days(30)	length
diagnoses	placenta_previa	baseline	dichotomous	days(90)	days(30)	length
diagnoses	placental_abruption	baseline	dichotomous	days(90)	days(30)	length
diagnoses	polyhydramnios	baseline	dichotomous	days(90)	days(30)	length
diagnoses	preeclampsia_with_severe_features	baseline	dichotomous	days(90)	days(30)	length
diagnoses	preeclampsia_without_severe_features	baseline	dichotomous	days(90)	days(30)	length
diagnoses	premature_rupture_of_membranes	baseline	dichotomous	days(90)	days(30)	length

diagnoses	preterm_labor	baseline	dichotomous	days(90)	days(30)	length
diagnoses	prior_cesarean_delivery	baseline	dichotomous	days(90)	days(30)	length
diagnoses	prior_pph	baseline	dichotomous	days(90)	days(30)	length
diagnoses	seizure_disorder	baseline	dichotomous	days(90)	days(30)	length
diagnoses	spontaneous_labor	baseline	dichotomous	days(90)	days(30)	length
diagnoses	superimposed_preeclampsia	baseline	dichotomous	days(90)	days(30)	length
diagnoses	thyroid_disease	baseline	dichotomous	days(90)	days(30)	length
diagnoses	trial_of_labor	baseline	dichotomous	days(90)	days(30)	length
flowsheets	amb_r_bp_diastolic_diastolic	baseline	continuous			min, max, median, first, last
flowsheets	amb_r_bp_systolic_systolic	baseline	continuous			min, max, median, first, last
flowsheets	height_height	baseline	continuous			min, max, last
flowsheets	pulse_oximetry_sp_o_2	baseline	continuous			min, max, median, first, last
flowsheets	pulse_pulse	baseline	continuous			min, max, median, first, last
flowsheets	r_bmi_bmi_calculated	baseline	continuous			min, max, last
flowsheets	r_map_map_mm_hg	baseline	continuous			min, max, median, first, last
flowsheets	respirations_resp	baseline	continuous			min, max, median, first, last
flowsheets	temperature_temp	baseline	continuous			min, max, median, first, last
flowsheets	um_ip_r_magnesium_sulfate_weight_magnesium_sulfate_dose_weight	baseline	dichotomous			sum
flowsheets	um_ip_r_pulse_pressure_pulse_pressure_systemic	baseline	continuous			min, max, median, first, last
flowsheets	um_ip_r_urine_output_output_m_l	baseline	continuous			sum
flowsheets	um_r_fio_2_fi_o_2	baseline	continuous			min, max, last
flowsheets	um_r_oxygen_flow_rate_o_2_flow_rate_l_min	baseline	continuous			min, max, last
flowsheets	um_r_oxytocin_volume_volume_m_l	baseline	continuous			sum
flowsheets	um_stk_r_qbl_calculated_qbl_calculated_qbl_m_l	baseline	continuous			sum
flowsheets	um_stk_r_qbl_running_total_qbl_running_total	baseline	continuous			sum
flowsheets	urine_output_urine	baseline	continuous			min, max, median, first, last
flowsheets	weight_scale_weight	baseline	continuous			min, max, last

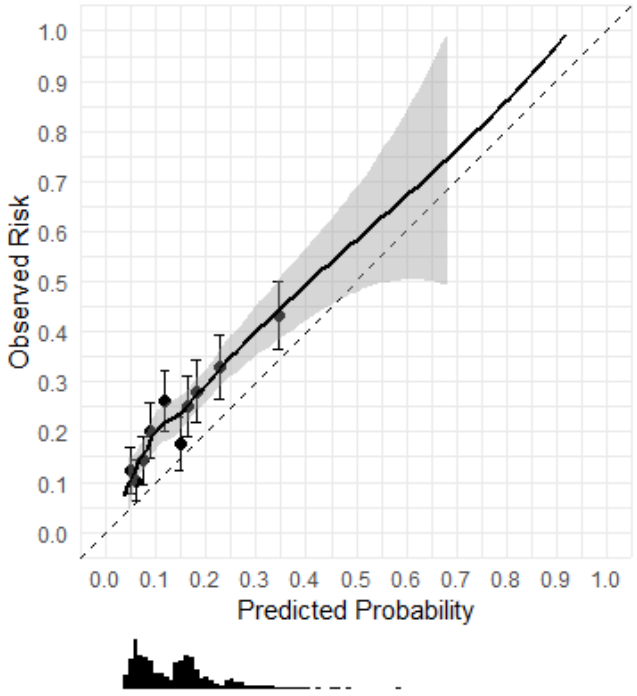
flowsheets	amb_r_bp_diastolic_diastolic	growing	continuous			min, max, median, first, last
flowsheets	amb_r_bp_systolic_systolic	growing	continuous			min, max, median, first, last
flowsheets	height_height	growing	continuous			last
flowsheets	pulse_oximetry_sp_o_2	growing	continuous			min, max, median, first, last
flowsheets	pulse_pulse	growing	continuous			min, max, median, first, last
flowsheets	r_bmi_bmi_calculated	growing	continuous			last
flowsheets	r_map_map_mm_hg	growing	continuous			min, max, median, first, last
flowsheets	respirations_resp	growing	continuous			min, max, median, first, last
flowsheets	temperature_temp	growing	continuous			min, max, median, first, last
flowsheets	um_ip_r_magnesium_sulfate_weight_magnesium_sulfate_dose_weight	growing	dichotomous			sum
flowsheets	um_ip_r_pulse_pressure_pulse_pressure_systemic	growing	continuous			min, max, median, first, last
flowsheets	um_ip_r_respiratory_effort_depth_resp_effort_depth	growing	categorical			last
flowsheets	um_ip_r_urine_output_output_m_l	growing	continuous			sum
flowsheets	um_r_fio_2_fi_o_2	growing	continuous			last
flowsheets	um_r_oxygen_flow_rate_o_2_flow_rate_l_min	growing	continuous			last
flowsheets	um_r_oxytocin_volume_volume_m_l	growing	continuous			sum
flowsheets	um_stk_r_qbl_calculated_qbl_calculated_qbl_m_l	growing	continuous			sum
flowsheets	um_stk_r_qbl_running_total_qbl_running_total	growing	continuous			sum
flowsheets	urine_output_urine	growing	continuous			min, max, median, first, last
flowsheets	weight_scale_weight	growing	continuous			last
ht_wt_bmi	bmi	baseline	continuous			min, max, last
ht_wt_bmi	median_height_cm	baseline	continuous			min, max, last
ht_wt_bmi	median_weight_kg	baseline	continuous			min, max, last
labor_categorical	delivery_type	baseline	categorical			length
labor_categorical	initial_labor_status	baseline	categorical			length
labor_categorical	labor_status	baseline	categorical			length

labor_duration	labor_duration	baseline	continuous			max
labs	creatinine	baseline	continuous			min, max, median, last
labs	fibrinogen	baseline	continuous			min, max, median, last
labs	gbs_colonization	baseline	dichotomous	days(30)		length
labs	hemoglobin	baseline	continuous			min, max, median, last
labs	international_normalized_ratio	baseline	continuous			min, max, median, last
labs	magnesium_level	baseline	continuous			min, max, median, last
labs	neutrophil_number	baseline	continuous			min, max, median, last
labs	partial_thromboplastin_time	baseline	continuous			min, max, median, last
labs	platelet_count	baseline	continuous			min, max, median, last
labs	urine_protein_creatinine_ratio	baseline	continuous			min, max, median, last
labs	white_blood_cell_count	baseline	continuous			min, max, median, last
labs	creatinine	growing	continuous			last
labs	fibrinogen	growing	continuous			last
labs	hemoglobin	growing	continuous			last
labs	international_normalized_ratio	growing	continuous			last
labs	magnesium_level	growing	continuous			last
labs	neutrophil_number	growing	continuous			last
labs	partial_thromboplastin_time	growing	continuous			last
labs	platelet_count	growing	continuous			last
labs	urine_protein_creatinine_ratio	growing	continuous			last
labs	white_blood_cell_count	growing	continuous			last
labs	creatinine	rolling	continuous	hours(12)		last
labs	fibrinogen	rolling	continuous	hours(12)		last
labs	hemoglobin	rolling	continuous	hours(12)		last
labs	international_normalized_ratio	rolling	continuous	hours(12)		last
labs	magnesium_level	rolling	continuous	hours(12)		last
labs	neutrophil_number	rolling	continuous	hours(12)		last
labs	partial_thromboplastin_time	rolling	continuous	hours(12)		last

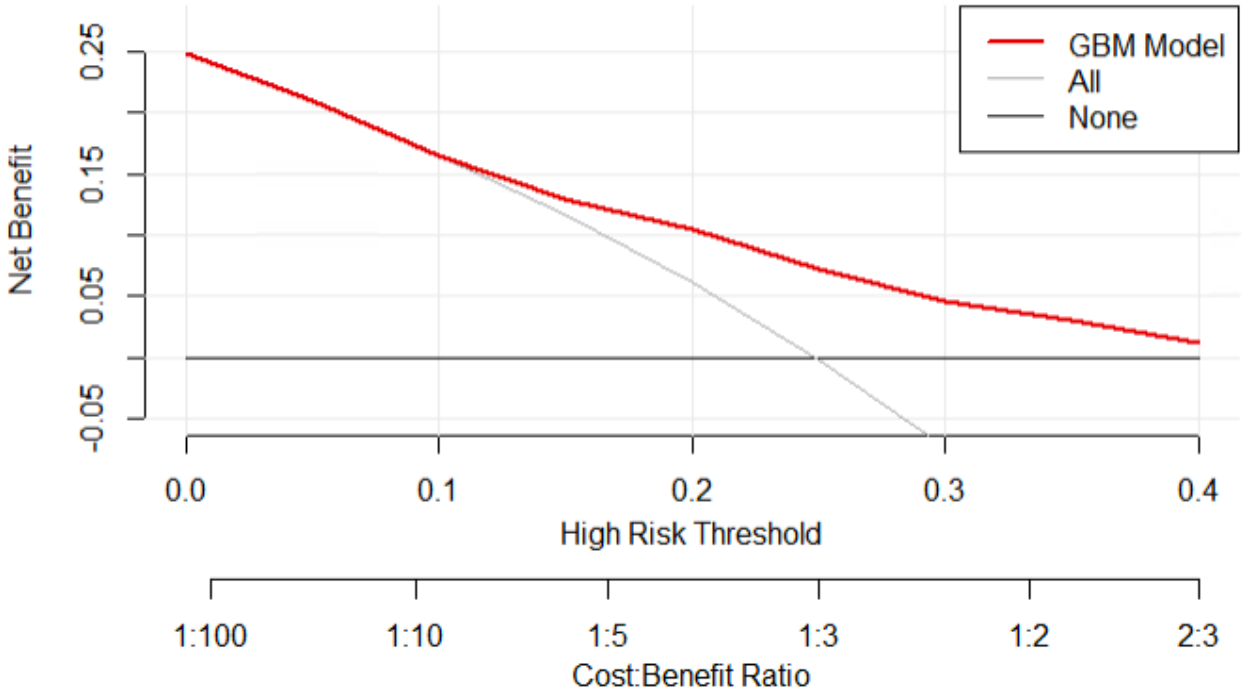
labs	platelet_count	rolling	continuous	hours(12)		last
labs	urine_protein_creatinine_ratio	rolling	continuous	hours(12)		last
labs	white_blood_cell_count	rolling	continuous	hours(12)		last
meds	betamethasone	baseline	continuous	hours(48)		sum
meds	carboprost	baseline	dichotomous	hours(48)		sum
meds	dexamethasone	baseline	continuous	hours(48)		sum
meds	indomethacin	baseline	continuous	hours(48)		sum
meds	magnesium_sulfate	baseline	continuous	hours(48)		sum
meds	methylergonovine	baseline	dichotomous	hours(48)		sum
meds	misoprostol	baseline	continuous	hours(48)		sum
meds	nifedipine	baseline	continuous	hours(48)		sum
meds	oxytocin	baseline	continuous	hours(48)		sum
meds	terbutaline	baseline	continuous	hours(48)		sum
meds	betamethasone	growing	dichotomous			sum
meds	carboprost	growing	dichotomous			sum
meds	dexamethasone	growing	continuous			sum
meds	indomethacin	growing	dichotomous			sum
meds	magnesium_sulfate	growing	continuous			sum
meds	methylergonovine	growing	dichotomous			sum
meds	misoprostol	growing	continuous			sum
meds	nifedipine	growing	continuous			sum
meds	oxytocin	growing	continuous			sum
meds	terbutaline	growing	dichotomous			sum
physiobank	heart_rate	baseline	dichotomous			min, max, median, first, last
physiobank	nibp_dias	baseline	continuous			min, max, median, first, last
physiobank	nibp_diastolic	baseline	dichotomous			min, max, median, first, last
physiobank	nibp_mean	baseline	dichotomous			min, max, median, first, last
physiobank	nibp_pulse_ox	baseline	continuous			min, max, median, first, last
physiobank	nibp_pulse_rate	baseline	continuous			min, max, median, first, last
physiobank	nibp_systolic	baseline	continuous			min, max, median, first, last

physiobank	shock_index	baseline	dichotomous			min, max, median, first, last
physiobank	sp_o_2	baseline	continuous			min, max, median, first, last
physiobank	sp_o_2_quality_index	baseline	dichotomous			min, max, median, first, last
physiobank	heart_rate	growing	continuous			min, max, median, first, last
physiobank	nibp_dias	growing	continuous			min, max, median, first, last
physiobank	nibp_diastolic	growing	continuous			min, max, median, first, last
physiobank	nibp_mean	growing	continuous			min, max, median, first, last
physiobank	nibp_pulse_ox	growing	continuous			min, max, median, first, last
physiobank	nibp_pulse_rate	growing	continuous			min, max, median, first, last
physiobank	nibp_systolic	growing	continuous			min, max, median, first, last
physiobank	shock_index	growing	continuous			min, max, median, first, last
physiobank	sp_o_2	growing	continuous			min, max, median, first, last
physiobank	sp_o_2_quality_index	growing	continuous			min, max, median, first, last
social_hx	illegal_drug_use	baseline	dichotomous			last
social_hx	tobacco_use	baseline	dichotomous			last
stork	cesarean_delivery	baseline	dichotomous			length
stork	instrumental_vaginal_delivery	baseline	dichotomous			last
stork	cesarean_delivery	growing	dichotomous			last

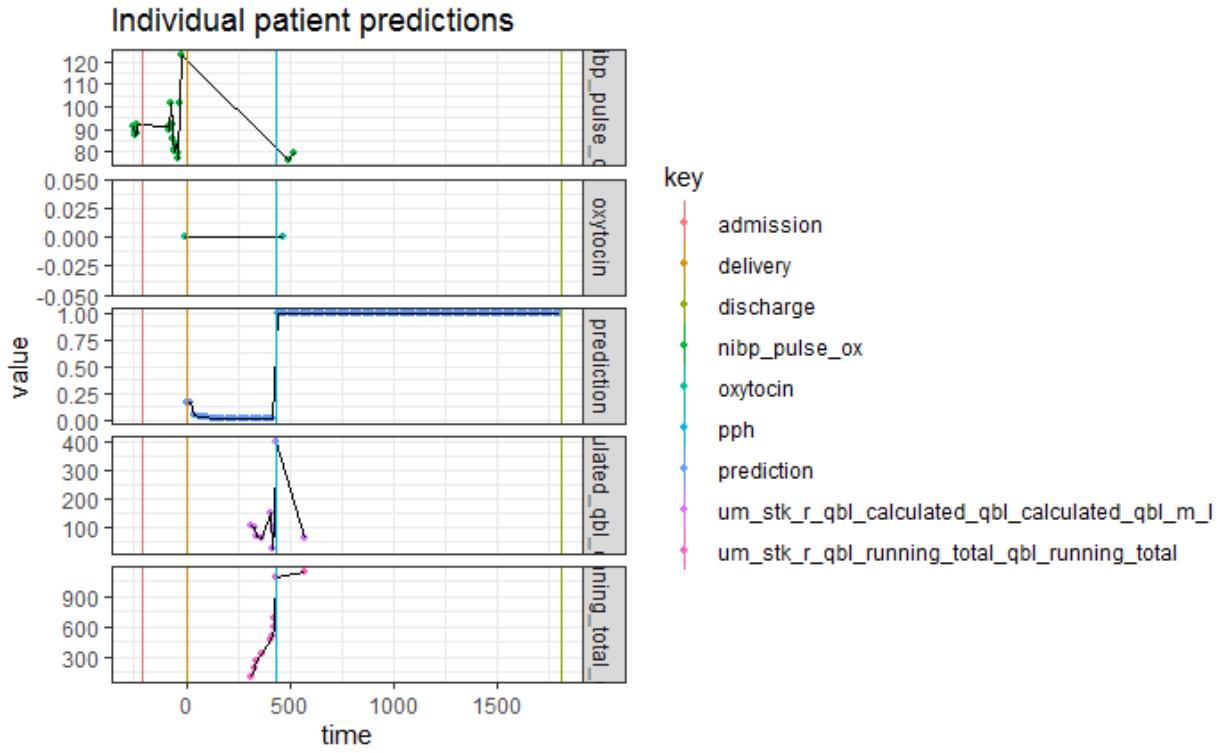
5.7 Supplemental Figures



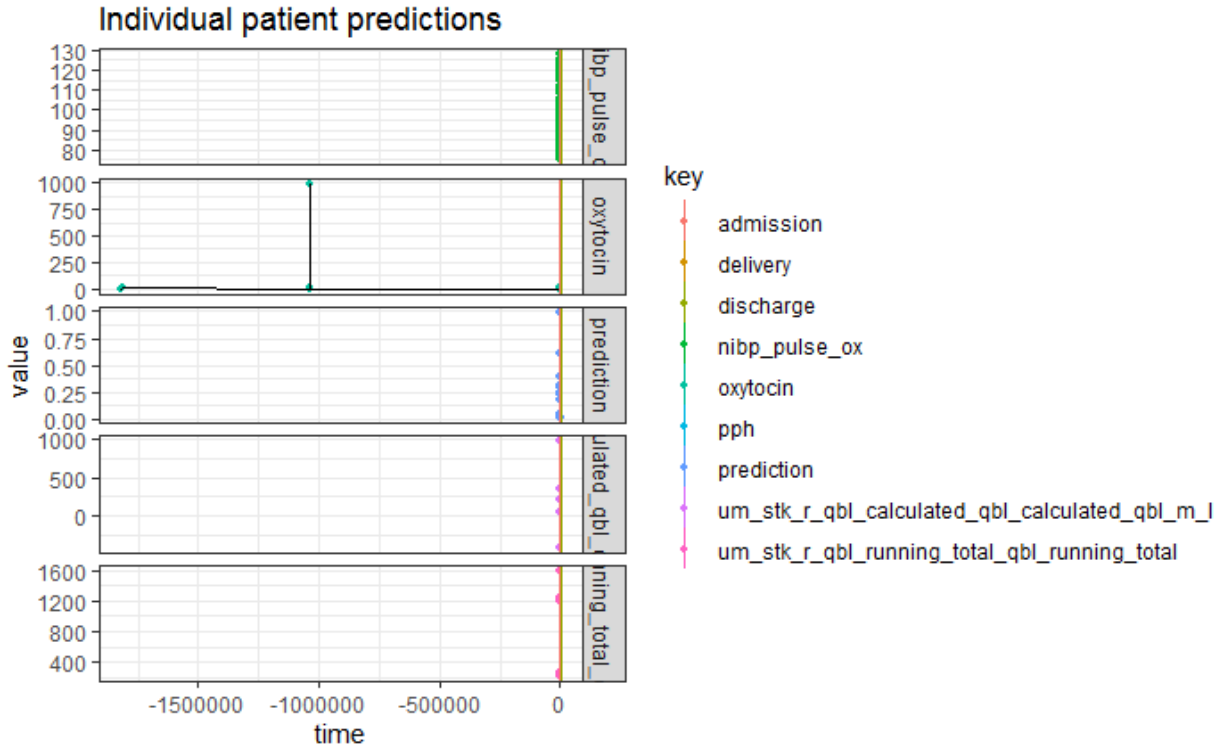
Supplemental Figure 5.6: Calibration of maximum probability per hospitalization



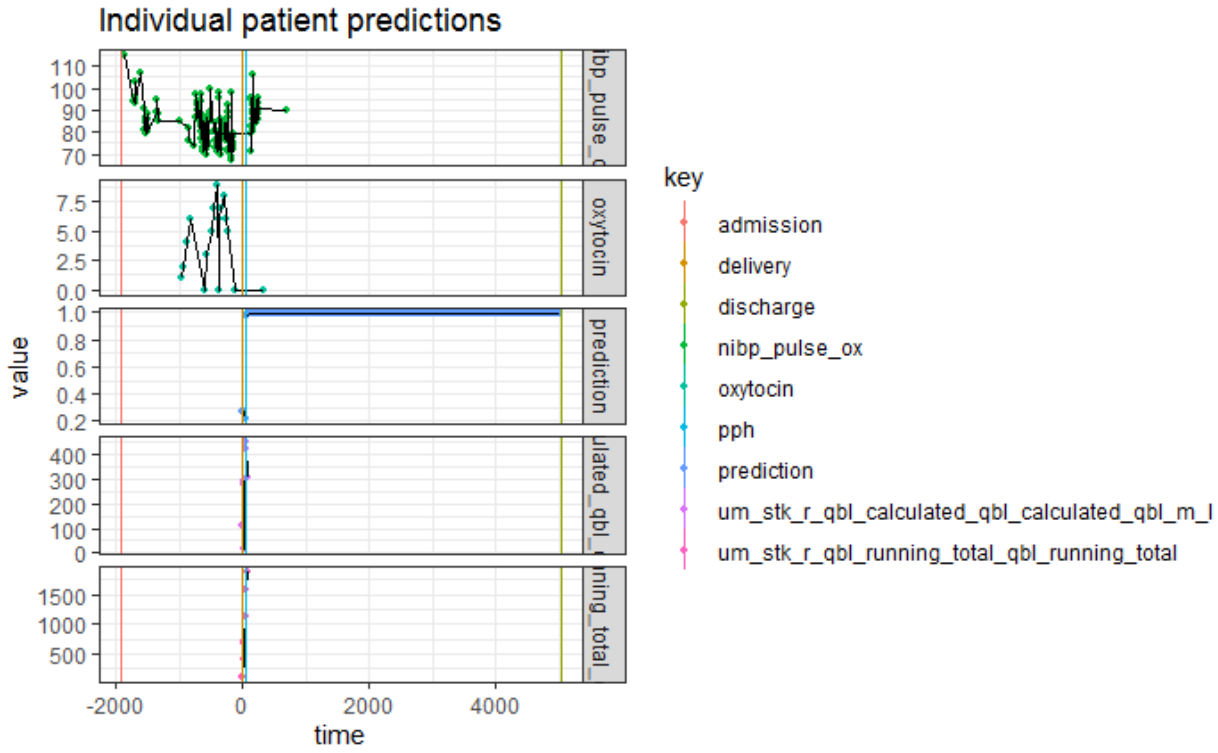
Supplemental Figure 5.7: Decision curve analysis
Decision curve analysis refers to the net benefit assessment for predictive models. More appropriate for a clinical setting where there are costs and benefits associated with treatment, net benefit assesses this balance which justifies treatment in the face of uncertainty [106], [107].



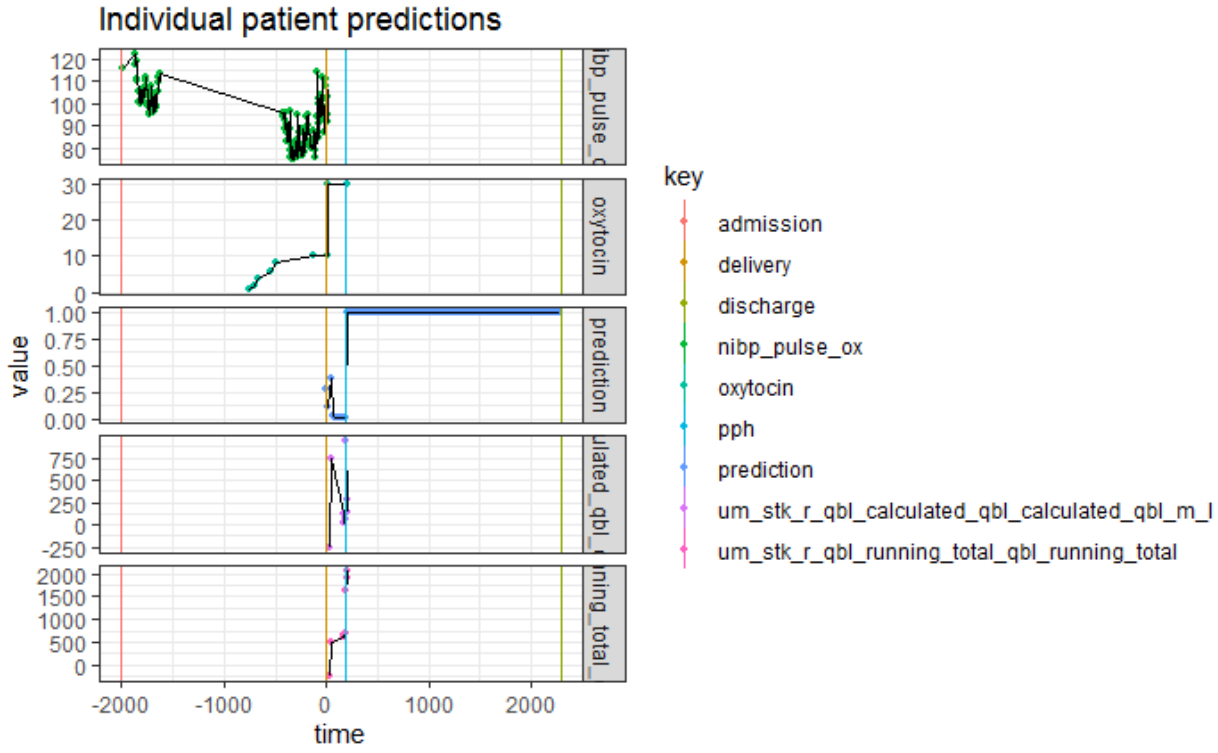
Supplemental Figure 5.8: Patient with highest overall risk who experienced outcome



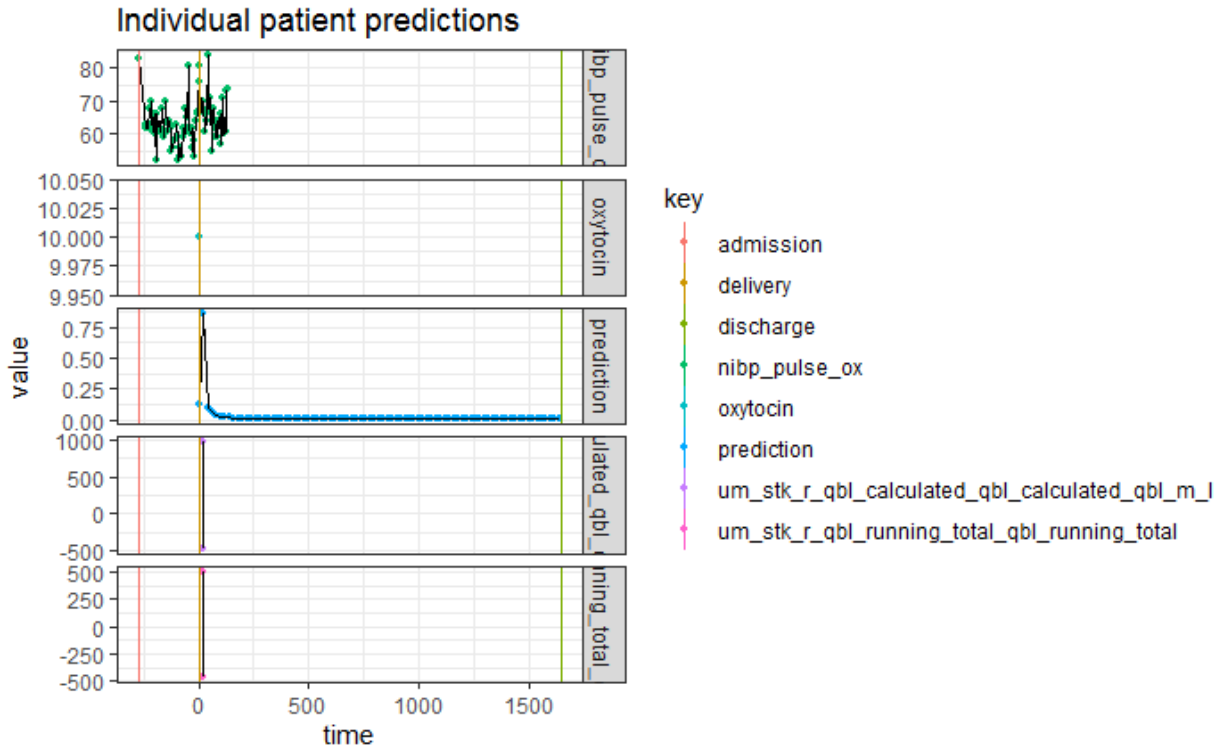
Supplemental Figure 5.9: Patient with highest overall risk who did not experience outcome



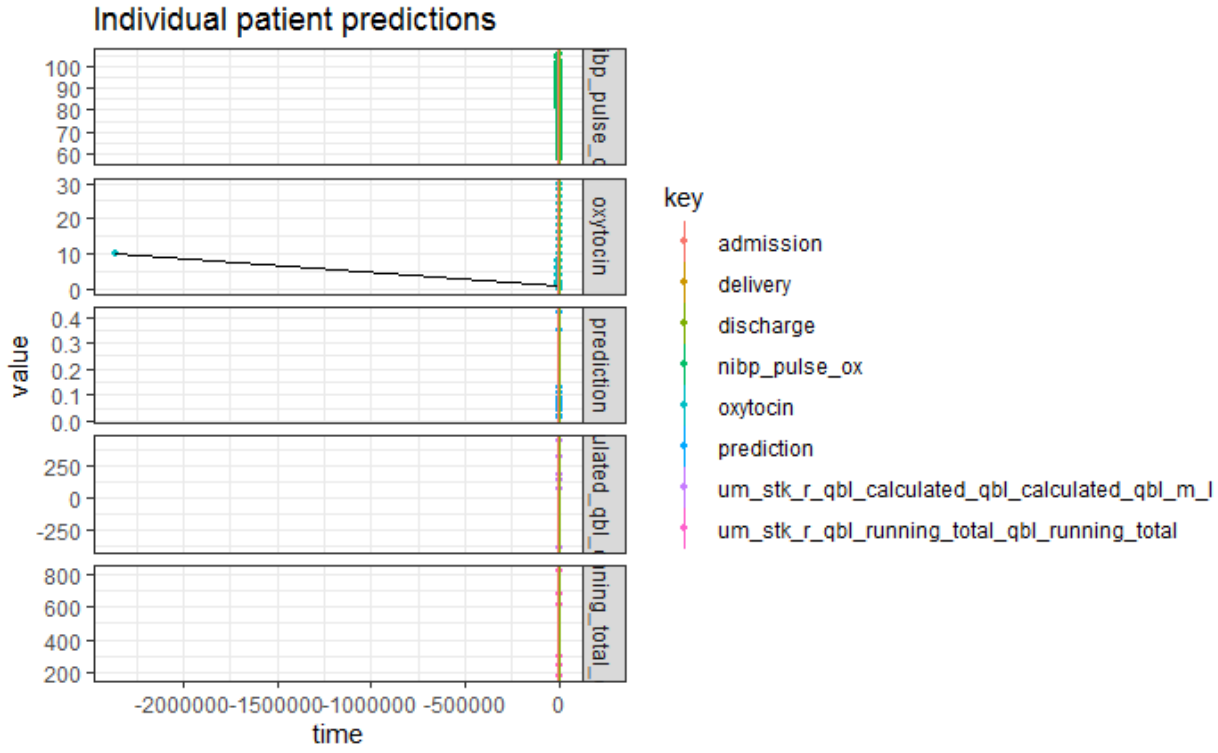
Supplemental Figure 5.10: Random patient who experienced outcome (1)



Supplemental Figure 5.11: Random patient who experienced outcome (2)



Supplemental Figure 5.12: Random patient who did not experience outcome (1)



Supplemental Figure 5.13: Random patient who did not experience outcome (2)

Chapter 6
Discussion

Chapters 3-5 identify the need for a time-series approach, introduce a novel methodological framework for structuring clinical time-series data, and apply this framework on preparing data for prediction of postpartum hemorrhage with a real-world dataset. This chapter summarizes the findings of these chapters, identifies implications and limitations of this work, and provides recommendations for future research.

6.1 Summary of Findings

In this work we set out to build a machine learning model to predict postpartum hemorrhage. Along the way we realized that a time-series approach may improve model performance but was more difficult to accomplish than we had anticipated. We discovered that data transformation was the most significant inhibitor to progress and this became the focus of the work. We found that contemporary frameworks used to pre-process clinical data are not ideal to build predictors with expert knowledge. I recognized this as the problem over the progress of this body of work and this became the design problem we attempted to solve.

Chapter 3 evaluated existing state-of-the-art models to predict postpartum hemorrhage at a single point, admission, in order to make two determinations. We wanted to know (1) if models developed elsewhere could be validated at our institution and (2) if a time-series approach was necessary. We found that these models did not perform well with data at our institution when compared to refit models, which also performed relatively poorly. Since these models did not generalize to our institution, we explored a time-series approach that would incorporate new information collected from the EHR and update risk predictions accordingly.

We discovered that, in addition to the analytical challenges inherent in team-based predictive modeling, a communication gap exists when incorporating domain expertise into data preparation techniques. In order to solve this problem, we developed an interface layer to communicate the important concepts of a time-series approach to clinical prediction modeling taking into account the common language used in team-based analytical studies. We sought to develop a framework which could design data preparation steps required for predictive modeling.

Our general approach was to informally study data preparation methods of recent predictive models intended for use in early warning systems. In a review of literature, recent inpatient prediction models were trained with methods which went unpublished or were difficult to easily generalize. Methods in transformation of time-series data remains an understudied area of research. We explored the possibility of condensing this knowledge into an easily readable format. We explored the inherent complexity in multivariate time-series and discovered a common language used in time-series data analysis which could be used as descriptors to build predictors for modeling.

Based on our findings, we designed and developed a framework focused on the communication aspect of time-series analysis, which we call a grammar language. This framework mirrors similar frameworks which break a data process down into intuitive elements such as the grammar of graphics. Our framework sought to require enough details to make accurate predictor descriptions while being simple enough that the combination of terminology can be used and understood by clinical domain experts.

We used a combination of autoregression, statistical summarization, and recursion techniques to transform source data using the grammar framework into a model-ready dataset capable of both single-point and interval-based prediction modeling. We then applied this approach to both a publicly available clinical dataset as well as a dataset sourced from our own research data warehouse. The intent of this application was to demonstrate a proof of concept and with a minimum amount of code, were able to achieve performance close to those attempted by other data preparation tools.

Addressing the original clinical problem of interest, we then applied our framework to domain expert selected predictors, preparing the data for an interval-based prediction model intended for maternal early warning systems to predict postpartum hemorrhage.

We found that our model outperforms existing early warning systems intended to predict postpartum hemorrhage. We also found that on average, we were able to obtain early warnings approximately 60 minutes prior to hemorrhage based on a specified risk threshold of greater than 10%. These findings indicate that a time-series approach to this clinical problem is likely more advantageous than a single-point prediction approach.

6.2 Implications

6.2.1 Maternal Early Warning Systems

Our model shows promising improvements in performance over existing early warning systems which predict postpartum hemorrhage. Despite being well accepted by institutions who implement trigger based systems, MEOWS, MEWC, and AWOB use thresholds based on general consensus [9], [31], [34]. The thresholds were determined by a consensus of domain experts on only a small number of variables. While this seems counterintuitive to our argument, given we are in support of domain expertise, to be clear, it is more advantageous to develop models which make a distinction between variables deemed to be important by domain experts from models developed using machine learning. Instead of risk determined by threshold, we take a hybrid approach using input from domain experts and literature and output using machine learning algorithms to determine risk.

This could lead the way to a shadow implementation running in the background while assessing its performance using streamed data and using existing notification systems both built within and outside the EHR system. Considering the two contemporary models already published, our time-series approach is likely more applicable to use in an early warning system for prediction of postpartum hemorrhage and more generalizable to other institutions with a similar patient population than the model proposed by Venkatesh et al. and Escobar et al. [37], [108].

While the model proposed by Escobar et al. evaluated multiple outcomes related to maternal care, the dataset had a very low prevalence of postpartum hemorrhage with virtually no alerts actually occurring postpartum [37]. This could be an indication that while performance may be acceptable at the composite level, for general maternal

comorbidities, it is unlikely to be effective specifically for postpartum hemorrhage which by definition is an acute condition which can only occur postpartum. Therefore, their model is unlikely to be generalizable to our institution with the sole purpose of predicting PPH.

A valid concern is the inability for our model to generalize to other institutions as we found in Chapter 3. While it is unclear whether the model is able to generalize at other institutions, the predictor mapping using wizard is quite detailed and allows other institutions to use similar time-series data preprocessing methods to refit local models calibrated to local data.

6.2.2 Opportunities to Improve Time-Series Modeling

Time-series modeling introduces many challenges, especially in the area of communication. Our framework is intended to not only enhance communication with a common language used to prepare data for modeling but allow this language to improve transportability of methods. The ability to use high performance computing can speed up processing time immensely as well as allow rapid prototyping to improve model performance.

6.3 Domain-informed predictors

While it is not known if using domain-informed predictors outperforms automated predictors, using them *can* improve the trust and credibility in the resulting model [109]. While this is largely dependent on the data collected and the outcome predicted, it may be difficult to determine if the domain experts do not know what good predictors are or if the outcome is not predictable by domain experts when modeling performance is poor.

Often in other disciplines the next step is to use critical thinking skills to conduct a root cause analysis. Some of this may be observed indirectly in the data through chart reviews of cases where patients experience the outcome to find predictors such as reviewing operating room reports, unverified vitals, nursing flowsheets etc. But sometimes the root cause is not captured in a secondary data source and often requires additional data gathering such conducting interviews or observational shadowing with clinicians to find clues. With consideration of postpartum hemorrhage, it may be beneficial to shadow

stakeholders to determine if there are considerations which are not directly observable in the data but may be revealed in behavioral observations. For example, alerts may be ignored during nursing shift change causing more bleeds in certain intervals or clinicians may be more accommodating to patients in their care choices which can lead to a higher likelihood of adverse events.

6.4 Limitations

This body of work has limitations related to existing theoretical work in this area of research. There is currently no evidence to suggest that domain expert-informed data preparation techniques, in favor of interpretability, commonly outperform automated preparation tools. Our data preparation framework can be used to create expert-informed predictors which are more interpretable, it's not known as to whether they commonly lead to better models. While we tested our framework on both a de-identified and proprietary dataset using feedback from experts throughout the project, validation of our framework performance has yet to be done on a large scale.

It is unclear whether wizard improves communication with domain experts or reduces human-spent hours on manual data preparation. While this tool was developed to meet the needs of the scientific research community, a large-scale qualitative analysis on user testing on user experience and user interface was outside the scope of work. As we did not perform a qualitative analysis to interpret domain experts' interactions with wizard, this may be an area of future work.

6.5 Critical appraisal and Reflection

I led weekly meetings over the course of two years with domain in experts in Obstetrics Anesthesia as well as Obstetrics and Gynecology to develop modeling predictors and interpret the results. While we found favorable performance using a time series approach over predictions made at a single point in time, we do not know how well these predictions stack up against existing rule-based models. I did not meet with nurses or patients following delivery to gain insight on how our work would affect clinical practice, and it is possible that this knowledge could have impacted our approach. We also did not

clinically implement the model, so it is unknown as to how well it would work in clinical practice prospectively.

In considering what I would do differently and to inform future doctoral students interested in analytical work in the clinical domain, I would advocate for transparency not only in methodological applications but also transparency in the rationale for decisions that are made because the context of a problem matters. Transparency in analytical methods is an important topic and one that I am passionate about, which is hopefully perceived in this body of work as transparency in methods that goes beyond simply sharing code. It can also be captured in a narrative form, which is what I have tried to elaborate on in the rationale for study design, capturing not only the decisions we made as a team but also the reason behind those decisions, which is often lost or only found in a domain niche.

There is also a tension when writing for a clinical audience (e.g., clinical journal) as opposed to an analytical audience (e.g., informatics journal). One of the drawbacks of targeting Chapters 3-5 for a clinical journal is that clinical journals often underemphasize the importance of methodological decisions so long so as the selected methods are reasonable.

I believe it is important to follow basic practice of the discipline to maintain structure and clarity in writing but allow that writing to deviate when necessary to incorporate details which make a research project unique. Writing about the domain experts' rationale serves as a narrative which can explain not only to future researchers the reason certain analytical decisions were made, but also to preserve this work for your future self to reflect upon. It is easier to remove detail than to resurrect it from code or a design journal.

6.6 Recommendations for Future Research

Our PPH prediction model, despite showing better performance with contemporary early warning systems, is still only a small part of what is required to improve maternal postpartum care. Stakeholder buy-in has often been shown to be the biggest roadblock to any broad implementation strategy of machine learning models [110]. This means cooperation and input from doctors, nurses, support staff, and patients alike in the search for an early warning system that communicates the right information to the right person

at the right time in order to take action. Additional research needs to be done to understand how clinicians are expected to interact with the model. Further work investigating how to respond to an alert by this model, what an appropriate risk threshold with consideration of workload (patients willing to evaluate), and what kind of clinical response is expected compared to currently utilized systems.

Our PPH model is the result of a transparent method of data preparation, which improves transparency for front-end to users (i.e., bedside providers). Future work should continue to improve model transparency by adequately communicating to care providers why the model made its predictions. The ability to predict postpartum hemorrhage is only part of the equation in clinical decision making. Another important aspect of clinical decision-making is supporting the clinician's ability to make informed decisions. This can be supplanted with a method to calculate Shapley values [111], which calculates variable importance at the individual prediction instead of the across the entire model such is the common practice and coincidentally the method we used in our work. Shapley values are calculated by comparing what a model predicts based on the feature presence and absence. It is done in every possible order to overcome limitations in the order of how the model sees these features.

Our grammar framework, wizard, is the result of collaboration and critical thinking among clinicians and data scientists but the challenges in time-series prediction modeling are not unique to clinical data. While the demonstrated applications have been in the clinical domain, the potential for application is intended to be broad for any type of data which is unevenly spaced and requires data transformation for supervised prediction modeling. A broader use of the framework could establish common practice for a time-series approach in certain types of data for clinical prediction modeling.

Bibliography

[1] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, May 2003.

[2] D. Pyle, *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[3] J.-H. Lin and P. J. Haug, “Data preparation framework for preprocessing clinical data in data mining,” *AMIA Annu. Symp. Proc.*, pp. 489–493, 2006.

[4] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, “Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data,” *J. Am. Med. Inform. Assoc.*, vol. 27, no. 12, pp. 1921–1934, 2020.

[5] H. Wickham, “A layered grammar of graphics,” *J. Comput. Graph. Stat.*, vol. 19, no. 1, pp. 3–28, Jan. 2010.

[6] “The Grammar of Graphics: The ggplot2 Package,” *R Graphics*. pp. 173–200, 2018.

[7] L. Wilkinson, *The Grammar of Graphics*. Springer Science & Business Media, 2013.

[8] T. T. Klumpner *et al.*, “Use of a Novel Electronic Maternal Surveillance System and the Maternal Early Warning Criteria to Detect Severe Postpartum Hemorrhage,” *Anesthesia & Analgesia*, vol. Publish Ahead of Print, Feb. 2020.

[9] T. T. Klumpner, J. A. Kountanis, E. S. Langen, R. D. Smith, and K. K. Tremper, “Use of a novel electronic maternal surveillance system to generate automated alerts on the labor and delivery unit,” *BMC Anesthesiol.*, vol. 18, no. 1, p. 78, Jun. 2018.

[10] S. C. Reale, S. R. Easter, X. Xu, B. T. Bateman, and M. K. Farber, “Trends in Postpartum Hemorrhage in the United States From 2010 to 2014,” *Anesth. Analg.*, vol. 130, no. 5, pp. e119–e122, May 2020.

[11] E. K. Main *et al.*, “National Partnership for Maternal Safety Consensus Bundle on Obstetric Hemorrhage,” *J. Midwifery Womens. Health*, vol. 60, no. 4, pp. 458–464, 2015.

[12] N. L. Downing *et al.*, “Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation,” *BMJ Qual. Saf.*, vol. 28, no. 9, pp. 762–768, Sep. 2019.

[13] S. Nemati, A. Holder, F. Razmi, M. Stanley, G. Clifford, and T. Buchman, “An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU,” *Crit. Care Med.*, vol. 46, no. 4, pp. 547–553, Apr. 2018.

[14] R. J. Delahanty, J. Alvarez, L. M. Flynn, R. L. Sherwin, and S. S. Jones, “Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis,” *Ann. Emerg. Med.*, vol. 73, no. 4, pp. 334–344, Apr. 2019.

[15] J. Futoma *et al.*, “An Improved Multi-Output Gaussian Process RNN with Real-Time

Validation for Early Sepsis Detection," *arXiv:1708.05894 [stat]*, Aug. 2017.
[16] J. Futoma, S. Hariharan, and K. Heller, "Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier," *arXiv:1706.04152 [stat]*, Jun. 2017.
[17] S. M. Brown, J. Jones, K. G. Kuttler, R. K. Keddington, T. L. Allen, and P. Haug, "Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department," *BMC Emerg. Med.*, vol. 16, no. 1, p. 31, Aug. 2016.
[18] S. W. Thiel *et al.*, "Early prediction of septic shock in hospitalized patients," *J. Hosp. Med.*, vol. 5, no. 1, Jan. 2010.
[19] A. M. Sawyer *et al.*, "Implementation of a real-time computerized sepsis alert in nonintensive care unit patients," *Crit. Care Med.*, vol. 39, no. 3, pp. 469–473, Mar. 2011.
[20] M. Moor, M. Horn, B. Rieck, D. Roqueiro, and K. Borgwardt, "Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping," *arXiv:1902.01659 [cs, stat]*, Feb. 2019.
[21] T. Desautels *et al.*, "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Med Inform*, vol. 4, no. 3, p. e28, Sep. 2016.
[22] N. Tomasev *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, Aug. 2019.
[23] H. Mohamadlou *et al.*, "Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data," *Can J Kidney Health Dis*, vol. 5, p. 2054358118776326, 2018.
[24] J. L. Koyner, R. Adhikari, D. P. Edelson, and M. M. Churpek, "Development of a Multicenter Ward–Based AKI Prediction Model," *Clin. J. Am. Soc. Nephrol.*, vol. 11, no. 11, pp. 1935–1943, Nov. 2016.
[25] J. L. Koyner, K. A. Carey, D. P. Edelson, and M. M. Churpek, "The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model," *Crit. Care Med.*, vol. 46, no. 7, pp. 1070–1077, 07 2018.
[26] H.-C. Lee *et al.*, "Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model," *J. Clin. Med. Res.*, vol. 7, no. 11, Nov. 2018.
[27] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, Jun. 2019.
[28] T. M. Deist *et al.*, "Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers," *Med. Phys.*, vol. 45, no. 7, pp. 3449–3459, Jul. 2018.
[29] D. Chestnut *et al.*, "Chestnut's Obstetric Anesthesia: Principles and Practice." 2019.
[30] B. Haibe-Kains *et al.*, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, Oct. 2020.
[31] S. Singh, A. McGlennan, A. England, and R. Simons, "A validation study of the CEMACH recommended modified early obstetric warning system (MEOWS)," *Anaesthesia*, vol. 67, no. 1, pp. 12–18, Jan. 2012.
[32] J. M. Mhyre, "The Maternal Early Warning Criteria: A Proposal from the National Partnership for Maternal Safety | Elsevier Enhanced Reader," 2014. [Online].

Available: https://reader.elsevier.com/reader/sd/pii/S0884217515316105?token=15240FDA2FD695D6F4847058C44F7403DD89BE1F5B03BF18A1135EA11A08402C289B77914E439D318E13649E8DEC8BBB. [Accessed: 05-Aug-2019].
[33] Y. Xiao, C. F. Mackenzie, F. J. Seagull, and M. Jaberi, “Managing the Monitors: An Analysis of Alarm Silencing Activities during an Anesthetic Procedure,” *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.*, vol. 44, no. 26, pp. 250–253, Jul. 2000.
[34] D. E. Arnolds, A. Smith, J. M. Banayan, R. Holt, and B. M. Scavone, “National Partnership for Maternal Safety Recommended Maternal Early Warning Criteria Are Associated With Maternal Morbidity,” *Anesthesia & Analgesia*, p. 1, Apr. 2018.
[35] A. M. Friedman, M. L. Campbell, C. R. Kline, S. Wiesner, M. E. D’Alton, and L. E. Shields, “Implementing Obstetric Early Warning Systems,” *AJP Rep*, vol. 8, no. 2, pp. e79–e84, Apr. 2018.
[36] K. K. Venkatesh *et al.*, “Machine Learning and Statistical Models to Predict Postpartum Hemorrhage,” *Obstet. Gynecol.*, vol. 135, no. 4, pp. 935–944, Apr. 2020.
[37] G. J. Escobar, L. Soltesz, A. Schuler, H. Niki, I. Malenica, and C. Lee, “Prediction of obstetrical and fetal complications using automated electronic health record data,” *Am. J. Obstet. Gynecol.*, vol. 224, no. 2, pp. 137-147.e7, Feb. 2021.
[38] G. J. Escobar, N. R. Gupta, E. M. Walsh, L. Soltesz, S. M. Terry, and P. Kipnis, “Automated Early Detection of Obstetric Complications: Theoretical and Methodological Considerations,” *Am. J. Obstet. Gynecol.*, vol. 220, no. 4, pp. 297–307, 2019.
[39] T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003.
[40] M. A. Munson, “A study on the importance of and time spent on different modeling steps,” *SIGKDD Explor. Newsl.*, vol. 13, no. 2, pp. 65–71, May 2012.
[41] S. F. Crone, S. Lessmann, and R. Stahlbock, “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,” *Eur. J. Oper. Res.*, vol. 173, no. 3, pp. 781–800, Sep. 2006.
[42] Y. Li, T. Jann, and P. Vera-Licona, “Benchmarking time-series data discretization on inference methods,” *Bioinformatics*, vol. 35, no. 17, pp. 3102–3109, Sep. 2019.
[43] J. C. Ferrão, M. D. Oliveira, F. Janela, and H. M. G. Martins, “Preprocessing structured clinical data for predictive modeling and decision support. A roadmap to tackle the challenges,” *Appl. Clin. Inform.*, vol. 7, no. 4, pp. 1135–1153, Dec. 2016.
[44] S. Lemeshow, R. X. Sturdivant, D. W. Hosmer Jr, and D. W. Hosmer Jr, *Applied Logistic Regression*. New York, UNITED STATES: John Wiley & Sons, Incorporated, 2013.
[45] J. A. K. Suykens and J. Vandewalle, “Least Squares Support Vector Machine Classifiers,” *Neural Process. Letters*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[46] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[47] L. Breiman, *Classification And Regression Trees*. Routledge, 1984.
[48] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[49] M. N. Wright and A. Ziegler, “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R,” *J. Stat. Softw.*, vol. 77, no. 1, 2017.

[50] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting,” p. 14, 1999.
[51] L. G. Valiant, “A theory of the learnable,” *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984.
[52] E. Bauer and R. Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Mach. Learn.*, vol. 36, no. 1, pp. 105–139, Jul. 1999.
[53] J. V. Tu, “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes,” *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996.
[54] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, Pittsburgh, Pennsylvania, 2006.
[55] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 281, Dec. 2019.
[56] B. T. Bateman, M. F. Berman, L. E. Riley, and L. R. Leffert, “The epidemiology of postpartum hemorrhage in a large, nationwide sample of deliveries,” *Anesth. Analg.*, vol. 110, no. 5, pp. 1368–1373, May 2010.
[57] S. F. Bell *et al.*, “Incidence of postpartum haemorrhage defined by quantitative blood loss measurement: a national cohort,” *BMC Pregnancy Childbirth*, vol. 20, May 2020.
[58] “Quantitative Blood Loss in Obstetric Hemorrhage: ACOG COMMITTEE OPINION SUMMARY, Number 794,” *Obstet. Gynecol.*, vol. 134, no. 6, pp. 1368–1369, Dec. 2019.
[59] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, Jan. 2012.
[60] A. Subbaswamy and S. Saria, “From development to deployment: dataset shift, causality, and shift-stable models in health AI,” *Biostatistics*, vol. 21, no. 2, pp. 345–352, Apr. 2020.
[61] T. P. A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. M. Moons, “A new framework to enhance the interpretation of external validation studies of clinical prediction models,” *J. Clin. Epidemiol.*, vol. 68, no. 3, pp. 279–289, Mar. 2015.
[62] “Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.” Food & Drug Administration, 01/2021.
[63] “reVITALize: Obstetrics Data Definitions.” [Online]. Available: https://www.acog.org/en/Practice Management/Health IT and Clinical Informatics/reVITALize Obstetrics Data Definitions. [Accessed: 29-Sep-2020].
[64] J. Zhang *et al.*, “Contemporary Cesarean Delivery Practice in the United States,” *Am. J. Obstet. Gynecol.*, vol. 203, no. 4, pp. 326.e1-326.e10, Oct. 2010.
[65] M. J. Pencina and R. B. D’Agostino, “Evaluating Discrimination of Risk Prediction Models: The C Statistic,” *JAMA*, vol. 314, no. 10, p. 1063, Sep. 2015.
[66] B. Van Calster *et al.*, “Calibration: the Achilles heel of predictive analytics,” *BMC Med.*, vol. 17, no. 1, p. 230, Dec. 2019.
[67] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W.

Steyerberg, “A calibration hierarchy for risk models was defined: from utopia to empirical data,” *J. Clin. Epidemiol.*, vol. 74, pp. 167–176, Jun. 2016.
[68] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[69] Team and TRDC, “The R Project for Statistical Computing,” *http://www.r-project.org*, 2008.
[70] “R Interface for the H2O Scalable Machine Learning Platform.” [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/index.html. [Accessed: 28-Sep-2020].
[71] K. Singh, *ML4LHS/runway*. Machine Learning for Learning Health Systems Lab, 2020.
[72] X. Robin *et al.*, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 12/2011.
[73] S. R. Meyer, *seanrmeyer/meyer_et_al_2021a*. 2021.
[74] M. Kuhn, B. Butcher, and B. J. Smith, *Feature Engineering and Selection: A Practical Approach for Predictive Models: by Max Kuhn and Kjell Johnson. Boca Raton, FL: Chapman & Hall/CRC Press, 2019, xv + 297 pp., $79.95(H), ISBN: 978-1-13-807922-9*, vol. 74. 2020.
[75] R. E. Schapire, “The strength of weak learnability,” *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, Jun. 1990.
[76] N. Tangri *et al.*, “Multinational Assessment of Accuracy of Equations for Predicting Risk of Kidney Failure: A Meta-analysis,” *JAMA*, vol. 315, no. 2, p. 164, Jan. 2016.
[77] E. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, Second. New York: Springer-Verlag, 2019.
[78] K. G. M. Moons *et al.*, “Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration,” *Ann. Intern. Med.*, vol. 162, no. 1, p. W1, Jan. 2015.
[79] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, “Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review,” *J. Am. Med. Inform. Assoc.*, vol. 24, no. 1, pp. 198–208, Jan. 2017.
[80] H. Tsai and K. S. Chan, “Temporal Aggregation of Stationary And Nonstationary Discrete-Time Processes,” *J. Time Ser. Anal.*, vol. 26, no. 4, pp. 613–624, 2005.
[81] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, “Towards parameter-free data mining,” in *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’04*, 2004, p. 206.
[82] H. Wickham, “Tidy Data,” *J. Stat. Softw.*, vol. 59, no. 1, pp. 1–23, Sep. 2014.
[83] P. D. Allison, “Discrete-Time Methods for the Analysis of Event Histories,” *Sociol. Methodol.*, vol. 13, p. 61, 1982.
[84] “Summarise each group to fewer rows.” [Online]. Available: https://dplyr.tidyverse.org/reference/summarise.html. [Accessed: 08-Jul-2021].
[85] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci Data*, vol. 3, no. 1, pp. 1–9, May 2016.
[86] “mimic3_experiments · master · MLD3 / FIDDLE,” *GitLab*. [Online]. Available: https://gitlab.eecs.umich.edu/mld3/FIDDLE/tree/master/mimic3_experiments. [Accessed: 09-Jan-2020].

[87] “Yottabyte Research Cloud.” [Online]. Available: https://arc.umich.edu/ybrc/. [Accessed: 03-Jul-2021].
[88] “Armis2 (HIPAA-aligned HPC Cluster).” [Online]. Available: https://arc.umich.edu/armis2/. [Accessed: 02-Jul-2021].
[89] “Great Lakes.” [Online]. Available: https://arc.umich.edu/greatlakes/. [Accessed: 02-Jul-2021].
[90] S. Wang, M. B. A. McDermott, G. Chauhan, M. C. Hughes, T. Naumann, and M. Ghassemi, “MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III,” *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, Apr. 2020.
[91] E. K. Main, C. L. McCain, C. H. Morton, S. Holtby, and E. S. Lawton, “Pregnancy-related mortality in California: causes, characteristics, and improvement opportunities,” *Obstet. Gynecol.*, vol. 125, no. 4, pp. 938–947, Apr. 2015.
[92] “[No title].” [Online]. Available: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pregnancy-mortality-surveillance-system.htm. [Accessed: 30-Mar-2021].
[93] P. Toledo, R. J. McCarthy, B. J. Hewlett, P. C. Fitzgerald, and C. A. Wong, “The accuracy of blood loss estimation after simulated vaginal delivery,” *Anesth. Analg.*, vol. 105, no. 6, pp. 1736–40, table of contents, Dec. 2007.
[94] W. Prasertcharoensuk, U. Swadpanich, and P. Lumbiganon, “Accuracy of the blood loss estimation in the third stage of labor,” *Int. J. Gynaecol. Obstet.*, vol. 71, no. 1, pp. 69–70, Oct. 2000.
[95] A. Borovac-Pinheiro *et al.*, “Postpartum hemorrhage: new insights for definition and diagnosis,” *Am. J. Obstet. Gynecol.*, vol. 219, no. 2, pp. 162–168, Aug. 2018.
[96] Committee on Practice Bulletins-Obstetrics, “Practice Bulletin No. 183: Postpartum Hemorrhage,” *Obstet. Gynecol.*, vol. 130, no. 4, pp. e168–e186, Oct. 2017.
[97] J. M. Mhyre *et al.*, “The maternal early warning criteria: a proposal from the national partnership for maternal safety,” *Obstet. Gynecol.*, vol. 124, no. 4, pp. 782–786, Oct. 2014.
[98] K. K. Venkatesh *et al.*, “256: Machine learning-based prediction models for postpartum hemorrhage,” *Am. J. Obstet. Gynecol.*, vol. 222, no. 1, pp. S175–S176, Jan. 2020.
[99] T. T. Klumpner *et al.*, “User Perceptions of an Electronic Maternal Alerting System,” *A A Pract*, vol. 14, no. 11, p. e01308, Sep. 2020.
[100] A. C. J. W. Janssens and F. K. Martens, “Reflection on modern methods: Revisiting the area under the ROC Curve,” *Int. J. Epidemiol.*, vol. 49, no. 4, pp. 1397–1403, Aug. 2020.
[101] “Gradient Boosting Machine (GBM) — H2O 3.32.1.3 documentation.” [Online]. Available: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html. [Accessed: 20-May-2021].
[102] F. E. Harrell, “calibrate: Resampling Model Calibration in rms: Regression Modeling Strategies,” 18-Mar-2021. [Online]. Available: https://rdrr.io/cran/rms/man/calibrate.html. [Accessed: 30-Jun-2021].
[103] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013.
[104] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,”

Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
[105] "Variable Importance — H2O 3.32.1.3 documentation." [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html. [Accessed: 29-Jun-2021].
[106] A. J. Vickers, B. van Calster, and E. W. Steyerberg, "A simple, step-by-step guide to interpreting decision curve analysis," *Diagn Progn Res*, vol. 3, p. 18, Oct. 2019.
[107] A. J. Vickers and E. B. Elkin, "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 11/2006.
[108] K. K. Venkatesh *et al.*, "Supplemental Document I, Machine Learning and Statistical Models to Predict Postpartum Hemorrhage," *Obstet. Gynecol.*, vol. 135, no. 4, pp. 935–944, Apr. 2020.
[109] J. Wang, J. Oh, H. Wang, and J. Wiens, "Learning Credible Models," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, United Kingdom, 2018, pp. 2417–2426.
[110] J. Watson *et al.*, "Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers?," *JAMIA Open*, vol. 3, no. 2, pp. 167–172, Jul. 2020.
[111] A. Ghorbani and J. Zou, "Data Shapley: Equitable Valuation of Data for Machine Learning," in *International Conference on Machine Learning*, 2019, pp. 2242–2251.