

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Semiparametric Prediction, Variable Importance, and Effect Estimation in Critical Care

Permalink

<https://escholarship.org/uc/item/5x10t818>

Author

Decker, Anna

Publication Date

2014

Peer reviewed|Thesis/dissertation

Semiparametric Prediction, Variable Importance, and
Effect Estimation in Critical Care

By

Anna Decker

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Biostatistics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Alan E. Hubbard, Chair
Professor Mark J. van der Laan
Professor Lisa F. Barcellos

Spring 2014

Abstract

Semiparametric Prediction, Variable Importance, and Effect Estimation in Critical Care

Anna Decker

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan Hubbard, Chair

Trauma injury is one of the leading causes of death in the United States, accounting for over 120,000 deaths in 2010 according to the CDC. Understanding the underlying mechanisms and improving the treatment of trauma is of great clinical and public health interest. The systematic collection and study of critical care data originated in combat conflicts and wars and more recently to civilian centers. Improving patient outcomes, the quality of care received, and identifying high-risk patients are unmet needs in this field.

Clinicians rely on their intuition, training, and heuristic scoring systems to identify patients who are likely to die or experience other outcomes such as the need for a massive transfusion, which resuscitates the patient via the infusion of blood products such as plasma, platelets, and red blood cells. We assessed the ability of measured covariates to predict various clinical outcomes, demonstrate the utility of machine-learning prediction algorithms, and examined the predictive performance of a commonly-used score to predict massive transfusion. This highlights the need for a principled approach to predicting outcomes that does not rely only on *ad hoc* procedures.

In addition to the prediction of clinical outcomes, we defined a measure of variable importance for ranking predictors based on their relationship with the outcome of interest. This parameter was motivated by causal inference and requires a systematic approach to the question of interest that helps translate it into a parameter with a clinically meaningful interpretation rather and maintains transparency about the assumptions required to deem the parameter a causal effect. We apply this procedure to gene expression data from critically injured patients to illuminate how the coagulation and inflammation pathways react to trauma injury.

Finally, we compare the quality of care received at different trauma center types around the United States using another parameter motivated by causal inference. This allowed us to simulate what would have happened to a patient if they had been treated at a

different trauma center and obtain an objective comparison that identified sites where severely injured patients would benefit most from being treated.

This research highlights the utility of causal inference for framing problems, motivating clinically meaningful statistical parameters, and interpreting the results. We also advocate for the use of semiparametric prediction algorithms to allow for greater flexibility in modeling assumptions and demonstrate their performance in practice.

Acknowledgements

I would like to express my very great appreciation to Dr. Alan Hubbard for his valuable and constructive suggestions during the planning and development of this research. His willingness to give his time so generously has been very much appreciated and his enthusiastic encouragement throughout my doctoral studies has been invaluable.

I would also like to thank Dr. Mark van der Laan and Dr. Lisa Barcellos for their useful and constructive recommendations on this research. My grateful thanks are also extended to Sharon Norris and Burke Bundy for their administrative support throughout this process. Special thanks should be given to Dr. Mitchell Cohen and the research team at San Francisco General Hospital for their invaluable subject-matter knowledge and constructive feedback on this research.

Finally, I wish to thank my family and friends for their support and encouragement throughout my studies.

Chapter 1

Introduction

1.1 Introduction

Trauma injury is one of the leading causes of death in the United States, accounting for over 120,000 deaths in 2010 according to the CDC. Understanding the underlying mechanisms and improving the treatment of trauma is of great clinical and public health interest. Questions of interest in this area include the prediction of clinical outcomes using available data, determining which variables are the most informative over time with respect to these clinical outcomes, comparing the quality of care at different trauma centers, and assessing the efficacy of patient resuscitation via the infusion of blood products. Current practice in this field can benefit greatly from the use of causal inference to motivate clinically meaningful and interpretable statistical parameters, even if the assumptions required for a causal interpretation are not met. Further, the use of mis-specified parametric models is common in the critical care literature. We advocate the use of a machine-learning prediction algorithm that has desirable theoretical optimality properties and, in practice, frees the user from having to choose a single prediction model. We demonstrated the utility of causal inference and machine learning in addressing key questions in critical care and provided clinically meaningful results to help inform clinicians' decisions.

1.1.1 History of trauma care and data

Trauma care in the United States was historically and understandably linked to wars. While surgical care remained relatively primitive, the idea of moving field hospitals as close to the battle lines to reduce the time from injury to surgical care was implemented as early as the War of 1812 [1]. During the American Civil War, both the Union and Confederate

CHAPTER 1. INTRODUCTION

sides made great advances in the implementation of systematic care of injured people and publishing of the medical and surgical history of the war [1]. The importance of a sanitary environment in the treatment of open wounds was emphasized and conditions improved for the Union army [1]. In 1895 the invention of the xray greatly advanced the diagnosis of traumatic wounds [2]. By World War I, blood transfusions were being used extensively and with great success and a commission was appointed to study shock and resuscitation in United States military members [1, 3]. Before World War II, the United States began laying the foundations of modern trauma systems with the formation of committees and programs to enforce standardization of care across hospitals [1, 3]. In the Korean and Vietnam conflicts, soldiers benefited from shorter periods between injury and the initiation of treatment as well as improved surgical techniques. The lessons learned from military conflicts have been applied to the care of civilians. In 1966, two trauma centers were established in San Francisco and Chicago with the aim of taking a systematic approach to trauma care [1]. Trauma centers proliferated around the United States so extensively that new ranking systems, model care systems, and quality improvement programs had to be implemented. Currently, there are 407 trauma centers accredited by the American College of Surgeons with ranks from I (highest, n= 112) to III (lowest, n = 58). Level I trauma centers are capable of providing total care for every aspect of injury– from prevention through rehabilitation whereas lower level centers are focused on the initiation of definitive care and stabilization of patients for transfer to other hospitals. There is some evidence that patients treated at Level I centers have better survival than those at lower ranked hospitals [4] The collection and maintenance of patient data has improved, especially with the use of electronic medical records. However, there is still considerable variability between trauma centers with respect to clinical practice and record-keeping. Multi-center studies are commonly implemented in the study of critical care in order to study clinically relevant questions where large number of patients are required and to make the results more generalizable. However, objectively comparing the quality of care at each center has not been a priority in the trauma community.

The treatment of trauma injury has improved with the invention of new technologies, but the proliferation of data available on patients is simultaneously a boon and hindrance. Clinicians have access to data on patients as soon as they arrive in the emergency department and the patients are monitored constantly throughout their treatment with occasional lab measurements, xrays, surgeries, and blood transfusions. These measurements may or may not be used in the decision to allocate treatment or to predict whether a patient will die in the next few hours. The effective communication and condensation of these vast amounts of data in the chaotic environment of the emergency department, operating room and intensive care unit is vital to the improvement of treatment. Indeed, modern care is confounded by is a continuous stream of multivariate data consisting of demographic information, injury data, medical staff documentation, laboratory testing

and continuous multivariate physiologic monitors. The dramatic increase in available information has led to a data-rich care environment that can be cognitively burdensome to the practitioner. Despite the improvements in, and increasing reliance on monitoring technology, these multivariate data are still recorded and no current method exists to integrate, computationally refine and make sense of these data. Even in hospitals where the paper chart has been replaced by a computerized medical record, these systems are not adequate for the tracking and analysis of complex multivariate relationships. This antiquated data collection and presentation limits the clinicians' ability to understand the complex relationship between variables and precludes longitudinal analysis of trends and developing patient pathophysiology, resulting in univariate treatment of complex multivariate post injury physiology. Clinicians base treatment decisions on their intuition and experience developed throughout their training their career. Despite the proliferation of data, medicine remains an process based on clinician gestalt and experience rather than a science based on mathematical modeling, protocols, and predictions. Thus, there is an unmet need in critical care to sift through the vast amounts of available data, determine what information is important at a given time, and use this information in the most efficient possible way.

1.1.2 Current statistical practice

Current statistical practice in the trauma literature relies heavily on the use of possibly mis-specified parametric regression models, where the parameter of interest is the coefficient in the regression equation. There is little biological evidence that the functional relationships between predictors and outcomes would be linear, but the bias in the literature requires parameters with a familiar, although not clinically meaningful interpretation.

There is also heavy emphasis on establishing cutpoints for variables or simple combinations of variables with the aim of categorizing patients into low- and high-risk groups. For example, the Abbreviated Injury Scale (AIS) classifies every injury in each of nine body regions (head, face, neck, extremity, etc.) according to its severity on a six-point ordinal scale (minor to maximal). The top three AIS scores are then squared and summed to produce the Injury Severity Score (ISS), which is itself commonly categorized into mild (< 15), moderate (15-25) and severe (> 25) [5, 6]. This score is used extensively to identify individuals at high risk for mortality, complications and hospitalization time after trauma.

Another scoring system is called the Assessment of Blood Consumption (ABC) score, which is designed to predict the need for massive transfusion and standardize the initiation of massive transfusion protocols across hospitals [7]. This score was based on clinician

interviews regarding their clinical criteria for activation of massive transfusion, and consisted of four dichotomous components: whether the injury is of a penetrating (as opposed to blunt) nature, whether the patient’s systolic blood pressure was 90 mmHg or higher in the emergency department, whether their heart rate was 120 beats per minute or greater, and whether they had a positive Focused assessment with sonography for trauma (FAST) scan, which consists of a bedside ultrasound to screen for blood around the heart [7]. This score improved upon previously used scoring systems (TASH and McLaughlin scores) but not significantly and was not data-driven, but clinician decision driven, suggesting there may be patients who could have benefited from a massive transfusion that were not identified using this method [7].

In the trauma literature, little work has been done to on the ability to predict outcomes with measured patient data using any algorithm other than regression and the comparisons are rarely “fair”, for example, the performances of the scores were assessed using data that were used to build the score. Additionally, clinicians have been relying on empirical evidence and case studies to guide their decisions rather than systematic studies of trauma care with robust statistical methods. We advocate for the use of causal inference to motivate clinically meaningful parameters of interest.

1.1.3 Causal inference roadmap

Many of the questions of interest in critical care are causal in nature, that is, clinicians are interested how the outcomes of interest would change if, for example, the amount of blood products a patient received had been different rather than the association between these outcomes of interest and blood product usage. In practice, these so-called counterfactual outcomes are not observable and the “natural experiment” that generated our observed data data does not correspond to the ideal experiment we are interested in [8, 9]. Parameters in the causal inference paradigm summarize how parameters of the underlying distribution of the data would change if the experimental conditions changed, making inference about such parameters difficult since we do not fully observe this distribution [8]. In contrast, statistical parameters are based on the joint distribution of past data, which can be extended to future events only if we assume the experimental conditions did not change, which is unlikely to be true. Under some assumptions, we can link the causal and statistical paradigms to make causal inferences from observed data. The causal inference roadmap for this is as follows:

1. Specify the question of interest, that is, the target population and what we want to learn about it. For example, the effect of the infusion of plasma on the probability of future mortality in a population of trauma patients.

2. Specify a structural causal model, which encodes the background knowledge about the system under study. This may be a system of equations or a directed acyclic graph (DAG) [8]. In the blood product example, this would include factors that influence the decision to start treatment, confounders of the effect of blood products on future mortality, and other major determinants of mortality in trauma patients. This formal representation of the background knowledge helps maintain transparency about the relationships between observed and unobserved variables. Here, if the relationship between two variables is known, for example, if treatment is randomized, this information can be included in the causal model.
3. Specify the interventions of interest, which represent hypothetical changes on the causal model that correspond to the ideal experiment. For example, the interventions of interest in the blood product example would be to deterministically have everyone receive plasma or not. These interventions generate so-called counterfactual, or potential outcomes, which correspond to the outcome an individual would have had, if, possible contrary-to-fact, they had had a certain treatment [10].
4. Specify the causal parameter of interest, which is a function of the counterfactual outcome distributions. The mean change in the probability of mortality under interventions forcing every patient to have plasma infused as opposed to no patient having plasma infused corresponds to the average treatment effect (ATE).
5. The causal parameter of interest may or may not be identifiable as a parameter of the observed data distribution. In this step, we assess whether background knowledge about confounding and mediation of effects is sufficient or whether we need to make some assumptions such as a randomization or positivity assumption in order to use the observed data to draw causal conclusions.
6. If the causal parameter is identifiable as a parameter of the observed data distribution, we commit to that statistical parameter. It is still possible to proceed with estimation of a statistical parameter that is only causal under some assumptions as long as those assumptions are made clear and explicitly represent this limitation and use it to inform the interpretation of the parameter after estimation.
7. Estimation of the parameter of the observed data distribution should respect the limits of the available background knowledge by using non- or semi-parametric estimators. There are various estimation approaches we can use, such as regression, inverse weighting, or more robust approaches, but the estimation step should not introduce new, unneeded model assumptions at this stage. We introduce and examine the finite predictive performances of a machine-learning algorithm, SuperLearner, below, to avoid unnecessary assumptions.
8. Interpretation of the results after estimation requires assessing the statistical signif-

ificance as well as selecting among a hierarchy of interpretations ranging from purely associational to approximating a hypothetical experiment. If we believe all the assumptions required for a causal interpretation are met, then the parameter estimate is what we would have seen in a perfectly executed randomized control trial. If we do not believe them, the parameter may still be interpreted as a statistical parameter.

This roadmap allows for the definition of causal questions, rigorous expression of causal assumptions, and forces us to evaluate whether our data and assumptions are sufficient to answer the question of interest. Structural causal models are a useful tool for concatenating background knowledge and communicating with subject-matter experts. Additionally, this roadmap enforces transparency about what is and is not observed in the data as well as assisting clinicians in determining what their ideal experiment would involve. It allows us to derive clinically meaningful parameters with a straightforward interpretation that clinicians can utilize in practice.

1.2 Data sources

The systematic collection of data from emergency departments requires substantial coordination and organization [11, 12]. Even with the rising utility and efficiency of electronic medical records, data collection and maintenance remains a challenge due to the high-dimensional nature of available patient measurements. This section details the sources of clinical and gene expression data used for analysis.

1.2.1 The PRospective Observational Multi-center Massive Transfusion sTudy (PROMMTT)

The PRospective, Observational Multi-center Massive Transfusion sTudy (PROMMTT) was a prospective, multi-center, observational cohort study that enrolled 1,245 individuals at ten level-one trauma centers from around the United States [13]. This study was motivated by the fact that uncontrollable hemorrhage after injury is the leading cause of potentially preventable death (as opposed to traumatic brain injury or multiple organ failure), which occurs quickly and is associated with the massive transfusion of blood products [14, 15]. Historically, whole blood was used in the resuscitation of trauma patients until the 1970's, when the separation of blood into component parts (crystalloid, red blood cells (RBC), plasma, and plaetlets) became commonplace and the infusion of different ratios of blood products gained usage [16, 17]. Motivated by the treatment of United States military combat casualties with substantial bleeding in Afghanistan and Iraq, a new resuscitation strategy called damage control resuscitation started being used in civilian

hospitals that showed that infusing a 1:1:1 ratio of plasma:platelet:RBC and minimizing crystalloid could avert or reverse coagulopathy, acidosis, and hypothermia [18]. Given conflicting findings regarding the association of different ratios of blood products, the PROMMTT researchers aimed to guide uniform transfusion practices for trauma patients with substantial bleeding after injury across variable level I trauma centers.

Patients had to survive at least 30 minutes after injury to be enrolled in the study upon arrival to the emergency department. Demographic, health status, treatment, and outcome measurements were taken on these individuals. The primary exposure of interest was the resuscitation of patients via the use of blood products (plasma, platelets, and red blood cells), so cumulative units of each of these products were collected at 30 minute intervals during the first six hours of treatment. The primary outcome of interest was in-hospital mortality, but other clinical outcomes included the initiation of massive transfusion, multiple organ failure, substantial bleeding, and complications. Underlying patient populations and transfusion practices differed among the hospitals, which confounded the effect of the transfusion of blood products on clinical outcomes. PROMMTT was the largest study to collect real-time prospective data on trauma patients, enrolling on average, 5 patients a week [11]

The main PROMMTT analysis found that higher ratios of blood products conferred a survival benefit at 6 hours, when hemorrhagic death predominated, but were not associated with later mortality. Other papers that analyzed these data examine questions related to pre-hospital interventions, early resuscitation strategies, coagulopathy, and improvements on the existing scoring systems such as FAST. However, there has been relatively little focus on predicting outcomes in an efficient, unbiased way nor studies of variability across hospitals.

A map of the sites involved in PROMMTT, which includes the distribution of injury severity scores (ISS) at each site, is shown in Figure 1.1 and the centers are identified in Table 1.1. All centers had both severe and moderately injured patients but some, such as Brook Army Medical Center (BAMC), had substantially more severely injured individuals than moderate. The covariates and outcomes of interest are summarized in Tables 1.2 and 1.3. The covariates of interest are commonly measured in every emergency department and were identified as main predictors of the outcomes. Injury severity, described earlier, is commonly used to categorize injury types [5, 6]. Increased body mass index (BMI), measured in kg/m^2 , has been associated with outcomes such as respiratory failure, kidney failure, multiple organ failure, and excessive clotting [19]. The CDC reported that men were more 3.4 times more likely to die from traumatic brain injury, 6 times more likely to be injured using a firearm [20]. Additionally, there has been evidence of different brain chemistries between men and women but no consensus on whether men or women fare worse and injury [20]. Patient age, measured in years in PROMMTT, is

also an important factor that clinicians take into account during treatment [21]. Studies of race/ethnicity disparities in recovery after injury suggest that different race/ethnicity groups experience different propensities for clinical outcomes, likely due to a combination of genetic factors and the quality of care that is accessible [22, 23]. Penetrating trauma is usually more severe and unpredictable than blunt trauma and most papers adjust for or stratify by these groups [24]. The use of anticoagulants such as Warfarin affects the clotting time of patients and can have a detrimental effect on patient survival given the pro-hemorrhagic properties of the drugs [25]. Systolic blood pressure and heart rate are classical measurements of patient health, are measured almost continuously on patients, and are integral to the calculation of the ABC score, described above, which predicts massive transfusion [7]. The Glasgow Coma Scale was developed in 1974 to determine the conscious state of a person and ranges from 3 (indicating deep unconsciousness) to 15 [26]. It combines categories of eye, verbal, and motor response to identify high-risk patients after injury [26]. The International Normalized Ratio (INR) and Prothrombin Time (PTT), which are a measurement of blood clotting time have been used to identify patients who may require a massive transfusion and experience early mortality [27, 28]. Low platelet counts have been associated with hemorrhage and injury severity [29]. Determining a patient’s hemoglobin can help estimate blood loss [30]. The acidity of circulating blood, as measured by base deficit is commonly used as a predictor of transfusion requirements and risk of complications [31, 32]. Finally, the Focused Assessment with Sonography for Trauma (FAST) ultrasound to detect blood around the heart is a method for assessing cardiac, abdominal, and throacic injuries [33]. The outcomes of interest included mortality at 2 hours, 6 hours, 24 hours, and overall, massive transfusion (as reported by each center and calculated from the available infusion data), a substantial bleeding indicator, multiple organ failure (failure of two or more vital organ systems), and the units of plasma, platelets, and red blood cells infused by 24 hours (which are simultaneously a summary of the treatment of interest) [34].

Hospital name (abbreviation)	City, State
University of Texas, Houston (UHT)	Houston, Texas
Brooke Army Medical Center (BAMC)	Houston, Texas
Froedtert Memorial Hospital (FH)	Milwaukee, Wisconsin
Oregon Health and Science University Hospital (OHSUH)	Portland, Oregon
University Hospital Cincinnati (UHC)	Cincinnati, Ohio
San Francisco General Hospital (SFGH)	San Francisco, California
University of Pittsburgh Medical Center (UPMC)	Pittsburgh, Pennsylvania
University of Texas Southwestern (UTSW)	Dallas, Texas
Harborview Hospital (HH)	Seattle, Washington
University of Texas Health Center at San Antonio (UTHSCSA)	San Antonio

Table 1.1: Trauma centers that enrolled PROMMTT patients

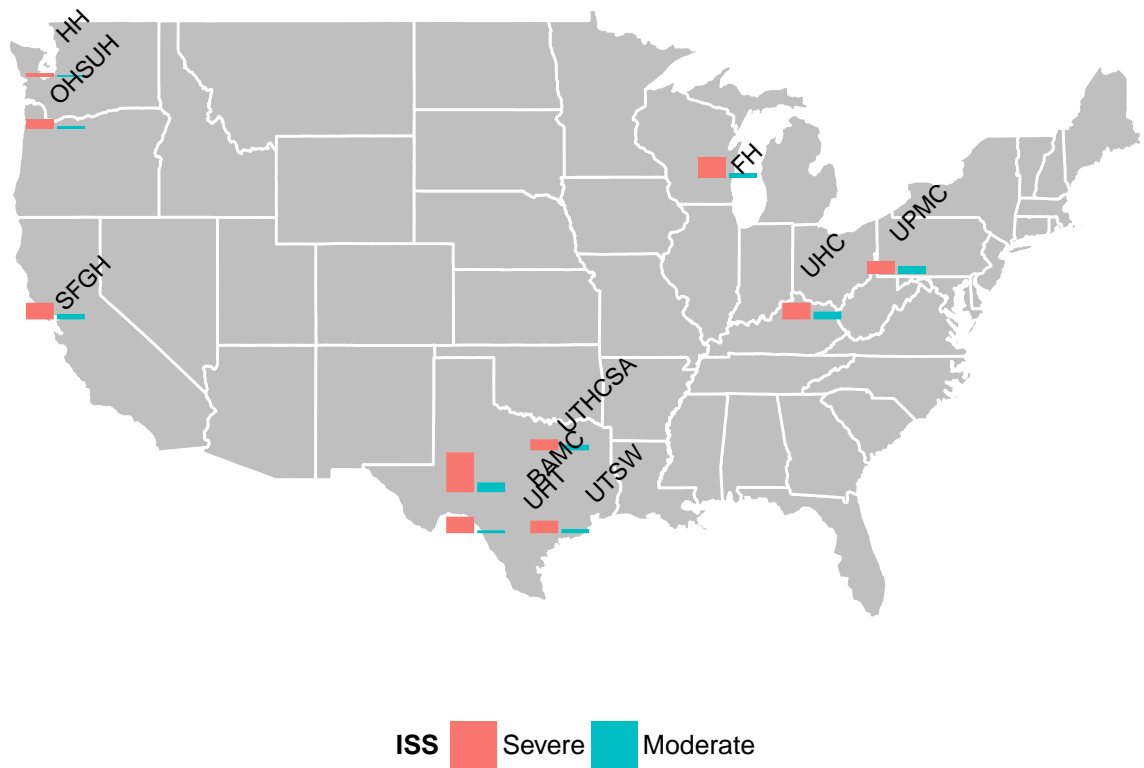


Figure 1.1: Map of PROMMTT hospitals with barplots showing the varied distributions of the injury severity of patients at each site

	Min	1st Qu	Median	Mean	3rd Qu	Max	Missing	Boxplot	Hist
ISS	0	16	25	26.2	34	75	0		
BMI	11.6	23.6	26.7	27.9	30.7	73.6	271		
Male	0	0	1	0.7	1	1	0		
Age (years)	16	24	38	40.7	54	97	1		
Hispanic	1	2	2	1.8	2	2	70		
Penetrating injury	0	0	0	0.4	1	1	0		
Anticoagulants	0	0	0	0.2	0	1	283		
Systolic BP	0	86	106	107.8	128	260	30		
Heart rate	0	86	105	106	124	199	25		
Glasgow coma score	3	3	14	9.8	15	15	104		
INR	0.8	1.1	1.2	1.5	1.4	18	162		
PTT	16	24.1	27.6	32	33	200	197		
Platelet count	1	180	226.5	231.5	277	938	68		
Hemoglobin	3	10.1	11.7	11.6	13.3	18.4	45		
Base deficit	-28.6	-10.1	-6.3	-7.1	-3	8	278		
White race	0	0	1	0.7	1	1	0		
Black race	0	0	0	0.2	0	1	0		
Asian/Pacific Islander race	0	0	0	0	0	1	0		
Unknown race	0	0	0	0	0	1	0		
FAST result	0	0	0	0.2	0	1	0		

Table 1.2: Summary statistics for covariates of interest in PROMMTT

	Min	1st Qu	Median	Mean	3rd Qu	Max	Missing	Boxplot	Hist
Massive transfusion (data)	0	0	0	0.2	0	1	0		
Massive transfusion (reported)	0	0	0	0.2	0	1	0		
Substantial bleeding	0	0	0	0.3	1	1	0		
Overall mortality	0	0	0	0.2	0	1	0		
24-hour mortality	0	0	0	0.1	0	1	0		
2-hour mortality	0	0	0	0	0	1	0		
6-hour mortality	0	0	0	0.1	0	1	0		
Complications	0	0	0	0.1	0	1	0		
Multiple organ failure	0	0	0	0	0	1	0		
Units of plasma by 24 hours	0	0	4	6.2	8	77	0		
Units of platelets by 24 hours	0	0	0	0.7	1	11	0		
Units of RBC by 24 hours	0	2	5	8.2	9	108	0		

Table 1.3: Summary statistics for clinical outcomes of interest in PROMMTT

1.2.2 The Inflammation and Host Response to Injury Cohort

In addition to utilizing clinical data to improve trauma injury treatment, understanding the critical features of response and recovery at a genomic level can help guide physicians in making treatment decisions and identify high-risk patients. One common reaction to traumatic injury is inflammation. This is a common process in the human body that guards against infection and can help heal injury, but can set off a cascade of potentially deadly events [35]. Excessive inflammation can lead to sepsis, which increases the chances

of mortality and is also related to the coagulation of blood [35]. Thus, understanding the mechanisms and pathways by which these processes act could help guide clinicians' decision making.

The study Inflammation and Host Response to Injury is a large-scale collaborative research program devoted to the systems level understanding of the key regulatory elements and their relative roles and importance that drive patients' response to serious injury and its accompanying severe systemic inflammation [12]. A subset of these data ($n = 167$), had gene expression measured in peripheral blood leukocytes at various time points during their treatment using an Affymetrix U133 microarray chip [36, 37]. The data are publicly available on the Gene Expression Omnibus (Accession GSE2328). Brownstein et al. (2006) established that certain genes are differentially expressed in the mixed leukocytes of mice that were exposed to traumatic injuries and that the patterns of the up and down regulation of these genes were different among three different animal models of inflammation and injury [38]. Previous studies using this data include [39], who aimed to create a score based on the entire probe set to predict negative outcomes in trauma patients. However, this composite score did not identify specific genes that were important nor does it respect the longitudinal nature of the collected data. [37] studied sources of variance in this dataset and found substantial changes in gene expression in response to trauma injury, suggesting that genomic information may be useful in treatment of trauma. Indeed, in a comparison of critically injured patients and healthy individuals, the stresses in circulating leukocyte transcriptomes after severe trauma and burn injury resulted in changes in over 80% of the cellular functions and pathways [40].

Motivated by the clinical interest in the inflammation and coagulation pathways, clinicians identified a subset of 24 genes of interest, summarized in Table 1.5 for which they were interested in obtaining measures of time-specific variable importance to see which genes were most important within and across time with respect to mortality and multiple organ failure. However, it has previously been shown that patients with different injury types and severity have substantially different gene expression profiles [40]. Thus, we adjusted for injury severity score (ISS), base deficit (BD), a measure of blood acidity that indicates overall patient health), and International Normalized Ratio (INR), which is a measure of clotting time. These covariates were dichotomized at clinically meaningful cutoffs: $ISS > 15$, $BD < -6$, and $INR > 1.3$, the distributions of which are summarized in Table 1.4.









	Min	1st Qu	Median	Mean	3rd Qu	Max	Missing	Boxplot	Hist
ISS < 15	0	1	1	0.9	1	1	0		
BD < -6	0	0	1	0.7	1	1	0		
INR > 1.3	0	0	1	0.6	1	1	0		
Death/MOF	0	0	0	0.4	1	1	0		

Table 1.4: Summary statistics for covariates and outcome in the gene expression data

Symbol	Name	Description
THBS1	Thrombospondin 1	mediates cell-to-cell interactions
F8	Coagulation factor VIII	procoagulant component
ANGPT1	Angiopoetin 1	important for angiogenesis and vascular development
TFPI	Tissue factor pathway inhibitor	coagulation inhibitor
MMP2	Matrix metalloproteinase 2	encodes enzyme involved in breakdown of extracellular matrix
PROS1	Protein S (alpha)	involved in inhibition of blood coagulation
THBD	Thrombomodulin	involved in inhibition of blood coagulation
SERPINE1	Plasminogen activator inhibitor 1	involved in inhibition of blood coagulation
F7	Coagulation factor VII	coagulation factor VII (serum prothrombin conversion accelerator)
ANGPT2	Angiopoetin 2	important for angiogenesis and vascular development
CPB2	Carboxypeptidase B2	involved in collagen biosynthesis
F2	Coagulation factor II	coagulation factor II (thrombin)
F2R	Coagulation factor II receptor	coagulation factor II (thrombin) receptor
F2RL3	Coagulation factor II Receptor-like 3	coagulation factor II (thrombin) receptor-like 3
F3	Coagulation factor III	coagulation factor III (thromboplastin, tissue factor)
MMP9	Matrix metalloproteinase 9	encodes enzyme involved in breakdown of extracellular matrix
NOS2	Nitric oxide synthase 2	involved in inflammation response to trauma
NOS3	Nitric oxide synthase 3	involved in inflammation response to trauma
PF4	Platelet factor 4	involved in platelet aggregation
PLAT	Tissue plasminogen activator	plasminogen activator, which is associated with excessive bleeding
PROC	Activated protein C	protein C (inactivator of coagulation factors Va and VIIIa)
PROCR	Activated protein C receptor	protein C receptor, endothelial
S1PR1	Sphingosine-1-phosphate receptor 1	involved in the regulation of lymphocytes trafficking
SERPINC1	Serpin Peptidase Inhibitor, Clade C	inhibits thrombin production, reducing clotting

Table 1.5: Gene names and descriptions for genes involved in the coagulation and inflammation pathway

1.3 Outline

The subsequent chapters are organized as follows. In Chapter Two, we introduce and motivate the use of data-adaptive machine-learning (SuperLearning) for the prediction of clinical outcomes and as the basis of estimators for estimating parameters motivated by causal inference. In Chapter Three, we examine time-specific variable importance measures applied to the genomic data. In Chapter Four, we compare the quality of care at different PROMMTT hospitals. Throughout, we highlight the utility of causal inference in motivating clinically relevant parameters of interest and the importance of careful estimation of these parameters.

Chapter 2

Semiparametric prediction of clinical outcomes using SuperLearner

2.1 Introduction to the prediction problem

Throughout our exploration of the factors driving patient response to trauma and improving patient care, we utilized a machine-learning prediction algorithm called SuperLearner. The prediction of an outcome Y_i using covariates W_i is a common problem in data analysis and many choices exist for the algorithm to use. Generally, an algorithm is an estimator that maps a data set of n observations $X_i = (W_i, Y_i)$, $i = 1, \dots, n$ into a prediction function that can be used to map W into a predicted value for Y . Algorithms may differ in the number of covariates or basis functions used, the loss function being minimized, and may depend on tuning parameters of their own so the choice of prediction algorithm is a challenge. Often, the prediction of quantitative outcomes has been called regression while the prediction of qualitative outcomes has been called classification, both of which we explored in the prediction of clinical outcomes using the PROMMTT data. Our goal was to build the best possible predictor of the clinical outcome summarized in Table 1.3. While there is great clinical interest in predicting eventual outcomes of trauma patients, these predictions are often limited to heuristic scoring systems and simple regressions in the interest of maintaining interpretability. We advocate for the use of a machine-learning prediction algorithm called SuperLearner, which has desirable theoretical optimality properties as well as good performance in the prediction of outcomes in PROMMTT.

Formally, the parameter of interest in the context of loss-based estimation is denoted by ψ_0 . This function is the minimizer of the risk (the expected loss), denoted by

$$\psi_0 = \operatorname{argmin}_{\psi' \in \Psi} \int L(x, \psi') dP_0(x) \quad (2.1)$$

where $L(x, \psi')$ represents any generic loss function, e.g. squared error (L_2) loss: $(Y - \psi(X))^2$, and P represents the true data-generating distribution. However, we do not know the true value of ψ , requiring estimation of the risk. We could use a simple substitution estimator of ψ_0 based on the empirical distribution of the data, P_n , which would yield

$$\psi_n = \operatorname{argmin}_{\psi' \in \Psi} \int L(x, \psi') dP_n(x) \quad (2.2)$$

Here, our parameter of interest was the conditional expected value of Y given W . We could consider competing prediction algorithms that model this conditional distribution and select the one that minimizes the empirical risk if we knew the true underlying distribution of the data over which the expectation is taken. However, this can lead to overfitting, where prediction algorithms have overly optimistic estimates of predictive performance because the same data were used to assess performance as were used to build the model [41, 42]. To avoid overfitting, we use cross-validation, a procedure that involves splitting the data into folds and creating learning and validation sets with the aim of evaluating a prediction algorithm's performance in data that were not used to build it. In 10-fold cross-validation, the data are partitioned into 10 subsets. A candidate algorithm is fit, or trained, on 9/10 of the data and the predictive performance is measured, or validated, on the remaining 1/10 of the data. This process is repeated 10 times, with each of the folds taking a turn being the validation set. Cross validation is used to avoid overfitting and overestimating the predictive ability of a given algorithm. The asymptotic and finite sample optimality properties of cross-validation samples has been established previously [43–46]. The cross validated risk assesss the performance of a candidate algorithm across the validation sets, allowing for a “fair” comparison of different algorithms. This suggests an procedure for the selection of a prediction algorithm based on its cross validated risk.

Many candidate prediction algorithms exist ranging from simple, such as a main-terms regression, to more complex, such as neural nets. In the critical care literature, since the focus is on interpretability of the prediction algorithm, main-terms or stepwise regressions are often used and are rarely cross-validated. The true functional form of the relationship between clinical outcomes and covariates of interest is typically unknown *a priori*. It may be that a main-terms regression will describe the true underlying data-generating

distribution, but we want to avoid relying on unnecessary model assumptions, especially if we plan to use these prediction functions as the basis for causal effect estimation. Thus, it is prudent to consider many options for candidate predictors.

An extension of the cross-validation selector allows for the combination of all the candidate algorithms into an ensemble predictor, which avoids having to choose a single candidate prediction algorithm. Some examples of ensemble methods include bootstrap aggregation of trees (bagging), random forests, and boosting [41]. The aim of ensemble learning is to build a prediction model by combining the strengths of several base models [41]. Generally, ensemble learning consists of first developing a population of base learners in training data and then combining them into a composite predictor. Bagging makes use of the bootstrap as a way to assess the accuracy of a prediction and averages predictions in the training set over a collection of bootstrap samples, resulting in a reduction of the variance around the prediction [41]. Any model can be “bagged” and the results can be averaged in various ways, for example, a “committe method” takes a simple average of the predictions for each model in a classification problem [41]. Random forests improved on bagging by reducing the correlation between sampled trees and averaging the predictions [41, 47]. Boosting is another ensemble method that combines “weak” prediction algorithms (those whose error rate is only slightly better than guessing randomly) across modified versions of the data and concatenates the sequence in a weighted combination where more accurate algorithms are given higher weights [41]. These procedures all aim to avoid model misspecification by combining a set of candidate prediction algorithms into a composite algorithm. Given the extensive choices of candidate prediction algorithms and the added layer of choice of ensemble methods for the combination of these candidates, choosing a procedure is a challenge and it is unknown *a priori* which procedure is correct. Thus, we advocate for the use of SuperLearning, which is based on the machine-learning principle of stacking, where all candidate learners, including any ensemble methods, can be considered as candidate prediction algorithms and included in the resulting ensemble predictor. The SuperLearner algorithm proceeds as follows, where \mathcal{L} represents the library of candidate prediction algorithms and $K(n)$ represents the number of candidate

prediction algorithms:

Algorithm 1: SuperLearner algorithm

for $i \in 1 \dots K(n)$ **do**

 Fit each algorithm in \mathcal{L} on the entire data set $\mathcal{W} = W_1, \dots, W_n$ to obtain
 $\hat{\Psi}_k(W), k = 1, \dots, K(n)$;

end

Split the data set \mathcal{W} into V equal sized folds **for** $v \in 1 \dots V$ **do**

 Let the v -th fold be the validation sample $V(v)$ and the remaining folds be the
 training sample $T(v)$;

 Fit each algorithm in \mathcal{L} on the training sample $T(v)$;

 Save predictions for each algorithm on the corresponding validation data

$\hat{\Psi}_{k,T(v)}(W_{V_v})$;

end

Stack the predictions from each algorithm to create a n by $K(n)$ matrix of predictions where the predictions are taken across the validation sets, denoted by

$Z = \hat{\Psi}_{k,T(v)}(W_{V(v)})$ where $v = 1, \dots, V$ and $k = 1, \dots, K(n)$;

Consider a family of weighted combinations of the candidate estimators indexed by a weight vector α ;

$$m(z|\alpha) = \sum_{k=1}^{K(n)} \alpha_k \hat{\Psi}_{k,T(v)}(W_{V(v)}) \quad \text{where } \alpha_k \geq 0 \forall k \text{ and } \sum_{k=1}^{K(n)} \alpha_k = 1 \quad (2.3)$$

Determine the α that minimizes the cross-validated risk of the candidate estimator $\sum_{k=1}^{K(n)} \hat{\Psi}_k$ over all possible combinations of α using non-negative least squares ;

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - m(z_i|\alpha))^2 \quad (2.4)$$

Combine $\hat{\alpha}$ with $\hat{\Psi}_k(W)$ to create the final SuperLearner fit

$$\hat{\Psi}_{SL}(W) = \sum_{k=1}^K \hat{\alpha}_k \hat{\Psi}_k(W) \quad (2.5)$$

SuperLearner itself can also be cross-validated in order to obtain an honest risk estimate, which allows for the comparison of SuperLearner to the library of candidate prediction algorithms. One could use this procedure to select the single best prediction algorithm

(the so-called discrete SuperLearner) by selecting the one with the smallest cross-validated risk. Indeed, it is possible for a single algorithm to get all the weight in the convex combination of algorithms because it performs consistently better than all its competitors [48]. However, SuperLearner is guaranteed to perform as well or better than the best algorithm in the supplied library [49]. Thus, we advocate using the convex combination since it allows for more flexibility over possible functional forms under the mild constraint that the number of candidate prediction algorithms, $K(n)$, less than is polynomial in size [9, 49].

2.2 Application to PROMMTT data

We were interested in predicting clinical outcomes of interest in traumatically injured PROMMTT patients using observed covariate data that are commonly collected in the emergency department. The early identification of high-risk patients remains a challenge in critical care. We utilized SuperLearner with a large library of candidate prediction algorithms to predict the clinical outcomes of interest. We present results for all the clinical outcomes and then delve into the underlying algorithms for the prediction of massive transfusion and provide a comparison of SuperLearner to a prediction score commonly used in practice to identify patients who will require a massive transfusion.

Since some of covariates had missing values, we included indicators of missingness for predictors with missing values to create a new set of basis functions. This allowed for us to predict future observations with missing values for some covariates [50]. We compared the cross-validated performance (as measured by the area under the receiver operating characteristic curve, relative mean squared-error, and R^2) of SuperLearner, main-terms regression, and stepwise regression with variable selection via AIC, which are two commonly-used models in the trauma literature and also examined the ranking of each algorithm in the cross-validation folds. The cross-validation procedure allowed us to obtain "fair" comparisons of the algorithms by assessing their performance on data that were not used to construct each prediction algorithm. For non-binary outcomes, to ensure that our predictions remained in the observed range for the outcomes of interest, we transformed the continuous outcomes (the blood product infusion variables) to be between 0 and 1, predicted them using the `family = "binomial"` argument in the SuperLearner, and then backtransformed the predictions to return them to their original measurement scales.

2.2.1 SuperLearner implementation

The SuperLearner library included a variety of algorithms ranging from simple and interpretable to more complex, including some ensemble learners. We restricted the candidates to algorithms that can take in an outcome bounded between 0 and 1 but did not require it to be binary so that the same library of candidate prediction algorithms could be applied to the binary and scaled continuous outcomes. Note that the descriptions of the algorithms below would occur in the training sample for each learner supplied to SuperLearner.

Logistic regression

Logistic regression models the log-odds of an outcome as a linear combination of the predictors

$$\log \frac{\Pr(Y = 1|W = w)}{\Pr(Y = 0|W = w)} = \beta_0 + \beta 1^T(w) \quad (2.6)$$

which is usually fit by maximum likelihood using the conditional likelihood of Y given W . Starting with $\beta = 0$, the parameters are estimated with iterative least squares. The main-terms logistic regression included only the predictors and their indicators of missingness without any additional exponentiated or interaction terms. We also included a model with every possible two-way interaction between the predictors and two modifications of the main-terms regression where the full model was reduced or an intercept-only model was constructed using Akaike's Information Criterion (AIC), which is calculated as

$$AIC = 2k - 2\ln(L) \quad (2.7)$$

where k is the number of parameters in the statistical model and L is the maximized value of likelihood model. In backwards selection via AIC, a full main-terms regression model has terms removed based on those whose removal results in a decrease in the AIC. This procedure stops when the removal of a variable does not result in a decrease of the AIC. The forward selection procedure starts with an intercept-only model that adds terms based on how they affect the AIC.

Bayesian logistic regression

The Bayesian implementation of logistic regression involves placing independent weakly informative prior distributions on the coefficients [51]. One advantage of this approach

is that it will always give an answer, even in the case of complete separation in logistic regression, that is, when the outcomes values are perfectly determined by a predictor [51].

Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is an adaptive machine learning algorithm based on linear splines and their tensor products. They use so-called “reflected pairs” of piecewise linear basis functions for each predictor W_j that are based on the values of the observed data. These functions are then used in a forward stepwise regression procedure. At each step, functions of the pairs of piecewise linear functions are added based on how their addition affects the training error [41]. This results in a model that may overfit the data, so terms whose removal leads to the smallest increase in residual squared error are taken out and a penalty is added for model complexity [41, 52]. Using these piecewise linear functions ”allows a parsimonious prediction using locally non-zero components” [41]. In our implementation, we restricted the model to pairwise tensor products of basis functions and imposed a penalty of 3 for model complexity. This model was included as a candidate prediction algorithm because it allowed for more flexibility than simple regressions but still aimed to find the most parsimonious model.

Classification and regression trees

This procedure builds a decision tree by partitioning the data based on the top predictors and recursively splitting the data using all the predictors [41, 53]. The resulting tree can then be “pruned” by considering each pair of leaves (nodes) with a common parent and whether their removal would decrease the prediction error [53]. Class predictions are based on a majority vote across the different trees and the average is used for prediction of continuous outcomes [53]. Since this algorithm searches for binary cutoffs for the predictors, which is commonly implemented in critical care, it was included as a candidate algorithm in SuperLearner.

Random forest

Random forests are an ensemble prediction method based on averaging a forest of bagged decision trees [41, 54]. This resampling procedure draws repeated bootstrap sample from the data, and within each sample selects a random number of predictors, and builds a classification or regression tree based on the best predictor among the selected predictors [53]. Different splits and nodes are selected in each bootstrap sample based on the

variability in the resampled data as well as the number of predictors available to split on in each bootstrap sample resulting in a forest of decision trees. The tuning parameters for random forest include the number of trees, the number of predictors chosen at random in each bootstrap sample (m), and the minimum node size. We chose to grow 1000 trees, sample 10 of the predictors for each bootstrap tree, and split until the node size was 5. Random forest is a relatively nonparametric machine learning tool, so it provides flexibility in fitting the regression function.

Neural nets

Neural networks were originally developed as models for the human brain [41, 55]. They use linear combinations of the predictors to model the outcome rather than the predictors themselves. These linear combinations, which can be denoted Z_m , are hidden units (not directly observable) but serve as a bridge between the observed covariates W and the outcome Y 2.1. At its most basic, a neural network simplifies to a regression of the outcome on linear combinations of the predictors, but additional hidden layers and transformation functions allow for increased model flexibility, which is why it was included as a candidate in SuperLearner.

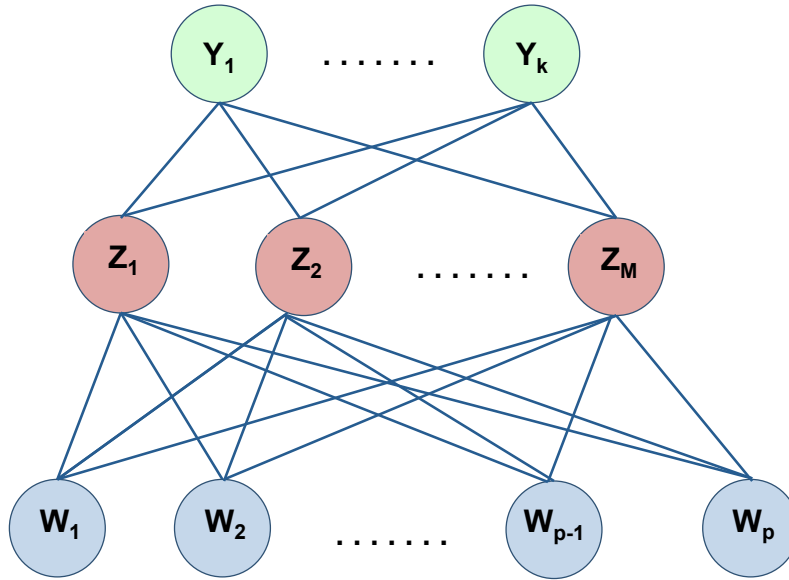


Figure 2.1: A schematic of a neural network where the W 's represent the predictors, linear combinations of which make up the hidden Z 's, which are in turn used to model Y

Leekasso

This algorithm fits a simple regression model for each of the predictors of the form

$$E[Y|W] = \beta_0 + \beta_k X_k \quad \text{for } k = 1, \dots, K \quad (2.8)$$

where K is the total number of predictors. The top ten variables with the smallest p-values from testing the β_k coefficients, then fit a linear model with only those ten variables. While this approach is *ad hoc* and there is little theoretical justification as to why the top ten predictors would yield better results than the top 9 or 11, it is one practical approach to dealing with the curse of dimensionality [56, 57].

2.2.2 Performance assessment

We assessed the performance of SuperLearner and its candidate algorithms using the cross-validated area under the receiver operating characteristic (AUROC) curve for the binary outcomes. The receiver operating characteristic curve is a plot of the true positive rate versus the false positive rate at various threshold settings for each classifier, resulting in a curve for which the area below is a one-number summary of the predictive performance of a classifier [58]. A classifier that chooses at random will have an AUROC of 0.5 and the goal is to have an AUROC of 1 (perfect prediction of both classes) (see Figure 2.2). Our performance measure here was the cross validated AUROC, which we describe below.

Recall that the main idea of cross-validation is to divide the available data into a training set and a validation set. The observations in the training set are used to build the prediction algorithms and the validation set is used to assess the risk of these algorithms. To distinguish these sets, we index the distributions of the training and validation sets with a split vector $B_n = B_n(i) : i = 1, \dots, n$

$$B_n(i) = \begin{cases} 0 & \text{if } i\text{th observations } X_i \text{ is in the training set} \\ 1 & \text{if } i\text{th observations } X_i \text{ is in the validation set} \end{cases} \quad (2.9)$$

Then P_{n, B_n}^0 and P_{n, B_n}^1 denote the empirical distributions of the training and validation sets. Then, as in Dudoit and van der Laan (2003), the general definition of the cross

validated risk estimator is

$$\begin{aligned}
 \hat{\theta}_{p_n, n} &\equiv E_{B_n} \Theta(\hat{\Psi}(P_{n, B_n}^0), P_{n, B_n}^1) \\
 &= E_{B_n} \int L(x, \hat{\Psi}(P_{n, B_n}^0)) dP_{n, B_n}^1(x) \\
 &= E_{B_n} \frac{1}{n_1} \sum_{i: B_n(i)=1} L(X_i, P_{n, B_n}^0)
 \end{aligned} \tag{2.10}$$

where $n_1 = \sum_i B_n(i)$ and $\hat{\Psi}(P_{n, B_n}^0)$ denotes the estimator of ψ_0 based on the training set. Dudoit and van der Laan (2003) proved the asymptotic linearity of the cross validated risk estimator.

Ledell, van der Laan, and Petersen (2012) showed that the AUC can be used as a loss function and also established the asymptotic linearity of the cross-validated AUC as an estimator. The AUC can be defined as

$$AUC(P_0, \psi) = \int_0^1 P_0(\psi(W) > c | Y = 1) P_0(\psi(W) = c | Y = 0) dc \tag{2.11}$$

The target, the true cross-validated AUC, is the mean of the AUC across the validation folds

$$EB_n AUC(P_0, \psi_{B_n}) = \frac{1}{V} \sum_{v=1}^V AUC(P_0, \psi_{B_n}^v) \tag{2.12}$$

where B_n^v is the split vector for fold v in the cross-validation procedure.

As shown in Ledell, van der Laan, and Petersen (2012) the cross validated AUROC is an asymptotically linear estimator of the true cross validated AUROC. This allows for the computation of confidence intervals for the point estimate of the AUROC and demonstrated that the cross validated AUC is indeed an asymptotically linear estimator of the true AUC [59]. The influence curve for the AUROC is given by

$$\begin{aligned}
 IC_{AUROC}(P_0, \psi)(O) &= \frac{I(Y = 1)}{P_0(Y = 1)} P_0(\psi(W) < x | Y = 0)|_{x=\psi(W)} \\
 &+ \frac{I(Y = 0)}{P_0(Y = 0)} P_0(\psi(W) > x | Y = 1)|_{x=\psi(W)} \\
 &- \left\{ \frac{I(Y = 0)}{P_0(Y = 0)} + \frac{I(Y = 1)}{P_0(Y = 1)} \right\} AUROC(P_0, \psi)
 \end{aligned} \tag{2.13}$$

For performance assessment of the prediction of continuous outcomes, we calculated the cross validated relative mean squared-error (relative to main-terms regression)

$$relMSE(k) = \frac{MSE(k)}{MSE(lm)}, k = 1, \dots, K \tag{2.14}$$

and percent correctly classified where the bins were created based on the range of each outcome.

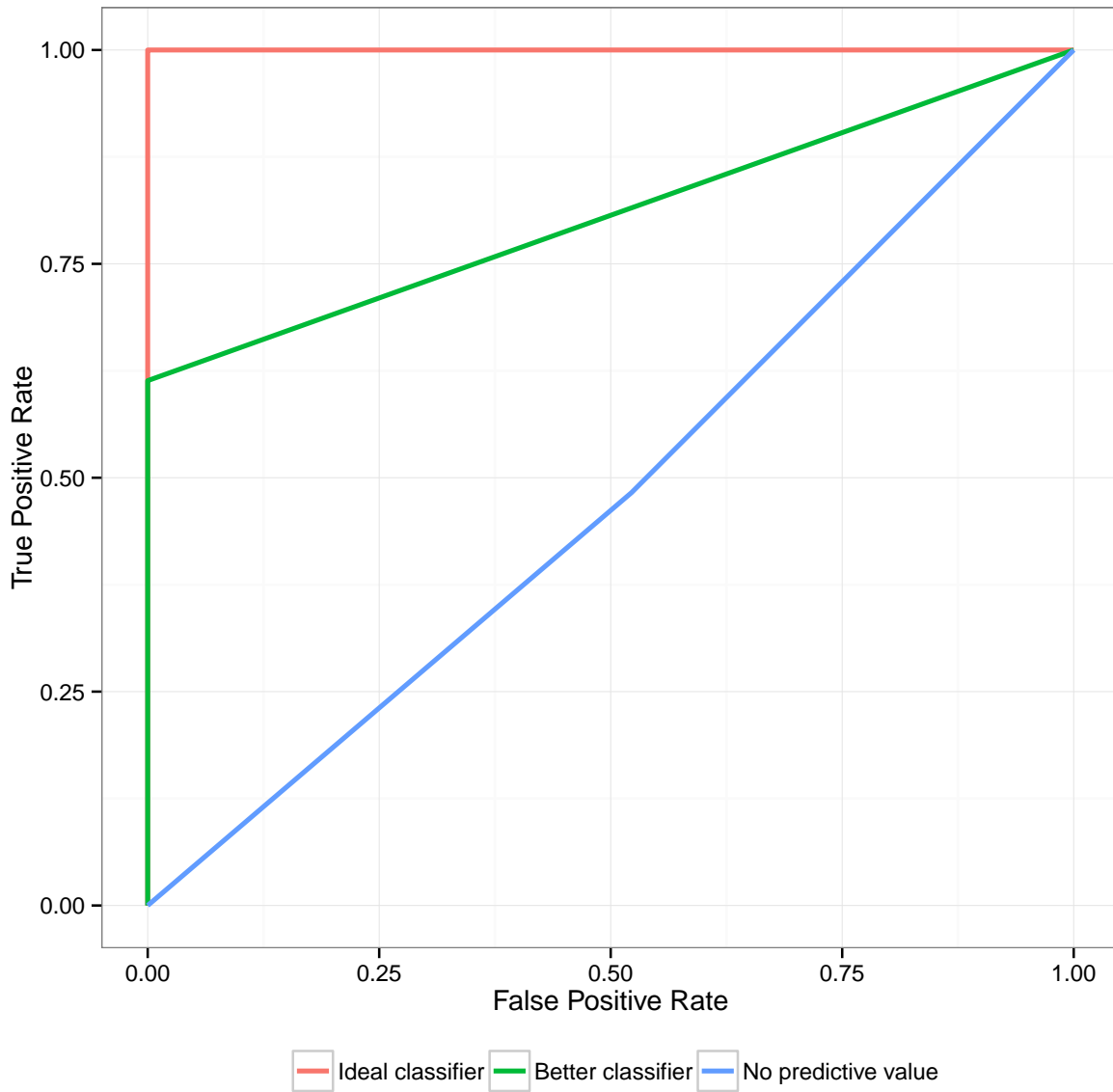


Figure 2.2: Receiver operating characteristic curves for ideal (red), better (green), and random (blue) classifiers

2.3 Results

The cross-validated ROC curves for SuperLearner, stepwise selection via AIC, and main-terms regression are shown in Figure 2.3. For the rarer outcomes, MOF and complications, the predictions were not much better than a random classifier but the other outcomes were predicted well by all three algorithms. Overall, using SuperLearner did not result in a substantial improvement in the predictive ability for any outcome, a result that can be confirmed in Figure 2.4. This figure shows the cross-validated AUROC values for SuperLearner and all its candidate algorithms as well as 95% confidence intervals computed using the influence curve. With the exception of multiple organ failure, SuperLearner (shown on the far left in every facet of the plot) performs significantly better than a random classifier. For every outcome, neural nets and regression with two-way interactions perform the worst, suggesting that the underlying functional form relationship was not captured by either of these algorithms.

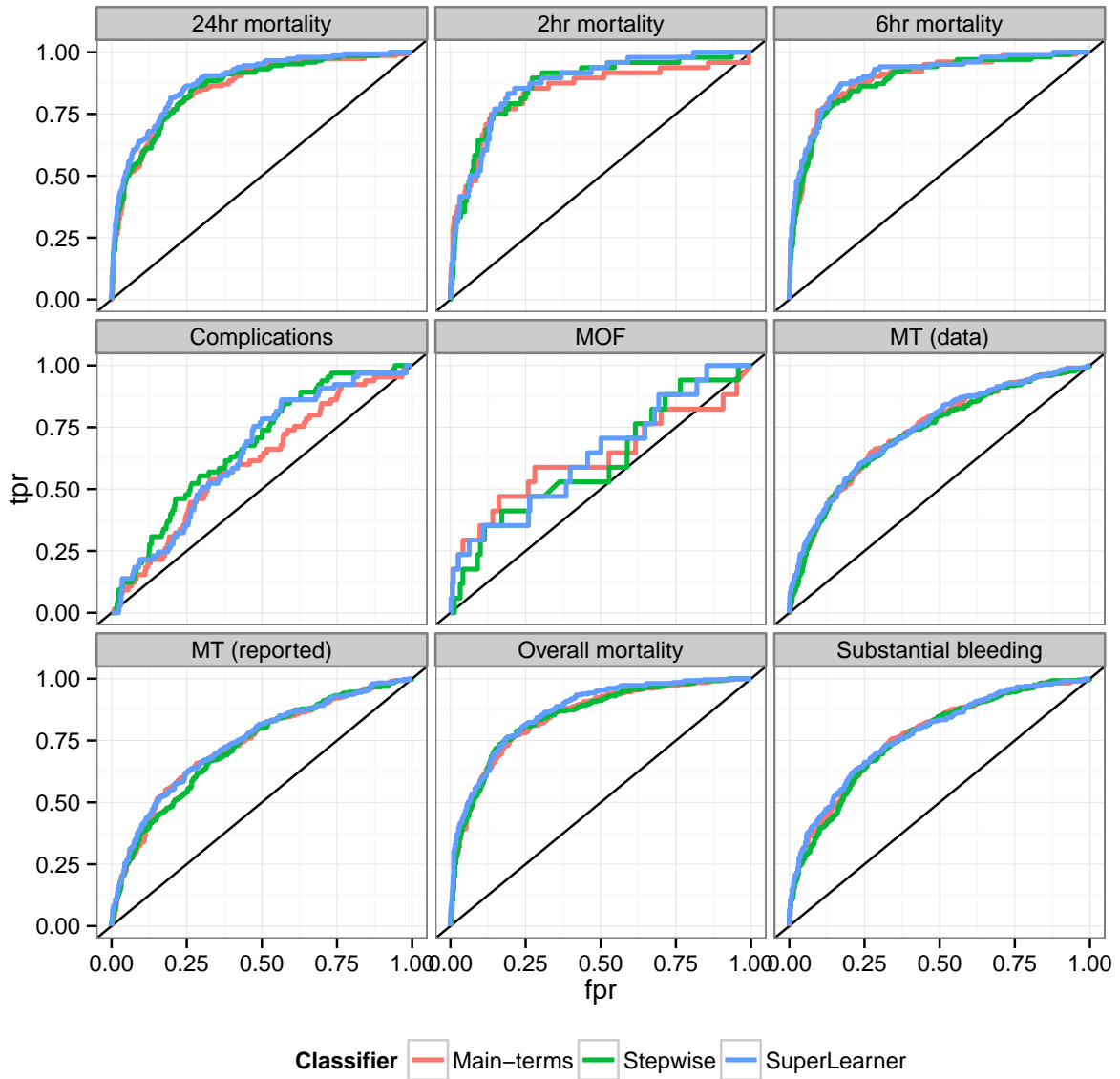


Figure 2.3: Comparison of the cross-validated receiver operating characteristic curves for SuperLearner, stepwise regression with variable selection using AIC, and main-terms regression.

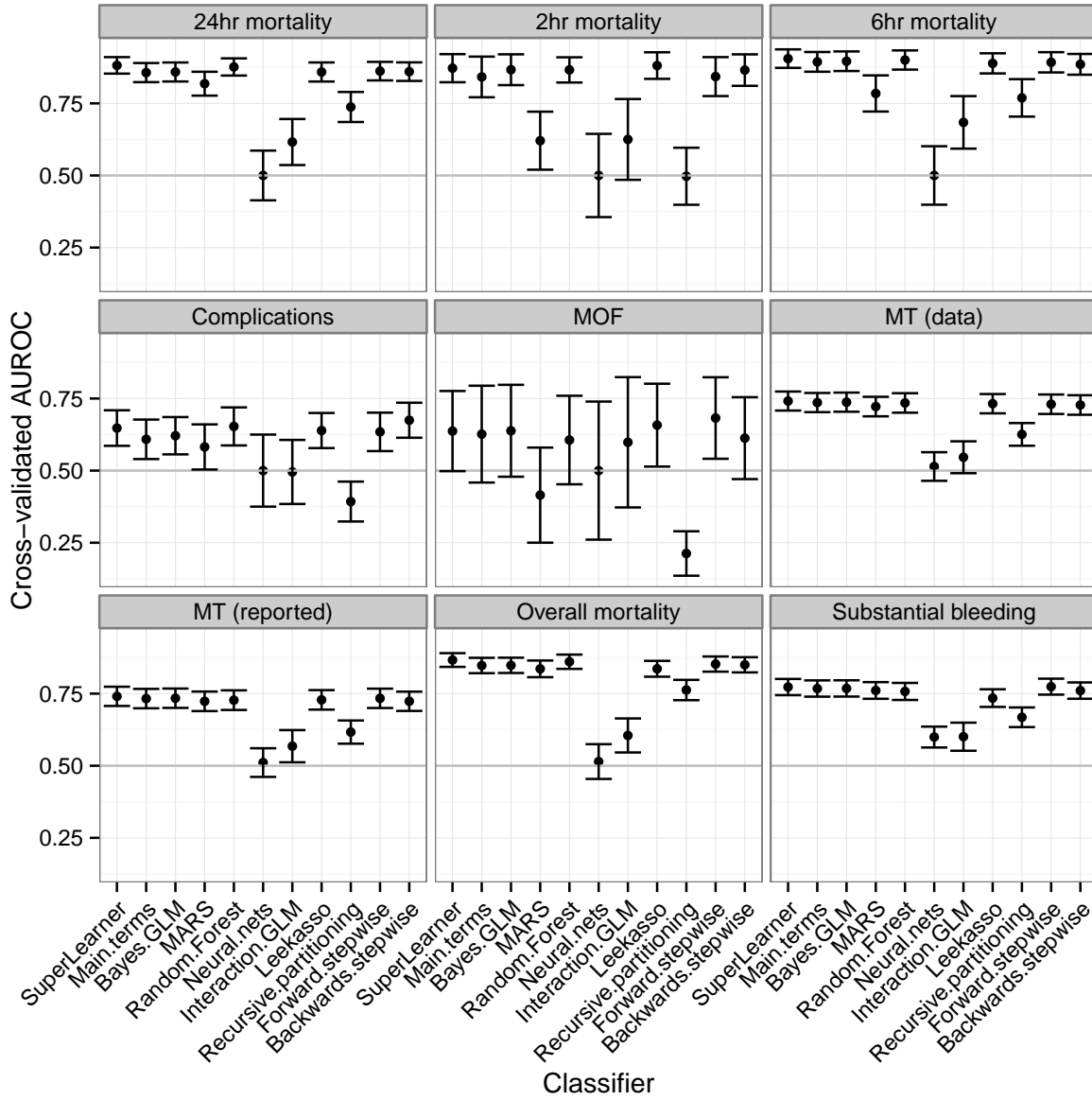


Figure 2.4: Cross-validated AUROC values and 95% confidence intervals for each of the outcomes and predictors

CHAPTER 2. SEMIPARAMETRIC PREDICTION

The continuous outcomes had performance measured by their cross validated relative mean squared-error and percent correctly classified, which are summarized in Table 2.1 and 2.2. Overall, SuperLearner had the lowest relative MSE for all three of the outcomes, performing marginally better than main-terms regression and stepwise regression with selection via AIC. SuperLearner did not perform appreciably better than any of its candidate learners in correctly classifying the blood products, although it was able to identify some of the more extreme categories unlike many of its competitors. The percent of plasma and platelets that were correctly classified were both substantially better than for red blood cells, where none of the algorithms did particularly well. This highlights the fact that SuperLearner is only as good as the supplied library.

	24-hour Plasma	24-hour Platelet	24-hour RBC
SuperLearner	0.91	0.98	0.98
Discrete SuperLearner	0.96	1.01	1.00
Main-terms	1.00	1.00	1.00
Bayes-GLM	0.99	0.99	1.00
MARS	1.20	1.22	1.24
Random-Forest	1.36	1.17	2.43
Neural-nets	1.64	1.32	1.99
Interaction-GLM	15.46	11.12	4.51
Leekasso	0.96	0.97	1.04
Recursive-partitioning	1.32	1.16	1.38
Forward-stepwise	1.02	1.03	1.09
Backwards-stepwise	1.07	1.03	1.17

Table 2.1: Relative mean squared-error for the prediction of continuous outcomes

CHAPTER 2. SEMIPARAMETRIC PREDICTION

	[0,20]	(20,40]	(40,60]	(60,100]
SuperLearner	0.99	0.04	0.00	0.00
Main.terms	0.98	0.07	0.05	0.00
Bayes.GLM	0.98	0.05	0.05	0.00
MARS	0.97	0.14	0.15	0.00
Random.Forest	1.00	0.00	0.00	0.00
Neural.nets	1.00	0.00	0.00	0.00
Interaction.GLM	0.79	0.00	0.00	0.50
Leekasso	0.97	0.02	0.00	0.00
Recursive.partitioning	1.00	0.00	0.00	0.00
Forward.stepwise	1.00	0.00	0.00	0.00
Backwards.stepwise	1.00	0.00	0.00	0.00

(a) Plasma

	[0,3]	(3,6]	(6,9]	(9,11]
SuperLearner	1.00	0.03	0.00	0.00
Main.terms	0.99	0.03	0.00	0.00
Bayes.GLM	0.99	0.03	0.00	0.00
MARS	0.98	0.08	0.00	0.00
Random.Forest	1.00	0.00	0.00	0.00
Neural.nets	1.00	0.00	0.00	0.00
Interaction.GLM	0.84	0.00	0.00	0.00
Leekasso	0.97	0.00	0.00	0.00
Recursive.partitioning	1.00	0.00	0.00	0.00
Forward.stepwise	0.99	0.03	0.00	0.00
Backwards.stepwise	1.00	0.00	0.00	0.00

(b) Platelets

	[0,2]	(2,5]	(5,9]	(9,108]
SuperLearner	0.19	0.55	0.54	0.01
Main.terms	0.27	0.51	0.48	0.02
Bayes.GLM	0.26	0.52	0.50	0.01
MARS	0.28	0.45	0.59	0.01
Random.Forest	0.28	0.14	0.28	0.02
Neural.nets	0.00	0.00	0.00	1.00
Interaction.GLM	0.31	0.41	0.29	0.17
Leekasso	0.18	0.63	0.45	0.07
Recursive.partitioning	0.06	0.50	0.40	0.01
Forward.stepwise	0.13	0.49	0.59	0.00
Backwards.stepwise	0.06	0.47	0.64	0.00

(c) Red blood cells

Table 2.2: Percent correctly classified for blood product outcomes

Regardless of the prediction algorithm used, the clinical outcomes in PROMMTT were predicted relatively well. None of the SuperLearner confidence intervals for the cross-validated area under the ROC curve crossed 0.5, suggesting that the at least the covariate information was related to the clinical outcomes. Since clinicians make treatment decisions based on these covariates, the strong predictive accuracy of the prediction models motivated additional analyses that identified the variables driving these predictions and assessed treatment efficacy while adjusting for these covariates. While simpler learners such as stepwise and main-terms regressions also performed well, we did not know *a priori* how those particular algorithms would perform and had no principled approach to comparing them that would generate “fair” comparisons. Additionally, SuperLearner had a comparable performance to each of these algorithms, there is little risk to using SuperLearner with a large library of candidate learners. Thus, we continue to advocate for the use of the cross validated SuperLearner with a rich library of algorithms to avoid having to choose a single algorithm.

Machine learning algorithms are often viewed as black boxes that that are not interpretable. One advantage of using SuperLearner is that the coefficients of each algorithm in the convex combination are a measure of that algorithm’s predictive performance and the components of each algorithm can be unpacked and examined further. For example, consider the reported massive transfusion outcome. Clinicians are very interested in identifying patients who would benefit from the infusion of blood products using patient data collected immediately after arrival in the emergency department and are very interested in the variables that might be driving the need for massive transfusion. When we examined the coefficients of each algorithm across the 20 cross validation folds, we found that MARS and Random Forest were selected the most often and given higher weights, suggesting that the relationship may be more complex than the simple regression is capturing. However, we did see good AUROC values for algorithms that were not heavily weighted in the SuperLearner, which suggests that if we built SuperLearner with the aim of maximizing the AUROC rather than minimizing the cross-validated risk, we would likely end up with different weights in the SuperLearner.

CHAPTER 2. SEMIPARAMETRIC PREDICTION

Main-terms	Bayes GLM	MARS	Random Forest	Neural nets	Itn. GLM	Leekasso	Recursive partitioning	Forward stepwise	Backwards stepwise
0.00	0.00	0.27	0.31	0.03	0.00	0.04	0.01	0.34	0.00
0.27	0.00	0.07	0.43	0.00	0.00	0.09	0.00	0.00	0.14
0.00	0.04	0.27	0.30	0.03	0.04	0.13	0.00	0.15	0.05
0.00	0.01	0.29	0.31	0.00	0.04	0.00	0.00	0.35	0.00
0.29	0.00	0.20	0.38	0.00	0.00	0.07	0.00	0.00	0.07
0.31	0.00	0.10	0.41	0.00	0.00	0.17	0.00	0.00	0.00
0.00	0.19	0.27	0.33	0.00	0.02	0.20	0.00	0.00	0.00
0.00	0.17	0.33	0.30	0.00	0.00	0.12	0.00	0.08	0.00
0.00	0.21	0.07	0.52	0.00	0.03	0.09	0.04	0.05	0.00
0.00	0.26	0.00	0.50	0.00	0.01	0.03	0.04	0.17	0.00
0.28	0.00	0.18	0.21	0.00	0.00	0.12	0.03	0.00	0.18
0.00	0.15	0.31	0.24	0.00	0.00	0.26	0.03	0.00	0.00
0.00	0.15	0.20	0.30	0.09	0.00	0.00	0.02	0.16	0.08
0.00	0.00	0.21	0.31	0.17	0.00	0.00	0.01	0.23	0.06
0.29	0.00	0.30	0.22	0.00	0.00	0.08	0.00	0.02	0.09
0.15	0.00	0.06	0.44	0.00	0.00	0.20	0.00	0.15	0.00
0.00	0.26	0.11	0.46	0.08	0.00	0.08	0.00	0.00	0.00
0.00	0.00	0.24	0.27	0.10	0.02	0.18	0.00	0.19	0.00
0.00	0.32	0.12	0.46	0.00	0.00	0.03	0.01	0.06	0.00
0.00	0.02	0.19	0.37	0.00	0.01	0.14	0.00	0.00	0.26

Table 2.3: Coefficients for each candidate prediction algorithm in SuperLearner across the 20 cross-validation folds

The cross-validated SuperLearner chooses the single best algorithm in each of the 20 folds (the so-called discrete SuperLearner), and in this case selected Bayesian regression once, random forest 17 times, and forward stepwise regression twice, suggesting that a single model may not be able to fully describe these data, although random forest gets close. We further explored the models with non-zero weights in SuperLearner in at least 10 of the 20 cross-validation folds that also had built-in variable selection methods: Random forest, forward stepwise regression, leekasso, and MARS. In the folds where they were given non-zero SuperLearner weights, we examined either the variables identified as important or the importance scores generated by the algorithm, which are described below and documented in subsequent tables and plots.

- Random forest.** The importance scores for each predictor are generated by calculating the mean decrease in accuracy (the proportion of true calls) over all the outcome classes that would occur if this variable were removed from each of the bootstrap samples. The variables can be ranked by these importance scores in each fold where random forest had a non-zero weight in SuperLearner (random forest had a non-zero weight in all 20 folds) and, to summarize, we took a weighted average of the importance scores across all the folds where the weights are the weights from the convex combination in SuperLearner to give a sense of the relative importance for all the variables. Plotting the ordered average importances showed a clear dropoff after BMI, identifying ten most important predictors of massive transfusion (see Figure 2.5).

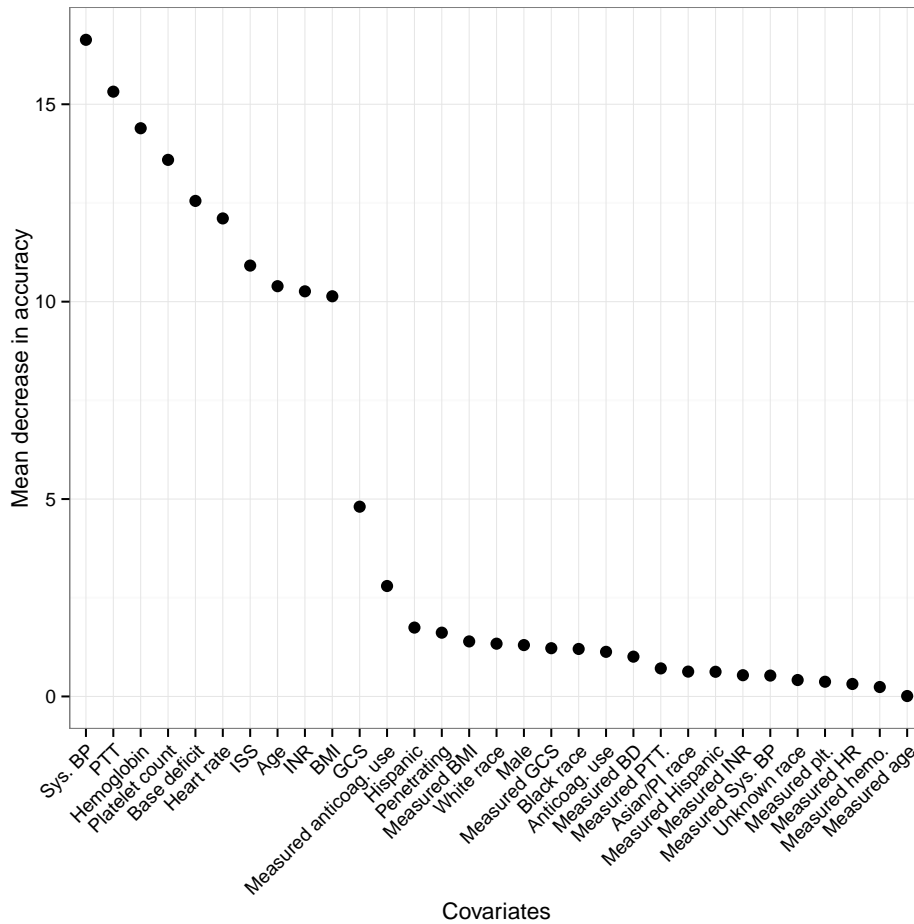


Figure 2.5: Plot of average importance scores across 20 cross validation folds for all predictors of massive transfusion

- Forward stepwise regression.** This procedure starts with an empty model and adds predictors based on how they change Akaike’s Information Criterion (AIC). Different predictors were added in each fold (forward stepwise regression had a non-zero SuperLearner weight in 12 of the 20 folds) and we ranked them based on how often they appeared, which are summarized in Figure 2.6. The top ten predictors were chosen in every fold, suggesting that they do indeed have a consistent association with the probability of massive transfusion.

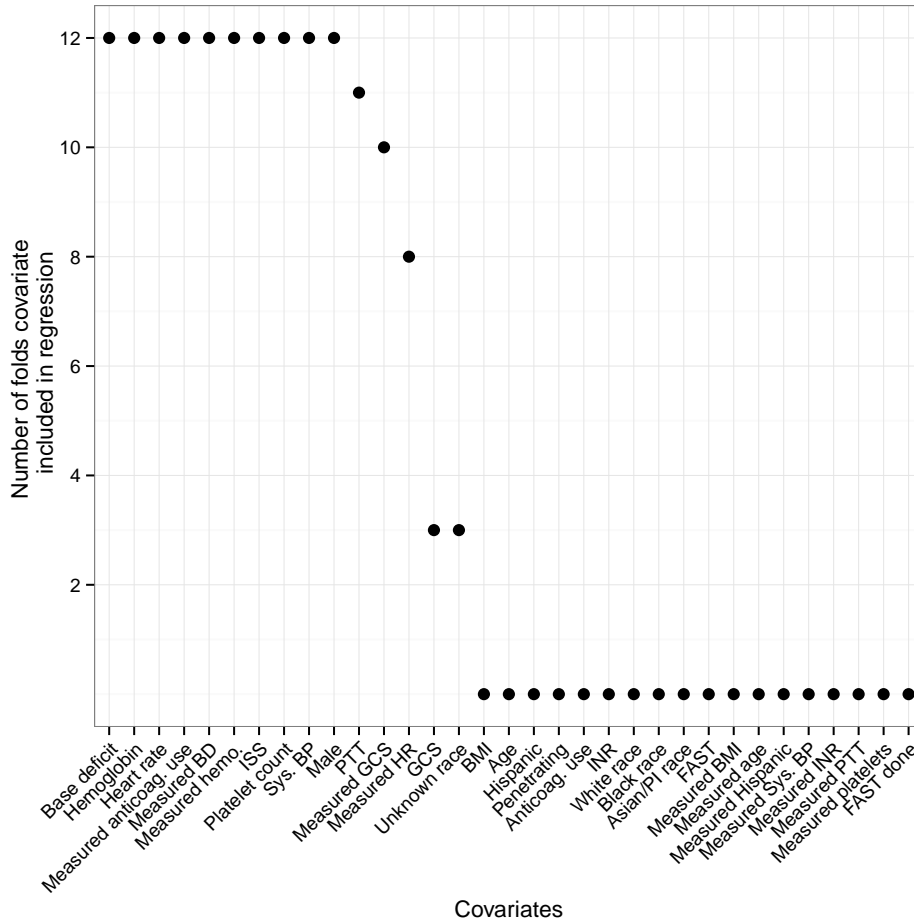


Figure 2.6: Plot of the number of folds each covariate was included as predictor by forward stepwise selection procedure

- Leekasso.** This prediction algorithm performs simple regression to rank the predictors individually by p -values and then includes the top ten in a multivariable regression model. We examined the ten predictors selected across the folds where leekasso had a non-zero weight in SuperLearner, of which there were 17, and plotted the number of folds that each covariate was included in the final multivariable regression, the results of which are summarized in 2.7. Nine variables were included in all 17 of the folds where leekasso had a positive weight in SuperLearner, suggesting that these have a strong association with massive transfusion.

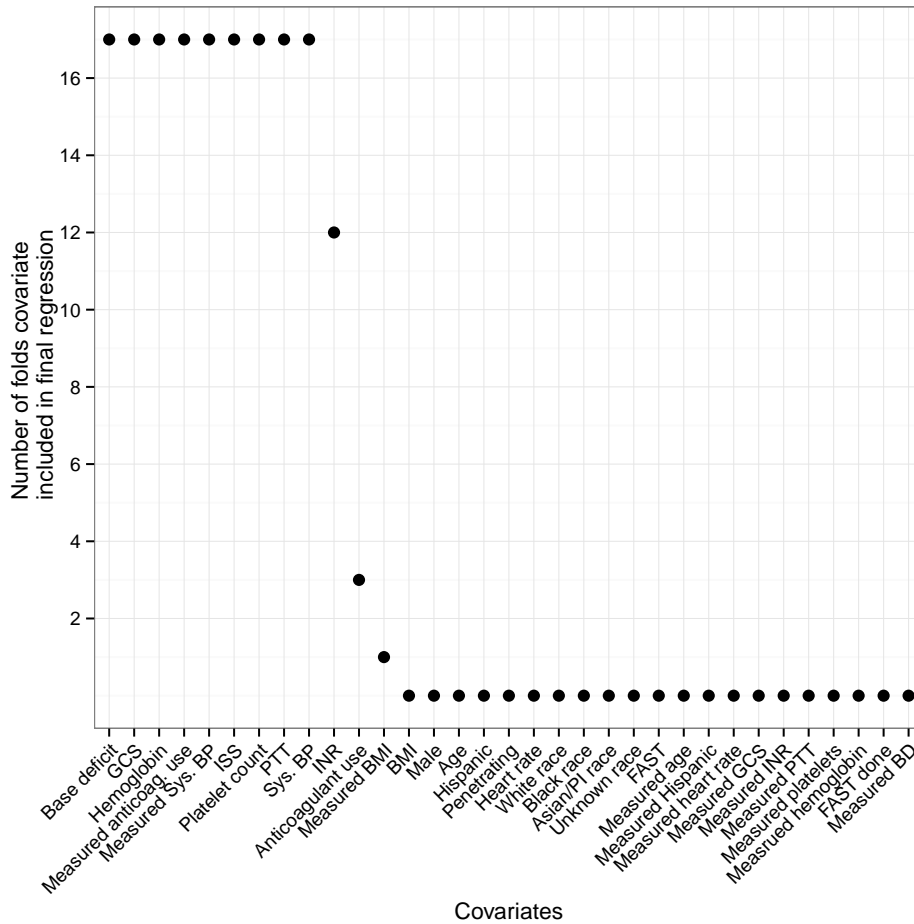


Figure 2.7: Plot of the number of folds each covariate was included as predictor by leekasso

- MARS.** Multivariate adaptive regression splines also utilize a stepwise procedure to build a regression model, but rather than using the raw predictors themselves, inputs stepwise constant functions of the predictors allowing for piecewise modeling of the outcome. Again, we counted the number of folds that each predictor appeared (out of the 19 folds where MARS had a non-zero weight in SuperLearner) and plotted the results to identify top predictors (see Figure 2.8. There is not a clear cutoff for identifying top predictors based on the number of folds each covariate appears as a predictor in a MARS model. However, seven of the predictors appeared in all 19 of the folds where MARS was used in SuperLearner.

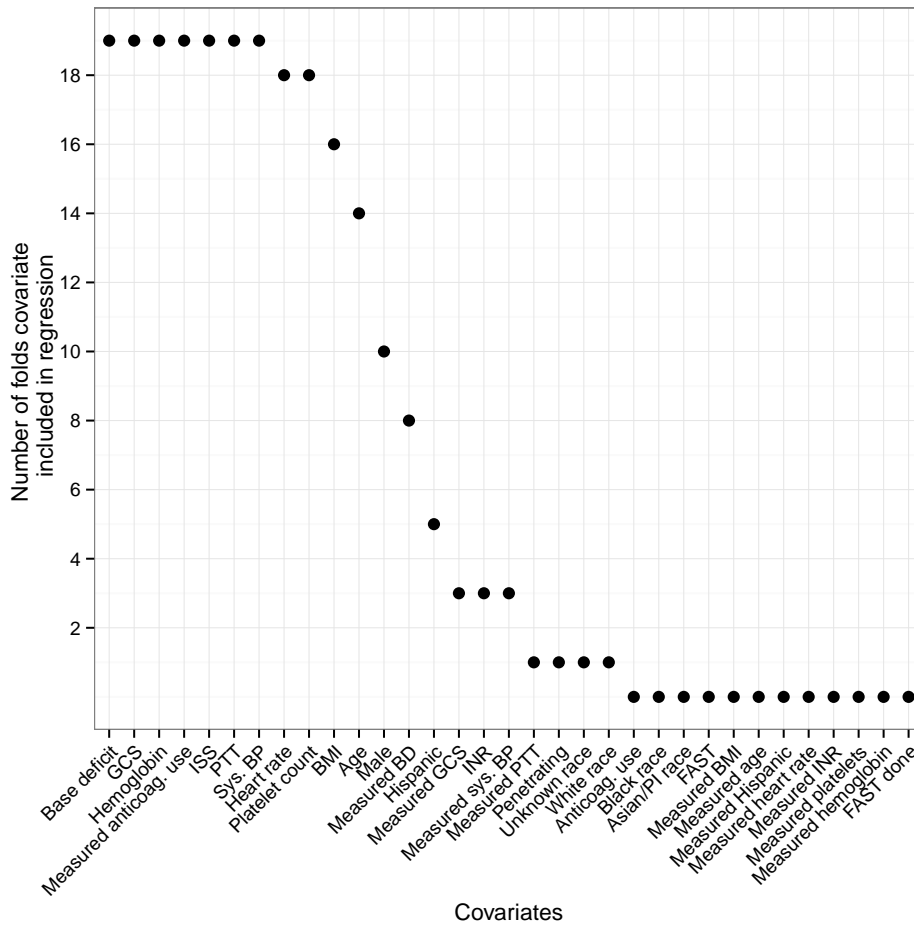


Figure 2.8: Plot of the number of folds each covariate was included as predictor by MARS

Table 2.4 summarizes the results of the exploration of predictors of massive transfusion for commonly-selected candidate classifiers in the cross validated SuperLearner that had built-in variable selection or importance measurements. For each classifier, the top predictors

were identified by plots of the ranked importance measures or how often the predictors were selected across the cross validation folds. Variables were identified as top predictors in these plots as long as they were selected before substantial drop off in the number of times they were selected (Figure 2.5 has a clear dropoff while the others are less dramatic). Several predictors are identified by every method: systolic blood pressure, prothrombin time (PTT), hemoglobin, base deficit, injury severity score (ISS) and platelet count. The variables with the "Measured" prefix are indicators of whether the variable following was measured, and several of them are identified by some of the classifiers as important predictors, suggesting that there may be some importance of missingness in the prediction of massive transfusion.

Predictor	Random forest	Forward stepwise	Leekasso	MARS
Systolic BP	✓	✓	✓	✓
Injury severity score (ISS)	✓	✓	✓	✓
Hemoglobin	✓	✓	✓	✓
Base deficit	✓	✓	✓	✓
International normalized ratio (INR)	✓			
Prothrombin time (PTT)	✓	✓	✓	✓
Platlet count	✓	✓	✓	✓
Heart rate	✓	✓		✓
Age	✓			✓
BMI	✓			✓
Male		✓		
Glasgow coma score (GCS)			✓	✓
Measured anticoagulant use		✓	✓	✓
Measured hemoglobin		✓		
Measured base deficit		✓		
Measured GCS		✓		
Measured systolic BP			✓	

Table 2.4: Top predictors identified by a subset of candidate learners in SuperLearner

2.3.1 Predicting the need for massive transfusion

The Assessment of Blood Consumption (ABC) score is designed to predict the need for massive transfusion and standardize the initiation of massive transfusion protocols across hospitals [7]. This score was based on clinician interviews regarding their clinical criteria for activation of massive transfusion, and consists of four dichotomous components: whether the injury is of a penetrating (as opposed to blunt) nature, whether the patient's systolic blood pressure was 90 mmHg or higher in the emergency department, whether

their heart rate was 120 bpm or greater, and whether they had a positive Focused Assessment with Sonography for Trauma (FAST) scan. Based on these dichotomous variables, the ABC score ranges from 0 to 4 and a score of 2 is used clinically to identify patients who will require a massive transfusion.

As a follow-up to the identification of top predictors using candidate algorithms in SuperLearner, we were interested in comparing the prediction of massive transfusion using only the ABC score to predictions obtained using the variables involved in the calculation of the ABC score in a SuperLearner, and also using all predictors, in order to see whether the ABC score is a useful scoring system. We built a SuperLearner with the four variables that make up the ABC score using the same prediction library as above and compared the area under the ROC curve for the the three prediction methods.

The ABC score dichotomized at two did not perform much better than a random classifier, with an AUROC of 0.532. The restricted SuperLearner built using only the predictors that go into the ABC score performed better with an AUROC of 0.64 but the full SuperLearner was the best at classifying massive transfusion patients with an AUROC of 0.718. This ordering of the predictive performances suggests that these variables may indeed be strong predictors of the need for future massive transfusion but the relationship is not as simple as seeing whether two of the four criteria are met, and that including other variables increases the ability to predict massive transfusion. While the confidence interval for the full SuperLearner did contain the point estimate of the AUROC for the restricted SuperLearner, the difference may be clinically meaningful since major hemorrhage is still a major cause of preventable death in trauma patients.

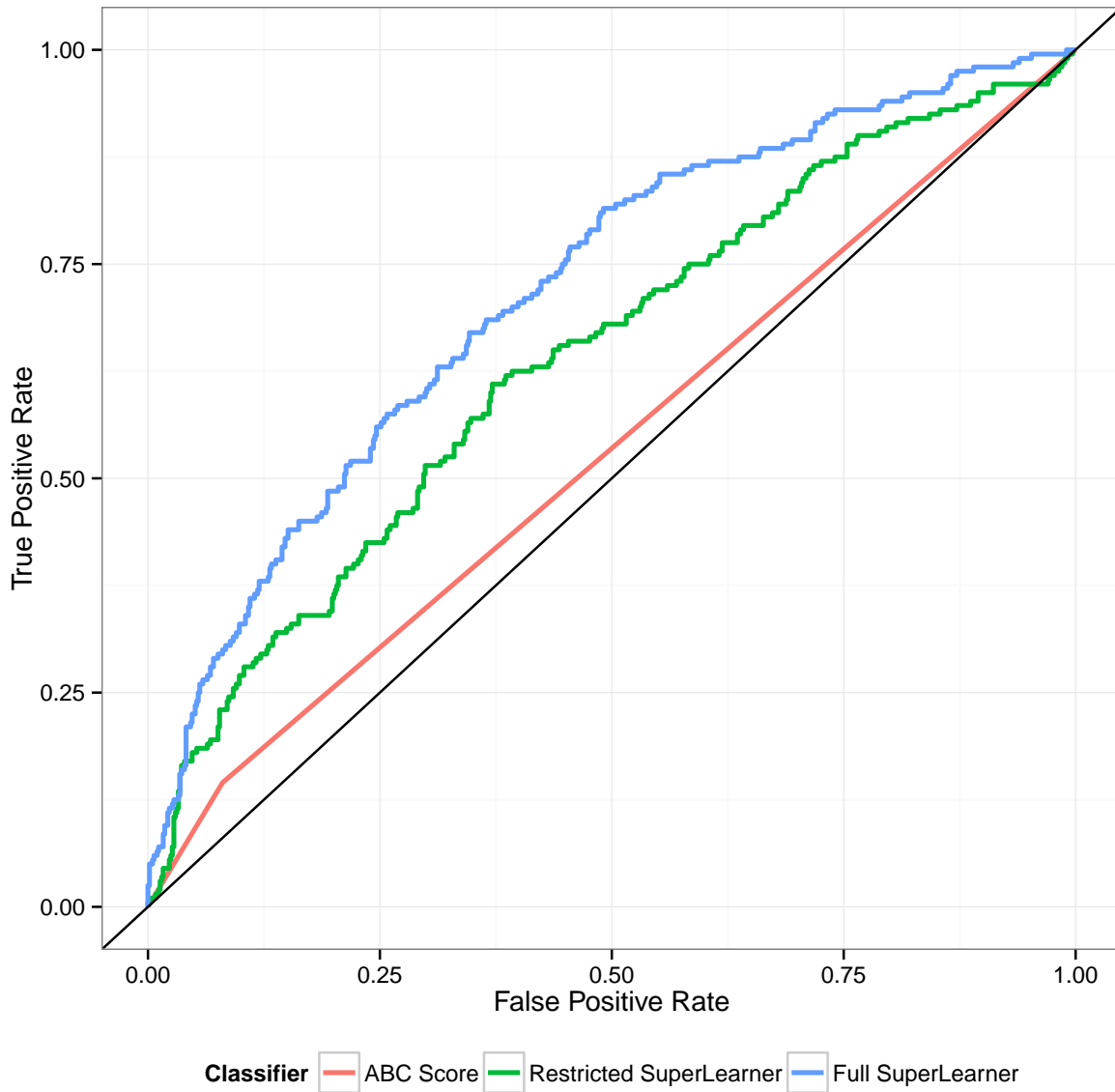


Figure 2.9: Receiver operating characteristic curves for the prediction of massive transfusion using the ABC score, a SuperLearner built using the ABC score variables, and the full SuperLearner

	Lower Bound	AUROC	Upper Bound
ABC Score		0.532	
Restricted SuperLearner	0.595	0.640	0.685
Full SuperLearner	0.677	0.718	0.758

Table 2.5: Comparison of the area under the receiver operating characteristic curve for predicting massive transfusion using the ABC score, a SuperLearner that uses the ABC score variables, and the full SuperLearner

2.4 Discussion

We have presented a principled approach to the prediction of clinical outcomes in critical care that has several desirable properties. SuperLearner allows the user to specify a library of candidate prediction algorithms that can range from very simple to more complex, ensemble learners. Including a variety of prediction algorithm strengthens SuperLearner because it allows for increased flexibility over possible functional forms to protect against model misspecification, and is simultaneously protected against overfitting by utilizing cross validation to obtain honest measures of predictive performance of the candidates and SuperLearner itself. Additionally, the individual candidate algorithms in SuperLearner can be unpacked and analyzed in greater detail, as we did with the massive transfusion outcome, allowing for increased interpretability and transparency, which machine-learning algorithms often obscure.

For the binary outcomes, we were also able to obtain inference for the AUROC performance measures using the influence curve, as derived in [59]. While we did not see a distinct advantage of using SuperLearner as opposed to main-terms regression or stepwise regression, we did not know how these prediction algorithms would perform *a priori*. In fact, this procedure allowed us to consider many competing candidate prediction algorithms, compare them in a systematic way, and combine them using a procedure designed to maximize the generalizability of the resulting prediction model, thereby making the most efficient use of data we had.

Upon delving into the algorithms given weight in SuperLearner based on their cross validated risk, we were able to identify particular predictors that had a strong association with the need for a massive transfusion. Interestingly, some variables identified as top predictors were in fact indicators of whether a particular variable was measured, suggesting an association between missing measurements and the massive transfusion. In such a chaotic environment where treatment decisions are made rapidly, there may not be time to take certain measurements and clinicians might rely on their previous experiences to

CHAPTER 2. SEMIPARAMETRIC PREDICTION

identify high-risk patients rather than take the extra time to collect data. Given these variables and the prediction algorithms selected by SuperLearner, this motivated us to examine the predictive ability of the ABC score, a diagnostic score commonly used to predict the need for massive transfusion and prized for its simplicity. We compared this score to a SuperLearner built using the four variables used in the ABC score and also a SuperLearner built using all the predictors and found that the ABC score barely outperformed a random classifier with an AUROC of 0.532. The two other methods performed substantially better, suggesting that the ABC score may be an oversimplification of the information contained in the four variables that go into its calculation and that further predictive ability is ignored when the ABC score is used.

We have demonstrated the utility of data-adaptive machine learning prediction of clinical outcomes in critical care and highlighted the importance of a principled approach to the prediction problem. While we did examine some *ad hoc* variable importance measures, we present a more targeted approach in the next chapter.

Chapter 3

Variable importance over time using irregularly measured genomic data

3.1 Introduction

In addition to utilizing clinical data to improve the treatment of trauma injury, understanding the patient response to injury at the gene expression level can help guide physicians in making treatment decisions and identify high-risk patients. As described in Chapter 1, common reactions to injury include inflammation and coagulopathy, which are partially regulated at the gene level. Thus, exploring the importance of gene expression after injury can help illuminate the underlying mechanisms of patient reaction to trauma as well as pathways by which these processes act to guide clinicians' decision making. Massive amounts of such data can be generated for each individual, resulting in a complex, high-dimensional data structure that can be further extended to include repeated measurements at irregular points in time during follow up. The aim of this paper was to present a principled approach to derive variable importance measures (VIM), motivated by causal inference with the aim of producing clinically interpretable statistical parameters and robust statistical inference to better understand the genomic response to injury and how the gene profiles vary over time. We present results for the entire Inflammation and Host Response to Injury cohort (which was described in Chapter 1) as well as more detailed results for a subset of genes involved in the coagulation and inflammation pathways (described in 1.5).

Obtaining a measure of variable importance or a ranking of predictors is not a new problem in statistics nor in the trauma literature. Some commonly-used methods of obtaining measures of variable importance first build a prediction model and the compute of variable

importance quantities based on how the predictions change when variables are used in the model. Some approaches include random forests and neural networks [47, 60–62]. Random forests use Gini and permutation variable importance measures that are based on the mean decrease in classification accuracy using resampling techniques or permuting the predictors [41, 47]. In neural networks, the weights that connect the predictors, hidden linear combinations of these variables, and the outcome of interest can be used as measures of variable importance [63]. The importance measures generated by these methods do not have a clear definition as a statistical or clinical quantity of interest nor do they have an interpretation based on the “natural experiment” that generated the observed data, i.e. how the patient’s outcome would have changed if they had had a different gene expression pattern. There is also no guarantee that an algorithm that will predict clinical outcomes well will do a good job estimating a VIM since they aim to achieve the optimal bias-variance tradeoff for the entire outcome model rather than for the particular parameter of interest (the VIM). In the trauma literature, VIMs are usually the coefficients associated with each predictor in a main-terms regression, which is not feasible for high-dimensional data where the number of predictors is much larger than the number of individuals for which predictors were measured. Additionally, little work has been done in the trauma literature on variable importance for genomic response to injury beyond differential expression analyses. Thus, a principled approach to obtain variable importance measures for many predictors is an unmet need in critical care that could have useful on the identification of high risk patients and have and widespread implications for the improvement of the treatment of trauma injury.

The data used for this analysis were a subset of the Inflammation and Host Response to Injury cohort ($n = 167$), who had gene expression measured in peripheral blood leukocytes at various time points during their treatment [36, 37]. This data set was described in detail in Chapter 1, specifying our outcome of interest as an indicator of whether the patient died or experienced multiple organ failure. Additionally, we wanted to adjust for a set of baseline covariates of interest: injury severity score (dichotomized at 15), base deficit (a measure of the acidity of the blood) dichotomized at -6, and INR (international normalized ratio, a measure of blood coagulation time) dichotomized at 1.3. Previous research using these data found that gene expression in trauma patients is substantially different from uninjured patients and established that the inflammation and coagulation pathways are significantly enriched in these individuals [37, 39, 40]. The Affymetrix U133 microarray chip measures the gene expression for over 45,000 probes, several of which were taken for each patient at irregular times after injury, resulting in a high-dimensional longitudinal data structure. Clinicians were interested in obtaining time-specific variable importance measures to better understand how patient response to injury changes within and across time. The dimensionality of the available data highlighted the need for a principled approach to deriving variable importance measures. The parameter estimate for these

data can be interpreted as the relative risk under two interventions that deterministically set gene expression to high versus low. This parameter, motivated by causal inference, allowed us to examine time-specific variable importance measures in a high-dimensional longitudinal data set. In addition to the entire probe set, we examined more closely a subset of genes involved in the coagulation and inflammation pathway (described in Table 1.5), which were of particular interest to clinicians since these pathways are often enriched in trauma patients. We found variability in the magnitude and direction for the association of some of these genes with death and multiple organ failure within and across time, but also found that these are not the only “important” genes associated with death and multiple organ failure, suggesting that the underlying mechanisms by which trauma patients react to injury are complex and require further research to fully understand.

3.2 Methods

3.2.1 Data

Our goal was to use irregularly spaced gene expression measurements to predict the probability of death or multiple organ failure at serial cross-sectional time points: 12, 24, 48, 72, and 96 hours, using data from the Inflammation and Host Response to Injury cohort (described in detail in Chapter 1). We used the closest observed gene expression measurements for each individual at each time point for which we wanted predictions, shown in Algorithm 2. For example, suppose an individual had expression measurements at 11 hours and 20 hours. Since the measurements taken at hour 11 are closest to hour 12 and also occurred before hour 12, they are carried forward. We did not extrapolate outside the minimum or maximum time measurements for a patient, but did use observed gene expression measurements from up to 120 hours (24 hours past the largest time over which we want to interpolate). This approach maintained the time-ordering of the observations, allowed us to examine the entire cohort of patients who were still alive by a given time point, and respected the fact that gene expression is highly variable across individuals.

Algorithm 2: Algorithm to smooth over the irregularly measured time points

```

for  $i \in 1 \dots n$  do
  if any observed times for individual  $i$  are greater than 120 hours then
    | Remove these observed values;
  end
  Calculate the absolute values of the distances between all possible combinations of
  observed time points and knots (time points where we want to obtain predictions);
  while possible combinations of observed time points and knots exist do
    | Determine the knot and observed time pair with the smallest absolute distance;
    | Assign that knot the expression measures from its paired observed time point;
    | Remove the knot and observed time point from the set of possible pairs;
    | Recalculate the absolute values of the distances between all remaining possible
    combinations of observed time points and knots;
  end
  if the number of observed times is smaller than the number of knots then
    | Assign missing value to the knots for which there is no closest observed time and
    expression
  end
end

```

For the variable importance analysis, we dichotomized the interpolated expression measures around $\log_2(100)$, defining “high” expression as above this level and “low” expression at or below this level, based on the distribution of the overall expression values, which is shown in Figure 3.1.

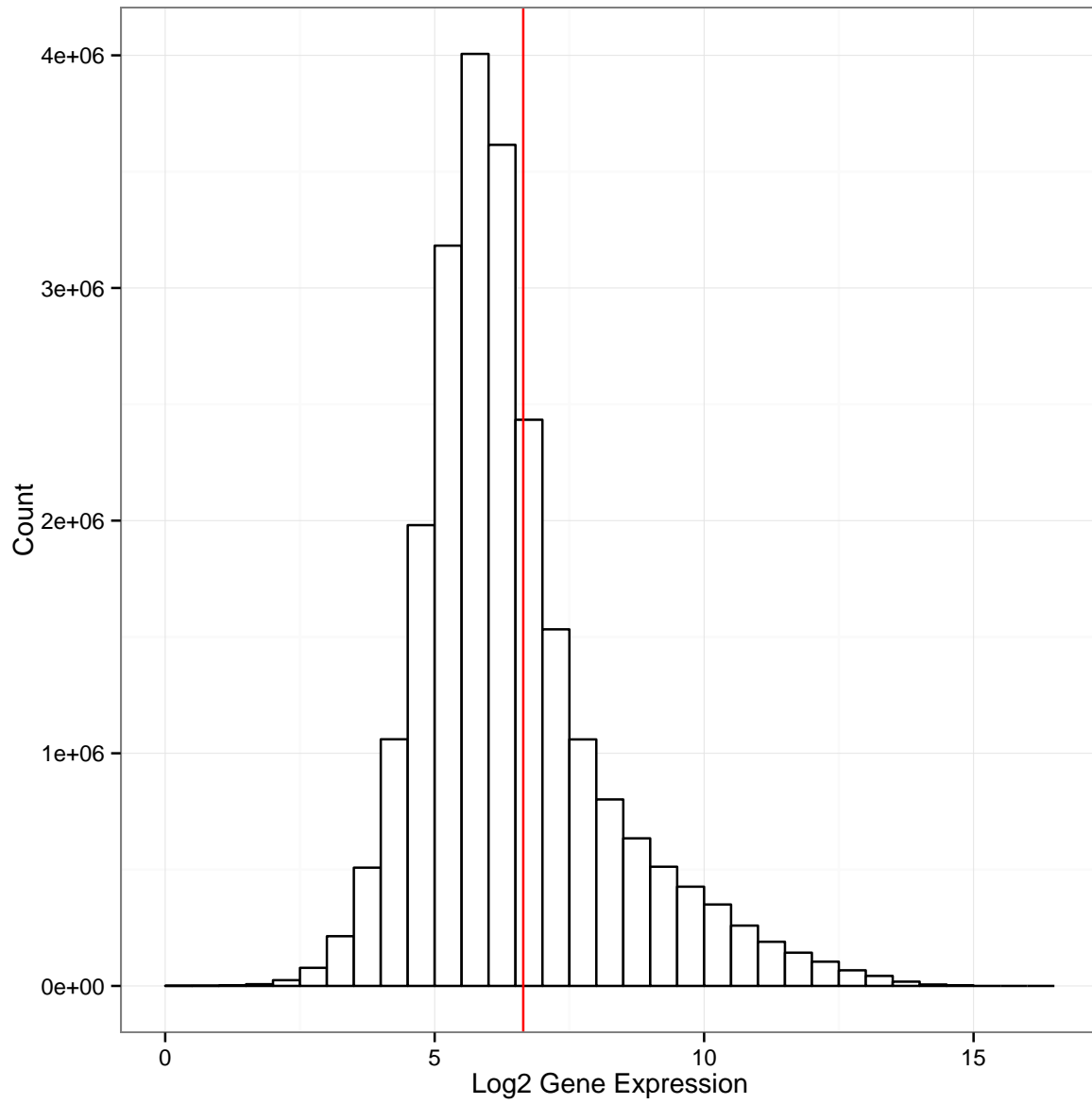


Figure 3.1: Histogram of gene expression with the $\log_2(100)$ cutoff shown in red

3.2.2 Parameter of interest and identifiability assumptions

Following the causal inference roadmap laid out in Chapter 1, we first defined our question of interest, specified the structural causal model that defined our background knowledge about the relationships between variables at each time point, and defined our parameter of interest, which can be translated into a statistical parameter (under some assumptions), which we call a variable importance measure (VIM). Let $Y(t)$ represent the binary outcome of interest (death or multiple organ failure) at time t , $A(t) = A_1(t), \dots, A_J(t)$ represent the vector of expression measurements for each probe of interest dichotomized as high or low at time t , and W represent a set of baseline covariates for which we need to adjust. Then at a given time point t , we represent the observed data for a given individual as $O(t) = (W, A(t), Y(t)) : t = 12, 24, 48, 72, 96$ hours.

Identifying informative variables in the context of a large number of candidate predictors was a challenge because of the high dimensionality of these data. Strategies for variable selection using gene expression data usually use a univariate measurement of how an individual gene is related to the outcome such as a t -test or rank test or a weighted combination of expression measurements (e.g. principal components analysis) to reduce the dimensionality of the data [64, 65]. The univariate approaches do not adjust for confounding variables and while the weighted combination approaches take into account the dependence between the genes, the results are difficult to interpret and assessing gene-level effects becomes challenging. Another approach is to build prediction algorithms and rank variables based on how their inclusion in these models affects the predictions. However, whether or not a gene is chosen by a prediction algorithm is not necessarily the best measure of its importance in relation to the outcome of interest.

We wanted to define a meaningful parameter in a semiparametric model to rank genes based on their relationship to the outcome (death or MOF). Causal inference provides such parameters by first requiring the specification of a structural causal model (SCM) that represents the relationships between variables of interest, then defining causal parameters based on interventions on this SCM and, through some assumptions, allows for the specification of statistical estimands. Our parameter of interest was ratio of the probability of death or MOF if each patient had had high gene expression over the probability of the outcome if each patient had had low gene expression, that is, a ratio of so-called counterfactual outcomes (shown below).

3.2.3 Structural Causal Model

To formally define the VIM, consider the following structural causal model (SCM), which specifies the relationship between each of the variables of interest at each time point t and

for each gene j .

$$\begin{aligned} W(t) &= f_{W(t)}(U_{W(t)}) \\ A_j(t) &= f_{A_j(t)}(W(t), U_{A_j(t)}) \\ Y(t) &= f_Y(W(t), A_j(t), U_{Y(t)}) \end{aligned} \tag{3.1}$$

where the U s represent unmeasured variables that affect each of $W(t)$, $A_j(t)$, and $Y(t)$ and are assumed to have their own distribution, denoted P_U [8]. This system of equations is non-parametric because we have made no statements about the functional form of any of the equations. Rather, they are generic functions of the random errors and each variable's parents. Then the parameter of interest could be defined in terms of counterfactuals generated by interventions on this system of equations, setting $A_j(t) = a$ where a is either 1 or 0, corresponding to high or background expression, respectively. These counterfactuals are denoted $Y_a(t)$ where a is the intervened expression value and the parameter of interest is some function of these counterfactuals. Our parameter of interest was the causal relative risk, which can be denoted by

$$\Psi(P_{U,X}) = \frac{E[Y_1]}{E[Y_0]} \tag{3.2}$$

where $P_{U,X}$ denotes the distribution of the full data, including the unmeasured variables. Since the analysis is repeated for each time t and gene j , we drop the indices from the notation in what follows.

If we wanted to estimate the causal relative risk, we would need the assumption of no unmeasured confounding (also known as the randomization assumption (RA)) [8]. That is, that there are no unmeasured common causes of the gene expression and the outcome, and that the same is true for unmeasured common causes of either gene expression and the covariates, or for the covariates and the outcome. Additionally, we need the consistency assumption, which assumes that $Y_A = Y$, that is, for each individual, their observed outcome corresponds to their counterfactual outcome under the observed intervention. Finally, we would need the assumption of positivity, that is, that there is some variability in the observed exposure in every strata of the covariates [8, 66]. If these assumptions were met, we could write

$$E[Y_a|W] \stackrel{RA}{=} E[Y_a|A = a, W] \stackrel{con.}{=} E[Y_A|A = a, W] \stackrel{pos.}{=} E[Y|A = a, W] \tag{3.3}$$

This shows how we can estimate, from the observed data under some assumptions

$$\Psi(P_0) = \frac{E[Y_1]}{E[Y_0]} \quad (3.4)$$

$$= \frac{E_0[E_0(Y|A=1, W)]}{E_0[E_0(Y|A=0, W)]} \quad (3.5)$$

$$= \frac{E_0[Q_0(1, W)]}{E_0[Q_0(0, W)]} \quad (3.6)$$

where Q_0 denotes the mean of the conditional probabilities of the outcome given the covariates and gene expression. Estimates of this relative risk can be sensitive to extreme values of the numerator and denominator and requires some variability in the gene expression in order to be defined. In this case, we did not feel comfortable making the necessary assumptions to give our statistical estimates a causal interpretation. We instead reported a purely statistical parameter, which is still clinically meaningful and interpretable but does not have a causal interpretation.

3.2.4 Substitution estimator

A simple substitution estimator of this statistical estimand would require estimating the conditional probability of the outcome given the covariates and gene expression for each gene at each time point, denoted by Q_n . Then, one could obtain predictions under each intervention by deterministically setting the gene expression to high and low and predicting the outcomes using the fit Q_n . Using the empirical distribution for W , a simple substitution estimator would be given by

$$\Psi(P_n) = \frac{\frac{1}{n} \sum_{i=1}^n Q_n(1, W)}{\frac{1}{n} \sum_{i=1}^n Q_n(0, W)} \quad (3.7)$$

Inference for this estimator would require use of the nonparametric bootstrap, for every probe at every time point, which was computationally infeasible. Additionally, since a simple-substitution estimator based on SuperLearner is not an asymptotically linear estimator, this inference would not be reliable [9, 67]. As described below, a follow-up step can be used to reduce the bias of the estimator and offers a non-computationally intensive estimate of the variance.

3.2.5 Estimation of Q

In order to estimate the conditional probability of the outcome given a gene's expression and the covariates, we used a machine-learning algorithm called SuperLearner, which is described in detail in Chapter 2. Since we do not know *a priori* the statistical model that will best describe the true underlying distribution of the data, SuperLearner offers an attractive approach to modeling Q_0 in a big model that has some theoretical justification [9, 49].

The dimensionality of W in this analysis was relatively low (only 3) and each variable was dichotomized. Thus, we specified a sieve of algorithms supplied to SuperLearner that ranged from the smallest possible model (only $A_j(t)$) all the way to a saturated model, using stepwise regression, making the library supplied to SuperLearner trivial.

3.2.6 Targeted maximum-likelihood estimation (TMLE)

While SuperLearner does the best possible job estimating the entire conditional mean of the outcome given the gene expression and covariates, it is not targeted towards estimating a particular feature of this distribution. Targeted maximum-likelihood estimation (TMLE) is a bias-reduction procedure that focuses in on the particular parameter of interest and does result in an asymptotically linear estimator. It requires estimation of the conditional probability of the gene expression (high versus low) given the covariates (the so-called treatment mechanism), denoted by $g_O(A|W)$. The advantages of the resulting estimator are that it is double-robust (i.e. the estimator is consistent if either the outcome regression or treatment mechanism estimator is consistent), locally efficient in a semiparametric model, and achieves the efficiency bound if both Q and g are estimated consistently [9, 68]. Additionally, its variance can be calculated with little computational cost using the influence curve.

One could estimate $g_O(A|W)$ using SuperLearner. However, since the estimates from TMLE can become unstable when $g_0(A|W)$ is fit aggressively we relied on the double-robust property of the TMLE and fit the initial estimators of $\bar{Q}_0(A, W)$ using SuperLearner and $g_0(A|W)$ using logistic regression. Note: an alternative to this approach would be to utilize collaborative targeted maximum-likelihood estimation (CTMLE), which adaptively chooses the adjustment set to achieve the optimal bias reduction [69]. The treatment mechanism can then used to augment the initial estimate of $\bar{Q}_0(A, W)$, as described below. We targeted the numerator and denominator of the VIM separately and then used the delta method to calculate the influence curve to obtain inference.

For each probe, j , at each time point t , we used the following steps to obtain the TMLE

for the treatment-specific means (the numerator and denominator of the VIM) separately and then used the δ method to derive the influence curve for the relative risk:

1. Generate an initial estimate, denoted $\bar{Q}_n^0(A, W)$ of $\bar{Q}_0(A, W)$ using SuperLearner and an estimate $g_n(A|W)$ of $g_0(A|W)$ using logistic regression. The superscript on the estimate indicates that it is an initial estimate.
2. Calculate the so-called clever covariate

$$H_n^*(A, W) = \frac{I(A = a)}{g_n(A|W)} \quad (3.8)$$

and the corresponding clever covariates under the interventions on gene expression, denoted $H_n^*(1, W)$ and $H_n^*(0, W)$. Note that in this step, extreme values of $g_n(A|W)$ can result in extreme values of the clever covariate, which results in instability of the final estimate.

3. Augment the initial estimator by regressing the outcome Y on the clever covariate, using the *logit* of the initial fit $\bar{Q}_n^0(A, W)$ as offset. Calculate ϵ_n , the coefficient attached to the clever covariate given by

$$\text{logit}(Q_n^1(a, W)) = \text{logit}(Q_n^0(a, W)) + \epsilon_n(H_n^*(a, W)) \quad (3.9)$$

4. The VIM is calculated using the updated estimates of Q

$$\hat{\Psi}(P_n) = \frac{\frac{1}{n} \sum_{i=1}^n Q_n^1(1, W)}{\frac{1}{n} Q_n^1(0, W)} \quad (3.10)$$

5. We targeted the numerator and denominator of the relative risk separately and then used the δ method to calculate the influence curve for this ratio of treatment specific means. The estimate of the variance of this estimator is given by the variance of the influence curve. For the relative risk, the influence curve is given by

$$IC_n = \left(\frac{1}{\mu_{a=0}} \left(\frac{I(A = 1)}{g_n(1|W)} \right) (Y - \bar{Q}_n^0(1, W)) + \bar{Q}_n^0(1, W) \right) - \left(\frac{\mu_{a=1}}{\mu_{a=0}^2} \left(\frac{I(A = 0)}{g_n(0|W)} \right) (Y - \bar{Q}_n^0(0, W)) + \bar{Q}_n^0(0, W) \right) \quad (3.11)$$

where $\mu_{a=1} = E[E(Y|A = 1, W)]$ and $\mu_{a=0} = E[E(Y|A = 0, W)]$

6. The standard error of $\Psi(\hat{P}_n)$ is given by

$$\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}} \quad (3.12)$$

where $S^2(IC_n)$ is the sample variance of the estimated influence curve values for each subject.

3.2.7 Implementation details

We first visualized the empirical distributions of gene expression values to determine whether there was sufficient variability in each gene to estimate our VIM, which is sensitive to extreme values in the numerator and denominator and used hierarchical clustering to examine potential multivariate groupings of individuals by their coagulation and inflammation gene expression patterns. Hierarchical clustering uses a dissimilarity matrix, in our case, based on the Euclidean distance between each subject, and joins the individuals iteratively until there is a single cluster. The results are displayed as a dendrogram, can be added to a heatmap of the entire data set to visualize the multivariate clustering. Thus, the heatmaps of the gene expression values before dichotomizing were a useful tool for both visualization of a subset of the data and exploration of the predictive ability of these expression values.

As a comparison to the VIM estimated using SuperLearner, we also performed unadjusted tests of association of the probability of death or MOF given gene expression levels and used Fisher’s exact test to obtain the p -value for that association. We accounted for multiple testing in both the adjusted and unadjusted analyses using the Benjamini-Hochberg adjustment and comparing these adjusted p -values to 0.05.

Finally, as an alternative to identifying important genes based on adjusted p -values, we used time-specific clustering of the individual components of the influence curve for the each VIM to identify important coagulation and inflammation genes as the medoids of these clusters. The VIMs were calculated adjusting only for baseline covariate measurements and not for other probes, but many of these genes belong to networks that may affect the outcome in a certain way. Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH), can identify groups of the top probes and genes that have similar relationship to the outcome and choose the probe and associated gene that best represents that group. Rather than clustering the raw gene expression values, however, we used HOPACH to cluster the individual contributions to the influence curve for the VIM (the relative risk) for each gene. Since the influence curve measures the impact that each observation has on the estimator, the individual components of the influence curve

themselves can be used in the identification of important coagulation and inflammation genes [68, 70].

HOPACH combines agglomerative (built from the bottom up like the hierarchical clustering algorithm described above) and divisive (built from the top down) clustering methods [71]. It uses the median split silhouette (a measure of cluster homogeneity) to collapse levels of a decision tree to unite smaller clusters and determine the optimal number of clusters [71]

We used HOPACH to cluster genes based on their individual contributions to the influence curve in order to group genes that had similar relationships with the outcome. At each time point, we clustered the values of the influence curve for the TMLE of the relative risk, shown in Equation 3.11, for each individual and probe in the coagulation and inflammation genes to determine the subset of genes that were representative of different impacts each gene had on the VIM. The clusters produced by HOPACH were centered at optimally-selected medoids and gave a reduced subset of the most important coagulation and inflammation genes for identifying patients at risk of death or multiple organ failure. Since the VIMs were calculated at the probe level, some of the clusters were centered at the same gene.

3.3 Results

We visualized the smoothed gene expression values using heatmaps at each time point with the aim of determining whether there was sufficient variability in each gene to estimate the VIM and also to explore the possibility of well-defined clusters based on the gene expression profiles. The heatmap for the coagulation and inflammation genes (described in 1.5) at hour 12 is shown in Figure 3.2. The gradient of colors in the plot was chosen based on the overall quantiles of the expression values at each time point, with warmer colors indicating larger expression values. The colors in the left-hand bar indicate whether individuals experienced multiple organ failure or death, with dark blue corresponding to a negative outcome. The dendrograms showing the results of the hierarchical clustering demonstrate that there is little clustering in relation to the outcome, but the genes themselves can be partitioned into several relatively distinct clusters based on their expression measures. Such plots can be constructed for each time point, but we did not see any striking clustering of the individuals using their gene expression values at any of the time points, suggesting that gene expression profiles alone cannot identify patients who experience death or multiple organ failure. These plots also demonstrated that, for some genes, the expression level is constant across individuals. Thus, when calculating the variable importance measures at each time point, we examined only probes that or

CHAPTER 3. VARIABLE IMPORTANCE

had some variability in expression, more specifically, the proportion of individuals with high gene expression had to be between 10% and 90%.

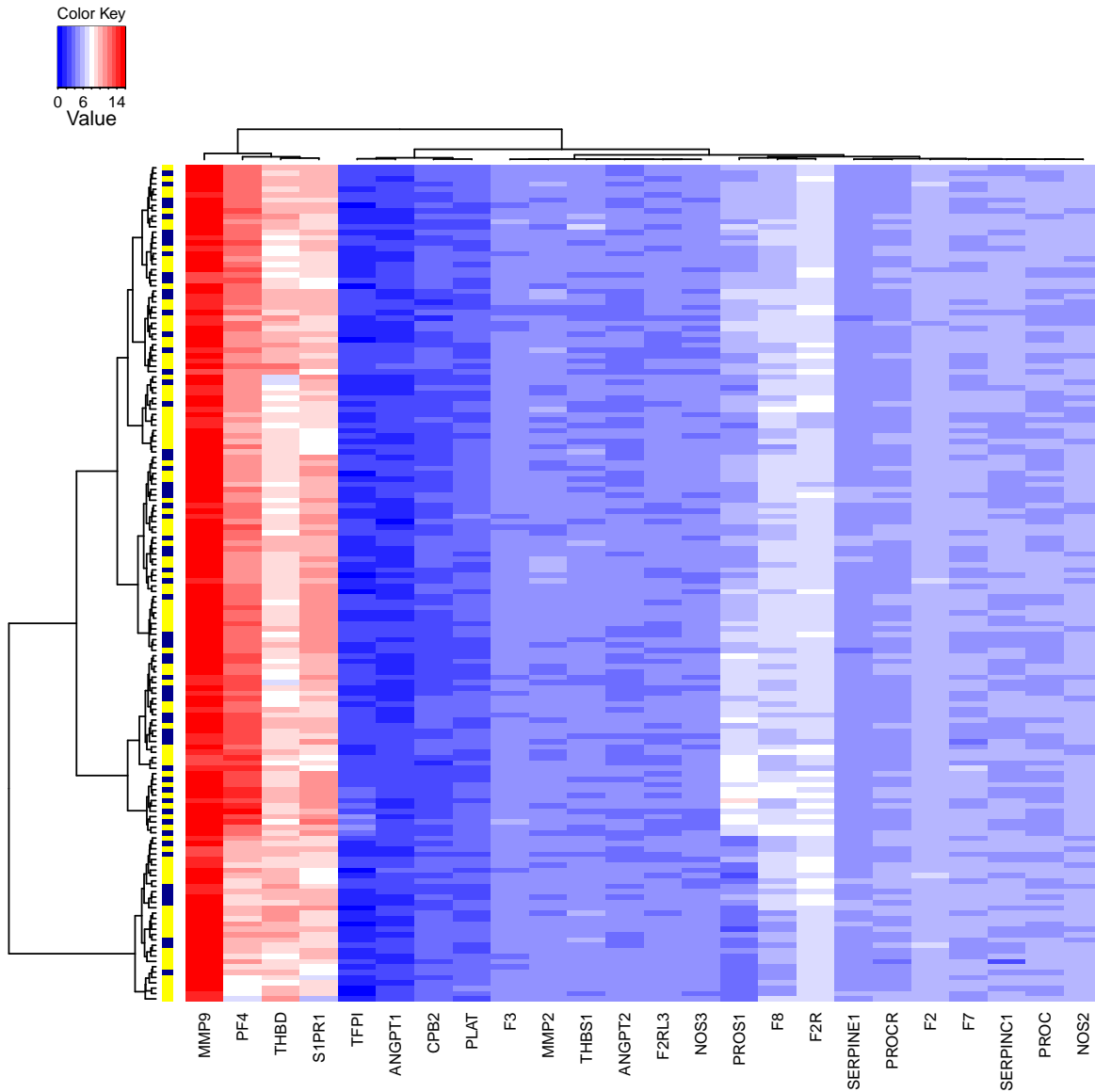


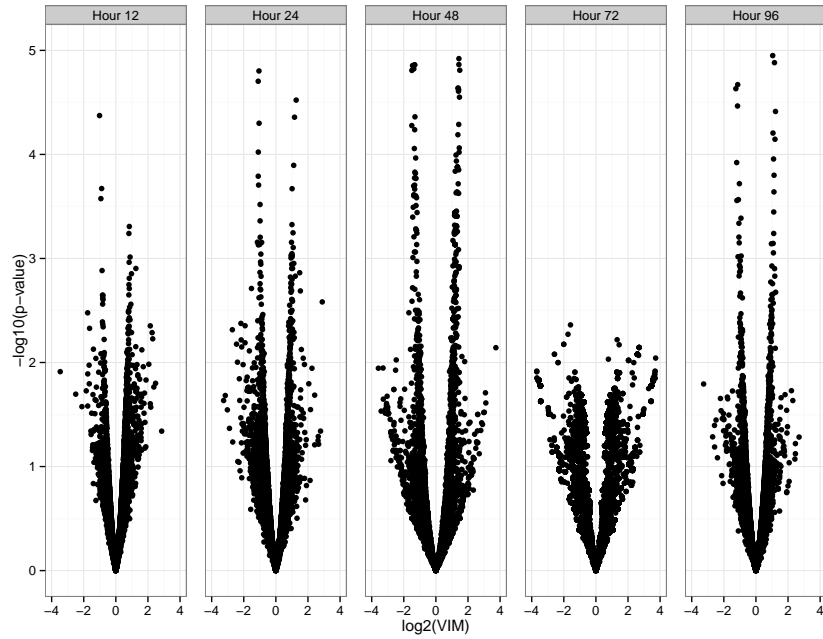
Figure 3.2: Heatmap of gene expression for coagulation genes at hour 12

The number of probes of each type (part of the coagulation set, the pathway set, or part of the entire probe set) that had any variability at each time point are shown in Table 3.1 with their corresponding gene class. Recall that probe had to be expressed in no fewer than 10% and no greater than 90% of the individuals at a given time point in order to be included in the analysis. The number that met the inclusion criteria varied over time, with the least amount of variability at hour 48. This narrowed down the number of probes for which we obtained measures of variable importance.

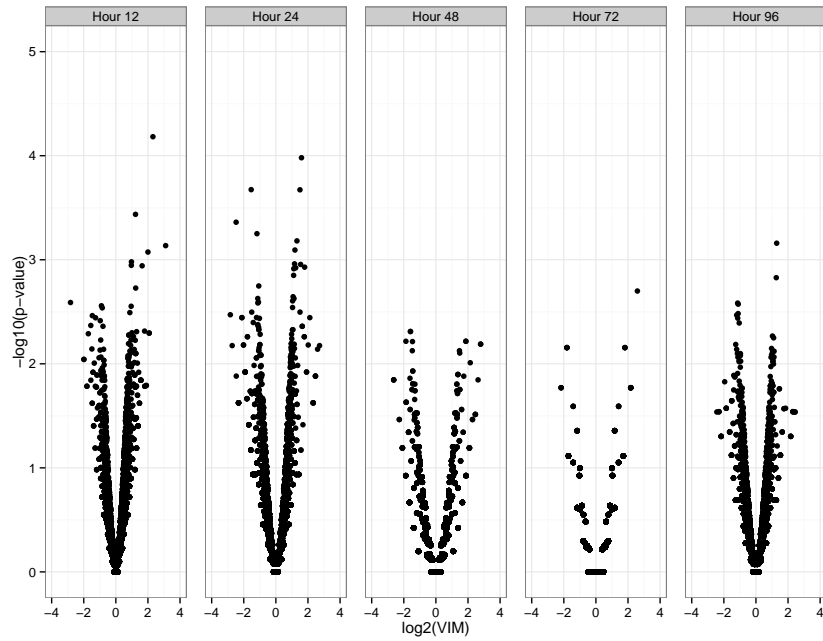
Time	Coagulation	Pathway	Other	Total
12	14	101	10171	10286
24	13	106	9726	9845
48	10	94	8838	8942
72	15	110	10285	10410
96	13	106	10104	10223

Table 3.1: Number of probes of each type that showed variability at each time point

We used two approaches to visualize the VIM results for all probes at each timepoint, one which allowed for the identification of probes with large magnitudes and small p -values and another which allowed us to compare the distributions of p -values for each probe type. These visualizations of all the results were carried out with the aim rapidly identifying specific stand-out probes and their associated genes. Comparing these plots for the adjusted and unadjusted analyses also helped us examine how adjusting for INR, base deficit, and injury severity affected the distributions of the p -values. Volcano plots, as shown in Figure 3.3, displayed the magnitude of the VIMs versus their $-\log_{10}$ transformed raw p -values and allow for the identification of probes with large magnitude VIMs and small p -values, which will appear in the upper corners of the plot. In both the adjusted and unadjusted analyses, a few probes stood out at each time point with small p -values but did not have extremely large magnitudes of the VIM. Boxplots of the distributions of $-\log_{10}$ transformed raw p -values shown in Figure 3.4 partitioned the distribution the p -values by the probe types. Based on these plots, it is not obvious that the proportion of statistical significant associations is higher among the coagulation genes than those in the larger pathway group or in the other genes. An exception could be at 72 hours, where the distribution of p -values appears to be smaller, which can also be seen in the volcano plots. Overall, from our visualizations of the results for the entire probe set, there was not a significant enrichment of the coagulation and inflammation pathways in these data. In fact as shown in Table 3.2, none of the coagulation or pathway genes survived the Benjamini Hochberg adjustment for multiple testing, but some of the other genes did most of them at hour 48, suggesting other pathways through which the patient reacts to injury. In the unadjusted comparisons, none of the genes had a significant association with death or multiple organ failure.



(a) Adjusted comparisons



(b) Unadjusted comparisons

Figure 3.3: Volcano plots of the log₂ transformed VIM and -log₁₀ transformed p-values over time for the adjusted and unadjusted comparisons

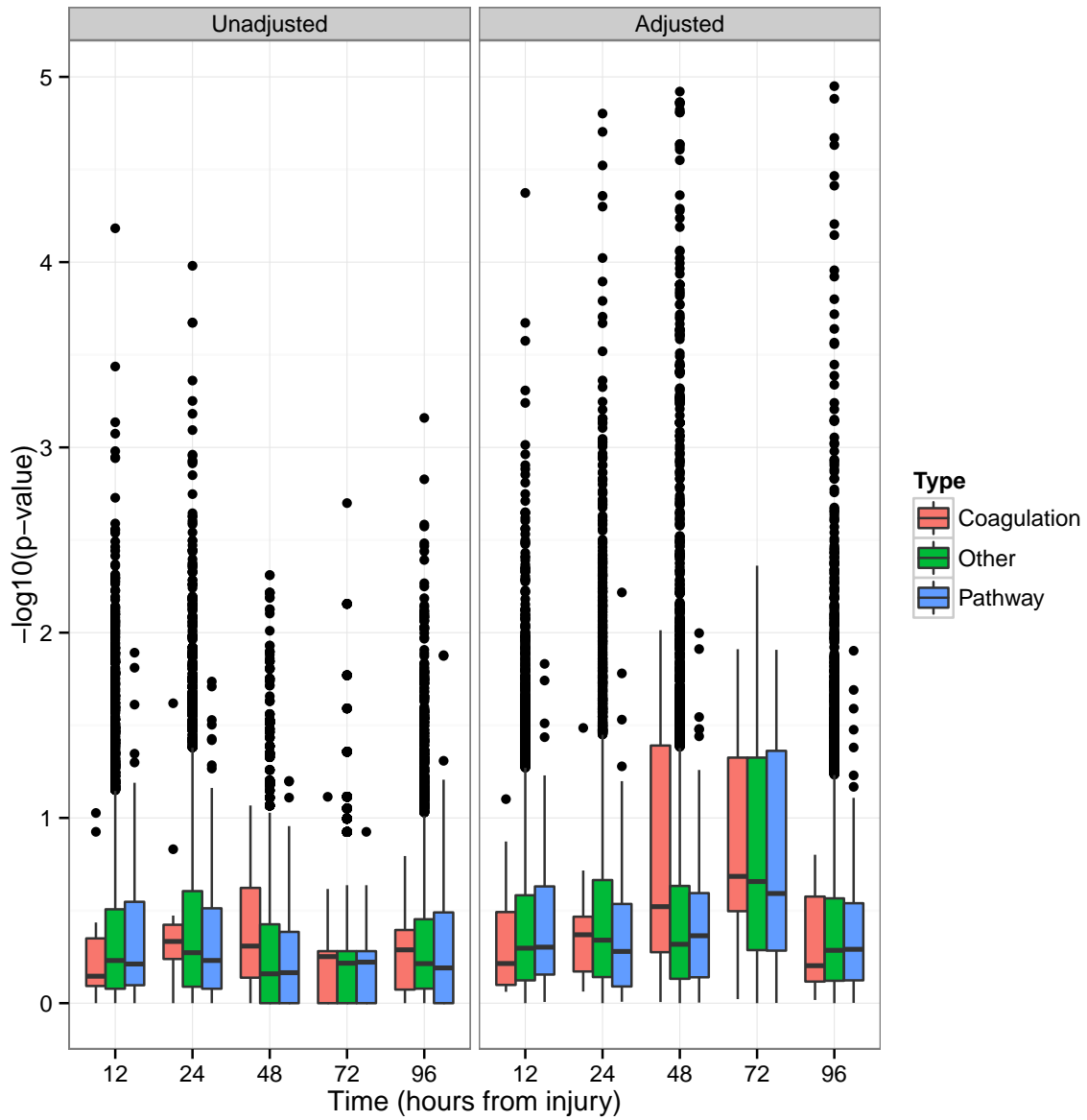


Figure 3.4: Boxplots of the unadjusted and adjusted \log_{10} transformed p-values for each gene type across time

Time	Coagulation	Pathway	Other
12	0	0	0
24	0	0	5
48	0	0	103
72	0	0	0
96	0	0	7

Table 3.2: Number of genes of each type that were significant at each time point

Although they were not identified as being significantly associated with death or multiple organ failure in the context of the entire probe set, we examined in more detail the specific VIM results for the subset of coagulation and inflammation genes because they had been previously identified by clinicians as some of the key drivers of immune response to trauma. Since each gene had several probes that mapped to it, we plotted the means of the VIMs for each of them for each time point in Figure 3.5, including the unadjusted relative risks for comparison. Some of the genes did not meet the inclusion criteria of being sufficiently variable at a given time point so they are missing points at some time points. While none of the coagulation genes had p -values that survived the adjustment for multiple testing, the magnitude of the VIMs may have been clinically meaningful. For example, at hour 48, *PROS1* was had a VIM value (relative risk) of approximately 2 in both the unadjusted and adjusted analyses, suggesting that having *PROS1* highly expressed is associated with a 2 times higher risk of mortality and multiple organ failure. For most of the genes, the adjusted and unadjusted comparisons were comparable with the exception of *THBD* at Hour 24, where the two are on different sides of a the null relative risk of 1, that is, once we adjusted for the covariates, having *THBD* highly expressed is associated with a higher risk of death or multiple organ failure rather than being protective. This gene encodes a thrombin binder, which reduces the amount of thrombin (the presence of which indicates clotting in the blood). Two of our covariates (INR and base deficit) were related to blood quality and ability to clot, and adjusting for them reversed the direction of the importance measure of *THBD*, demonstrating the impact of adjusting for these covariates.

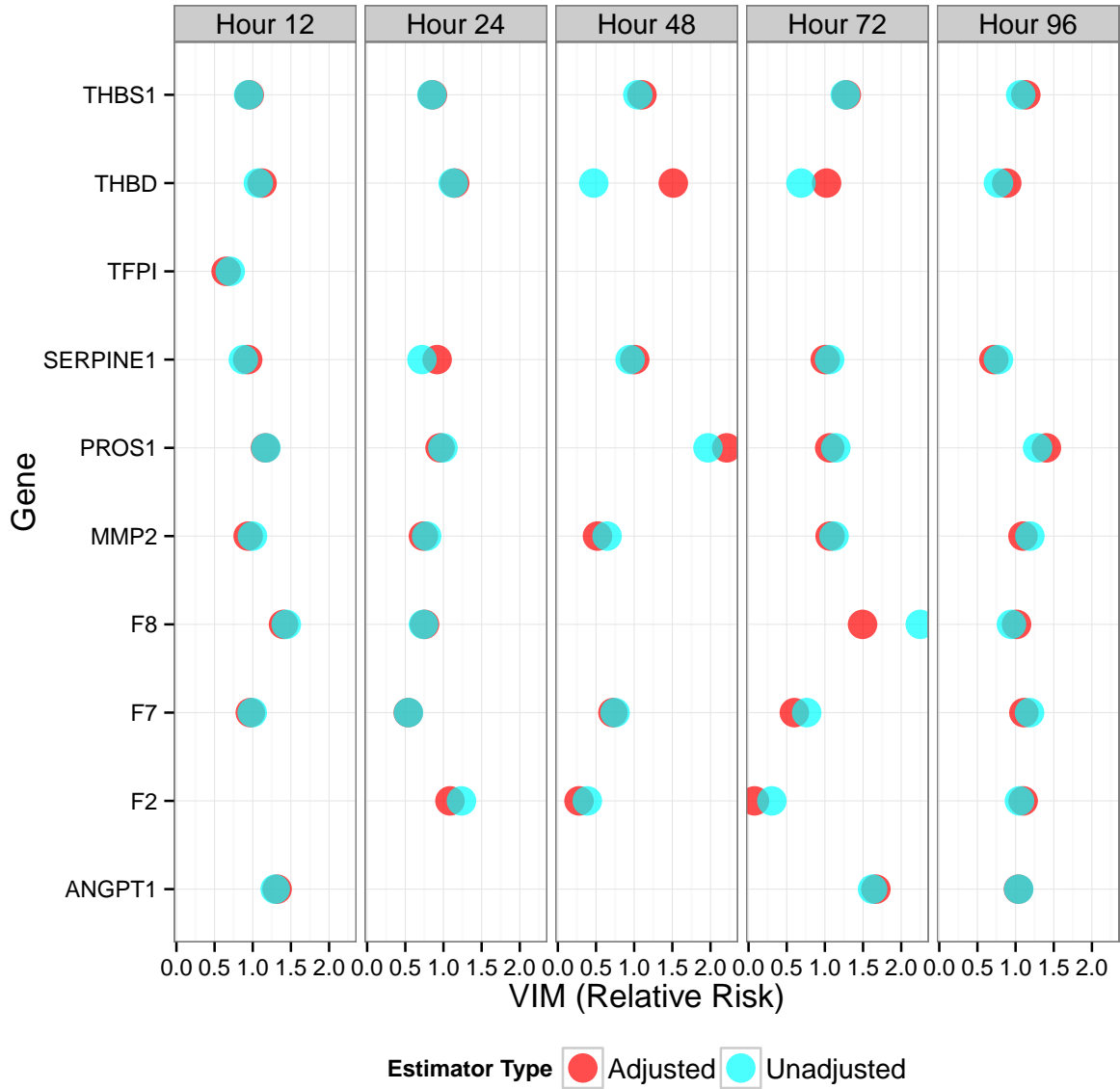


Figure 3.5: Adjusted and unadjusted variable importance magnitudes and significance levels for coagulation genes

The results from the clustering method of identifying the top genes at each time point are summarized in Table 3.3. The table indicates when and how often each gene was selected as a medoid in the clustering procedure described in the Methods section. While this procedure does not rely on the statistical significance of each variable importance measure, it is based on the impact each observation has on the TMLE of the relative risk and is a more multivariate approach to identifying the top genes at each time point that takes into account the other genes involved in the coagulation and inflammation pathway. However, it is less interpretable than ranking the genes by statistical significance. The number of clusters chosen at each time point ranged from 7 to 13, with some genes being selected as medoids more than once due to the mapping of multiple probes to each gene. Thrombospondin 1 *THBS1*, a gene involved in blood clotting, was identified as medoids at every time point, meaning that that it was important at every time point after injury. One of the main methods for the resuscitation of trauma patients is the infusion of blood products, especially if patients have lost a substantial amount of blood. Uncontrollable hemorrhage is one major cause of preventable mortality in trauma, and even the infusion of new blood products will not be successful if the blood cannot clot or clots too much, which supports the hypothesis that coagulation is important throughout follow up after injury.

Gene	Hour 12 9 clusters	Hour 24 7 clusters	Hour 48 8 clusters	Hour 72 7 clusters	Hour 96 6 clusters
THBS1	✓(3)	✓	✓(3)	✓(2)	✓(2)
SERPINE1	✓	✓		✓	
F7		✓	✓		✓
F2			✓		✓
ANGPT1	✓			✓	✓
F8	✓			✓	
THBD	✓	✓	✓	✓	
MMP2	✓	✓(2)	✓(2)	✓	✓
PROS1	✓	✓			

Table 3.3: Time-specific clustering results and genes identified as medoid centers. The numbers in parentheses indicate the number of times the gene was chosen as a medoid.

3.4 Discussion

We have presented an approach for assessing the importance of individual genes' expression levels over irregularly measured time points for critically injured patients, which produced statistical parameters with clinically relevant interpretations. The use of a

structural causal model enforced transparency regarding the relationships between variables of interest, and the variable importance measure derived based on interventions on this SCM has a causal interpretation under some assumptions. While we do not believe that the identifiability assumptions were met for these data, the VIMs are still interesting statistical parameters. This process highlighted the utility of a well-defined target parameter for variable importance analysis in high-dimensional data as well as the use of an approach motivated by a particular research question to derive a clinically meaningful parameter of interest. The results suggested that the drivers of these patients' responses to trauma were not limited to genes involved the coagulation and inflammation pathways, suggesting areas for future research to better aid clinicians in the identification of high-risk patients.

None of the unadjusted comparisons identified a significant association between gene expression and the likelihood of experiencing death or multiple organ failure. However, after adjusting for base deficit, INR, and ISS, and estimating the effect using TMLE and data-adaptive SuperLearning, we did identify a subset of genes that were significant at each time point, suggesting that a naive comparison that does not take into account the injury severity, clotting ability, and general health of the patient upon admission to the emergency department can be misleading. In the adjusted comparisons, no single gene was significant at every time point and the top genes changed over time. In the context of the entire probe set, none of the coagulation nor pathway genes were significantly associated with death and multiple organ failure, highlighting the complexity of the underlying mechanisms of patient response to trauma. It is possible that these genes are activated much earlier after injury and was not captured in our time scale, which started 12 hours after injury. Additionally, our analysis was limited to genes that had some variability in expression, genes with constant expression levels across individuals are not included here. For example, Protein C (*PROC*), a key gene in blood clotting, was not highly expressed at any of time points, possibly because patients were bleeding so severely that they there had been a depletion of the protein produced by the activation of this gene [72]. Indeed, the patients for whom we have expression measures were more likely to have a base deficit below -6, an INR of greater than 1.3, and more severe injuries, all of which have been implicated in injury-induced coagulopathy. Thus, a lack of variability in gene expression does not mean that gene is not important, but we could not obtain importance measures for such genes since there would be no support in the data for an intervention setting expression to be the missing expression level. However, a expansion of this analysis comparing these individuals with a healthy patient group could explore the importance of these consistent genes.

When examining the set of coagulation genes identified by clinicians *a priori*, we saw variability in the magnitude and direction of the importance measures within and across time, although these variations were not statistically significant. However, the variability

suggests that differences are occurring at the gene expression level in severely injured patients and that different pathways may be activated at different times. It also suggests that expanding the search for genomic markers of coagulopathy, multiple organ failure, and mortality might identify useful predictors to identify high-risk patients. The alternative method of identifying important coagulation genes does not depend on statistical significance but rather the influence each observation has on the estimator of the VIM. This procedure identified different numbers of optimal clusters at each time point as well as varied medoids of each cluster. However, Thrombospondin 1 (*THBS1*) stood out since it was chosen at every time point suggesting that it had a strong association with mortality and multiple organ failure. While this procedure did identify optimal clusters of genes based on individuals influence curve values for the relative risk, it does not have as clear of an interpretation as the significance of the VIM itself. However, it did identify two coagulation genes as being consistently important over time, which could be used as part of a genomic signature to identify high-risk patients, but would require prospective validation. Additionally, using HOPACH to cluster the influence curves for the entire probe set at each time point would be computationally infeasible, so some dimension reduction technique was necessary to use this method to identify important genes. We used genes identified by clinicians, but could have used some other criteria to identify candidate genes for the clustering algorithm, e.g. the genes with significant VIMs.

This analysis was limited to serial cross-sectional analyses of variable importance with a relatively small adjustment set of baseline covariates. This allowed for separate models for the relationship between the outcome given the gene expression and covariates to obtain time-specific rankings of genes. An expansion to longitudinal measures of variable importance would require the specification of a different structural causal model and parameter of interest and is another area for future research in the mechanistic understanding of trauma [73]. The use of CTMLE to adaptively choose the adjustment set W and perhaps include other genes as candidate covariates is another reasonable expansion of this analysis, which would result in a more realistic ranking of the genes [69].

While our variable importance measures do not have a causal interpretation due to unmeasured confounding, they were still a substantial improvement over the standard coefficients in a regression equation because they have a clinically meaningful interpretation, improved estimation, and robust inference. We were able to define these importance measures in the context of a complex, high-dimensional, longitudinal data structure and explore in more detail the relative importance of genes involved in the coagulation and inflammation pathways. The variability of these importance measures within and across time demonstrate the complexity of patient response to injury, highlighting the need for further research focused on the underlying mechanisms of injury. Furthermore, this analysis was limited to importance measures for gene expression only. A useful expansion would be to examine similar importance measures for gene or protein expression in com-

petition with clinical variables, which have so far been the focus on the trauma literature since they are easier to access during treatment, and would allow us to examine how clinical covariates are associated with the patient response to injury. While gene expression is not a modifiable patient characteristic, obtaining variable importance measures does assist with the mechanistic understanding of patients' response to trauma. Our results did not identify any single pathway or subset of genes that identify patients at high risk for mortality or multiple organ failure, highlighting the complexity of this response. We acknowledge that this analysis only scratches the surface of understanding the underlying human response to injury but we have demonstrated the utility of a principled approach to the analysis of complex, high-dimensional data.

Chapter 4

Quality of care comparison

4.1 Introduction

A large source of differences in trauma patient outcomes is hypothesized to reflect institutional-level variation in care. Comparing the quality of care at different hospitals trauma is a challenging question, since each center has unique doctors, resources, and patient types. For example, it would not be fair to compare two centers with different distributions of baseline injury severity since more severely injured patients experience a higher propensity for mortality. A vast literature exists suggesting that patients experience different outcomes based on the type, quality, and treatment practices practiced at different hospitals, for example, comparisons of large versus small, urban versus rural, and the use of a massive transfusion protocol [74–77]. Evaluation of the quality of care is usually based on comparisons of observed versus expected mortality and injury, which are relatively simple scoring systems that do not take into account the variability in underlying patient population served by each hospital or treatment practices. A qualitative ranking of trauma centers in the United States is carried out by the American College of Surgeons that ranks hospitals on a scale from level I to V (I being the highest, V being the lowest) [78]. In addition to the variation in care received, the effects of differences between hospitals is heavily confounded due to treatment by indication (that is, patients that are worse off tend to receive more treatment) as well as the heterogeneous nature of the patient cohorts. This analysis focused on finding an objective comparison that accounted for the underlying patient population and did not include treatment as a distinguishing factor among hospitals.

Using a statistical parameter motivated by the causal inference literature, we implemented four different estimators to compare quality of care at ten level one trauma

centers from around the United States. Our approach utilized two major advances in statistical methodology: (1) data adaptive machine learning tools for modeling clinical outcomes and the distribution of patient across different trauma centers, given a potentially large numbers of patient predictors, and (2) using the resulting models to estimate causal parameters, that provide relevant to summaries of how patient outcomes differ due to site. Specifically, we were interested in comparing large and small-volume sites involved in the study. Large-volume sites were identified based on the number of patients served and staff size. The statistical parameter we estimated allowed us to compare the large-volume patients outcomes at the center type at which they were observed with their so-called counterfactual outcome at a small-volume center [10]. This enabled us to identify outcomes where there was a benefit to being treated at a large-volume center as well as individuals who would be most affected by being treated at a different type of hospital.

4.2 Methods

4.2.1 Structural causal model and parameter of interest

We followed the causal inference roadmap laid out in Chapter 1 to motivate this analysis and estimate a parameter with a clinically meaningful clinical interpretation. Our question of interest was whether large-volume patients would have experienced different outcomes if they had instead been treated at a small-volume hospital. Consider the following structural causal model (SCM), which encodes the hypothesized relationship between the variables of interest, where W represents baseline covariates such as gender, race/ethnicity, or injury severity, A represents the site type (large- or small-volume), and Y represents the outcome of interest (various). The U 's represent unmeasured variables, which have their own distribution, denoted P_U .

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(A, W, U_Y) \end{aligned} \tag{4.1}$$

Our observed data can be denoted $O = (W, A, Y)$, and we are assuming that O_1, \dots, O_n are independent and identically distributed with distribution P_O . In this SCM, the baseline covariates confound the effect of site type on the outcome of interest, that is, we assume that the baseline covariates are not affected by site type. This is a reasonable assumption

given that severely injured patients are transported to the closest trauma center. Intervening on this system of equations to change the site type resulted in a modified SCM that generated a change in distribution of Y , the so-called counterfactual distribution. The resulting counterfactual we denoted Y_a where a is the value to which the site type was set in the SCM. The counterfactuals generated by interventions on this system are also known as potential outcomes [10]. Comparing some feature of the counterfactual distribution, such as the mean, under different interventions can illuminate the effect of site type on the outcome of interest.

The target causal parameter that addresses the effect of site type on a given outcome is the so-called effect of treatment among the treated (ETT), which is denoted by

$$E[Y_1|A = 1] - E[Y_0|A = 1] \tag{4.2}$$

that is, for the treated group, the average of their outcomes, minus their counterfactual outcome. In this case, the “treated” status refers to large-volume sites as opposed to small-volume sites. Under the randomization and positivity assumptions, the corresponding statistical estimand is given by

$$\psi(P_0) = E_0[E_0(Y|A = 1, W)E_0(Y|A = 0, W)|A = 1] \tag{4.3}$$

The randomization assumption states that $Y_a \perp\!\!\!\perp A|W$. This allowed us to link the counterfactual and observed data by writing

$$\begin{aligned} E_O[Y|A = a, W = w] &= E_{F_X}[Y_a|A = a, W] \\ &= E_{F_X}[Y_a|W] \quad \text{if } Y_a \perp\!\!\!\perp A|W \end{aligned} \tag{4.4}$$

where F_X represents the distribution of the full data, which is given by $X = (Y_1, Y_0, A, W)$ that is, all the potential outcomes, the site identifiers, and covariates. The positivity assumption requires that there be sufficient natural experimentation (variability) of each site type in each strata of the covariates, so that there will be support for each intervention. Under these assumptions, we can identify the ETT as a parameter of the observed data distribution.

Estimation of the statistical estimand required estimation of the conditional mean of the outcome among patients in large- and small-volume sites. Note that the covariates for

which we adjust are not on the causal pathway, i.e. they are not affected by site, nor are we adjusting for any treatment variables such as surgeries or infusion of blood products, which may be site-specific.

4.2.2 Estimation

The statistical parameter of interest can be estimated in several different ways. Before we expand on the details of each estimator, recall that we can represent our observed data as $O = (W, A, Y)$ and we assumed a common distribution, P_0 for O_1, \dots, O_n . The likelihood of the observed data can then be factorized as follows

$$\begin{aligned} P_0(O) &= P_0(W, A, Y) \\ &= P_0(Y|A, W)P_0(A|W)P_0(W) \\ &= \bar{Q}_0(A, W)g(A|W)Q_0(W) \end{aligned} \tag{4.5}$$

where the likelihood factorizes into an outcome regression of Y on A and W , the treatment mechanism, and the marginal distribution of W . We denote the outcome regression with $Q_0(A, W)$ and the treatment mechanism with $g(A|W)$.

Prediction-based estimator

The first estimator, called the prediction-based estimator, involved building a prediction model that utilized data from only the small-volume sites ($A = 0$ group), then predicting outcomes for the large-volume sites ($A = 1$ group) using that model. The motivation for this estimator was to model the relationship between the covariates and each of the outcomes and then examine how well this model predicted the outcomes for the large-volume patients using their covariate information. Subtracting the mean of the predictions based on the small-volume model from the mean of the observed outcome for the large-volume patients gave a substitution estimator of the ETT. Unfortunately, no formal inference exists for this estimator since it is not asymptotically linear.

Simple substitution estimator

For the simple substitution estimator we estimated the conditional mean of the outcome among both site types, then intervened and deterministically set the site variable to be the large volume indicator for all subjects ($A = 1$), obtained the predicted outcome under

this intervention, then did the same under an intervention that set the site to be low volume for all subjects ($A = 0$). Then, we examined the differences in means among only the large-volume patients, which yielded another substitution estimator of the ETT based on the G-computation formula [79]. This estimator is also not asymptotically linear, so we could not obtain inference.

Targeted maximum-likelihood estimator

The targeted maximum-likelihood estimator (TMLE) used the distributions of Y under each intervention from the simple substitution estimator described above as inputs into a bias-reduction step used to achieve the optimal bias-variance tradeoff for the ETT [9]. This follow-up step involved fluctuating the initial estimators using a clever covariate that involves the propensity score for each site type [9]. This estimator is also a substitution estimator that will respect the natural bounds of the outcome, double-robust (consistent if either the treatment mechanism or outcome regression is correct), and asymptotically linear. Thus, inference for TMLE can be carried out using the influence curve, which is given by

$$D^*(P_O) = \left(\frac{I(A=1)}{P(A=1)} - \frac{I(A=0)g(1|W)}{P(A=1)g(0|W)} \right) [Y - \bar{Q}(A, W)] + \frac{I(A=1)}{P(A=1)} [\bar{Q}(1, W) - \bar{Q}(0, W) - \psi(P_O)] \quad (4.6)$$

Propensity score matching

The propensity score matching estimator required estimation of the probability of each site type given the covariates (the so-called treatment mechanism), the inverse of which is then used to generate counterfactual outcomes for large-volume patients by matching each of them with a small-volume patient based on their propensity score and calculating the ETT in this new matched data set [80, 81].

SuperLearner

All of the estimators described above required building a prediction model either on the entire observed dataset or a subset of the data. The prediction, simple substitution, and TMLE required estimation of the outcome regression and the TMLE and propensity score matching also required estimation of the treatment mechanism. To avoid model

misspecification and do the best possible job of estimating each part of the likelihood, we used SuperLearner [9, 49]. This allowed us to use a library of prediction algorithms as well as cross-validation to protect against overfitting. The algorithms supplied to SuperLearner included main-terms regression, Bayesian regression, multivariate adaptive regression splines (degree = 2, penalty = 3), and generalized additive models. The first three were described in detail in Chapter 2. Generalized additive models were developed in 1990 by Hastie and Tibshirani and introduced a smoothing parameter into the form of a generalized linear model and models the outcome as a function of smoothed forms of the predictors [82].

In collaboration with clinicians, we identified several outcomes of interest, shown in Table 1.3 as well as potential confounders to adjust for, shown in Table 1.2. Some of the covariates did have missing values, so we used an indicator of whether the variable was measured to identify those individuals as a covariate in the prediction algorithm. The three large-volume sites treated 541 individuals while 681 were treated at the seven low-volume sites.

4.3 Results

The results for all of the estimators is shown in Table 4.1 with the unadjusted difference included for comparison. The parameter estimated the expected change in the outcome for individuals who were treated at high-volume sites had they been treated at low-volume sites. A negative value means that the predicted outcome would be higher than expected had patients from high-volume sites been treated at low-volume sites. For high-volume site patients, the probability of overall and 24-hour mortality would have been 7% and 5% higher, respectively, had they been treated at a low-volume site, as estimated by TMLE. In contrast, the probability of reported and data-based massive transfusion would have been 9% and 6% lower, respectively, for the same individuals. Additionally, the number of plasma and platelet units infused by 24 hours would have been lower at the low-volume centers. According to the estimates from TMLE, the remaining outcomes would not be significantly altered by a change in site. These included earlier mortality at 2 and 6 hours, substantial bleeding, multiple organ failure, complications, and the number of RBC units infused by 24 hours. The propensity score matching procedure also identified reported massive transfusion, overall mortality, and 24-hour mortality as outcomes that would be significantly altered by a site change in the same directions as the TMLE, but additionally found an 8% increase in the probability of mortality in large-volume patients if they had instead been treated at a small-volume center. While the unadjusted comparisons are not exactly analogous to the ETT (since the difference is calculated across all patients as opposed to only the large-volume patients), it does reflect the results of

CHAPTER 4. QUALITY OF CARE COMPARISON

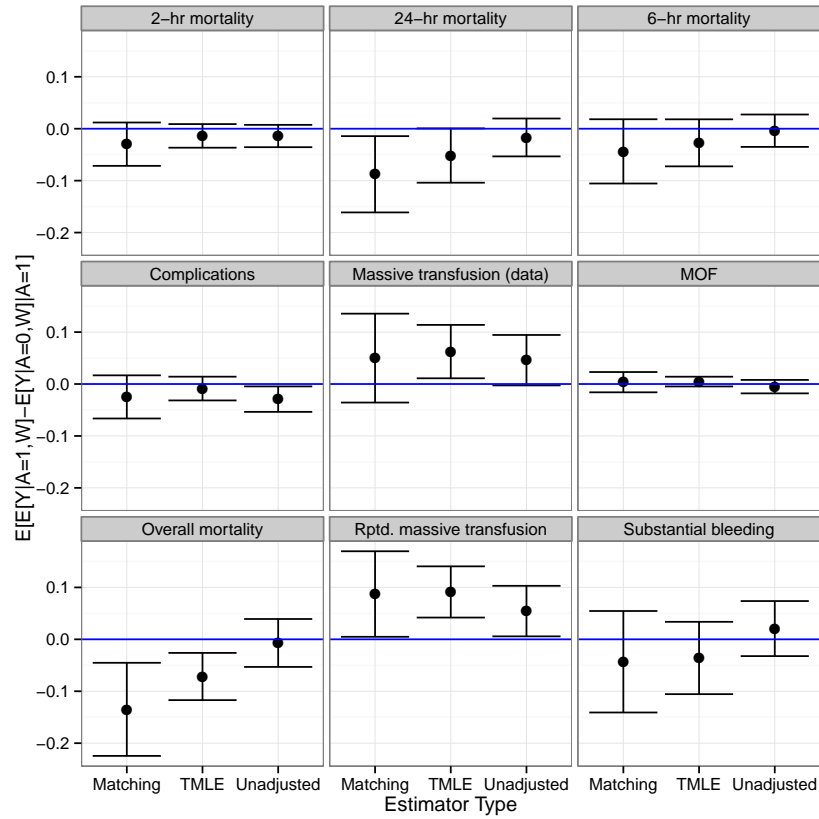
a naive analysis that does not take into account the underlying differences in patients at each site. This estimator identified plasma units infused by 24 hours and complications as differing significantly between the two sites, highlighting the importance of adjusting for confounders.

Overall, the direction of estimated effects from each of the four estimators was the same and the magnitudes were comparable. The significance of the results did not always agree, for example, unadjusted probabilities of complications differed significantly between the two groups but not in the TMLE or matching estimators of the ETT. The significant results by estimator are summarized in Table 4. Large-volume patients had a higher probability of being massively transfused and received more plasma and platelets and had lower probabilities of later mortality than they would have had if they were treated at a low-volume site.

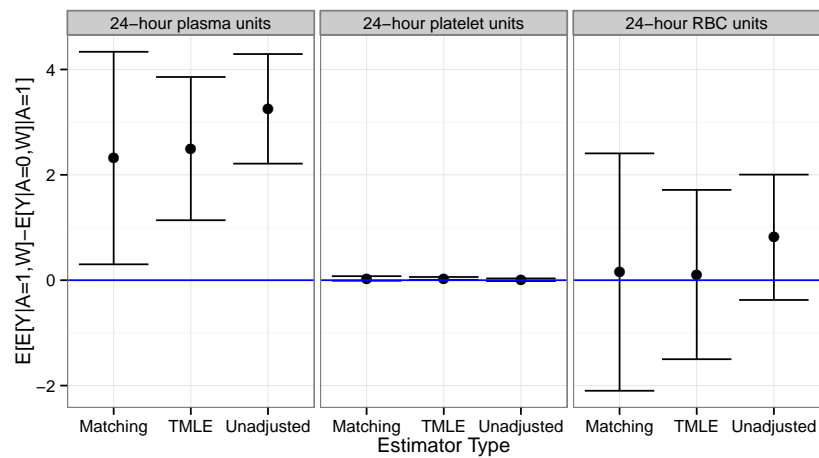
Variable	Unadj. Diff	Unadj. P-value	Pred.- based	Simple Subs.	TMLE	TMLE P-value	Matching	Matching P-value
MT (Rptd.)	0.05	0.03	0.05	0.08	0.09	0.00	0.06	0.06
Plasma 24h	3.25	0.00	3.22	3.55	2.50	0.00	3.20	0.00
Mortality	-0.01	0.82	-0.01	-0.01	-0.07	0.00	-0.02	0.51
Plt. 24h	0.05	0.52	0.04	0.16	0.21	0.01	0.07	0.48
MT (Data)	0.05	0.07	0.05	0.07	0.06	0.02	0.05	0.10
24h Mortality	-0.02	0.42	-0.02	-0.04	-0.05	0.05	-0.03	0.33
2h Mortality	-0.01	0.27	-0.01	-0.02	-0.01	0.23	-0.02	0.26
6h Mortality	-0.00	0.89	-0.00	-0.02	-0.03	0.24	0.00	0.92
Subst. Bleeding	0.02	0.48	0.02	0.03	-0.04	0.31	0.01	0.79
MOF	-0.01	0.61	-0.01	0.00	0.00	0.32	-0.02	0.13
Complications	-0.03	0.03	-0.03	-0.02	-0.01	0.45	-0.03	0.06
RBC 24h	0.81	0.18	0.81	1.26	0.11	0.90	1.01	0.23

Table 4.1: Results for all estimators

CHAPTER 4. QUALITY OF CARE COMPARISON



(a) Binary outcomes



(b) Continuous outcomes

Figure 4.1: Plots of the point estimates and confidence intervals for the propensity score matching estimator, TMLE, and unadjusted comparisons

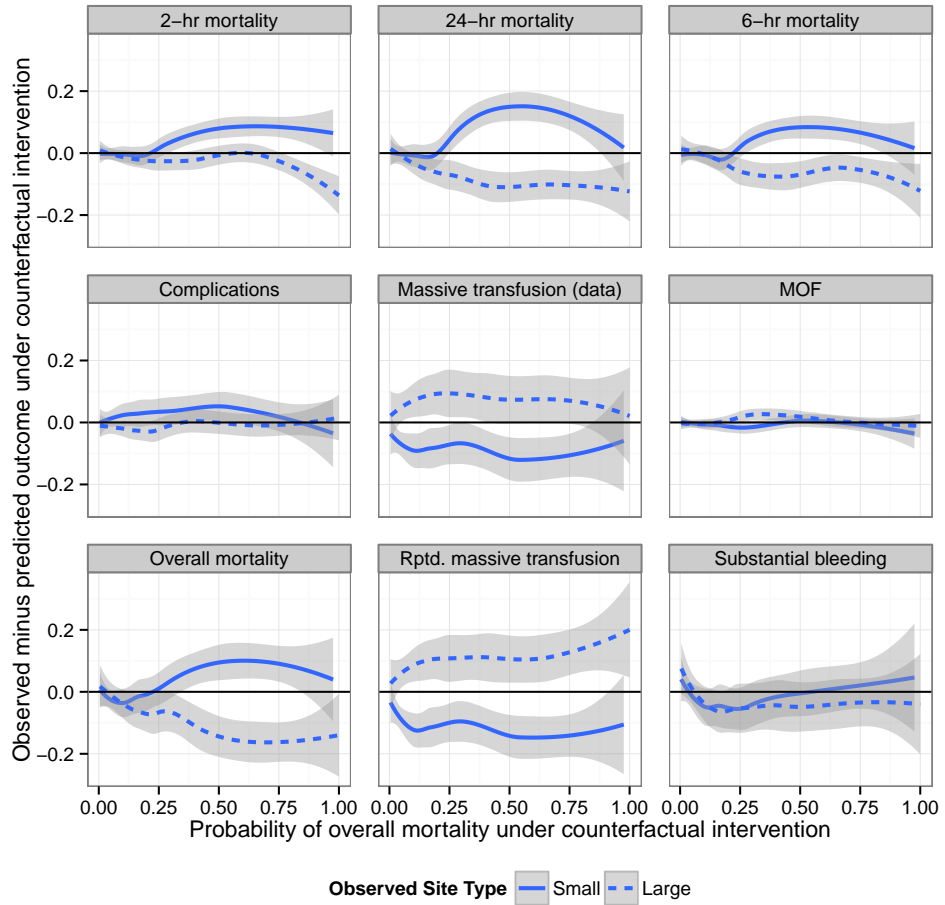
	Less at large-volume sites	More at large-volume sites
TMLE	Overall mortality	Massive transfusion (reported and from data)
	24-hour mortality	Plasma units by 24 hours Platelet units by 24 hours
Matching	Overall mortality	Massive transfusion (reported only)
	24-hour mortality	Plasma units by 24 hours
Unadjusted	Complications	Massive transfusion (reported only)
		Plasma units by 24 hours

Table 4.2: Significant results by estimator

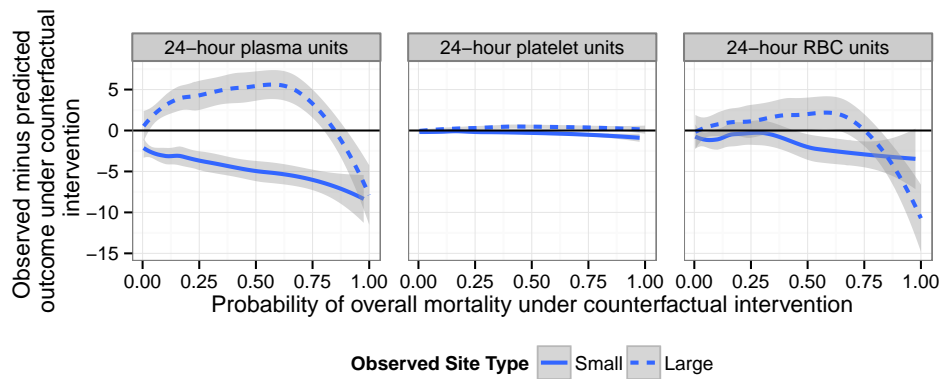
4.3.1 Effect of patient differences on 24-hour mortality

In order to identify patients who were driving the differences in the ETT, we calculated the residuals (the difference between the observed and predicted outcome) for each outcome where the prediction was obtained under the counterfactual intervention and plotted a loess curve of these residuals against the predicted probability of overall mortality, also obtained under the counterfactual intervention (Figure 4.2). The individuals with the largest residuals had counterfactual predictions that were further from their observed outcome. The most striking deviation in residuals by site type was 24-hour mortality, where the individuals with moderate predicted probabilities of mortality are responsible for the greatest deviance in the residuals, suggesting that these individuals would be most affected by the change in site type. Individuals at both ends of the spectrum (both having either relatively low or very high probabilities of death) appear to have little association with clinic size. The histogram of the predicted probabilities of overall mortality (the x-axis in the residual plots) shown in Figure 4.3 demonstrated that there are people from each site represented across the entire range of predicted probabilities, so the deviations in residuals we saw in the residual plots are driven by only a few individuals.

CHAPTER 4. QUALITY OF CARE COMPARISON



(a) Binary outcomes



(b) Continuous outcomes

Figure 4.2: Residual plots for all outcomes

CHAPTER 4. QUALITY OF CARE COMPARISON

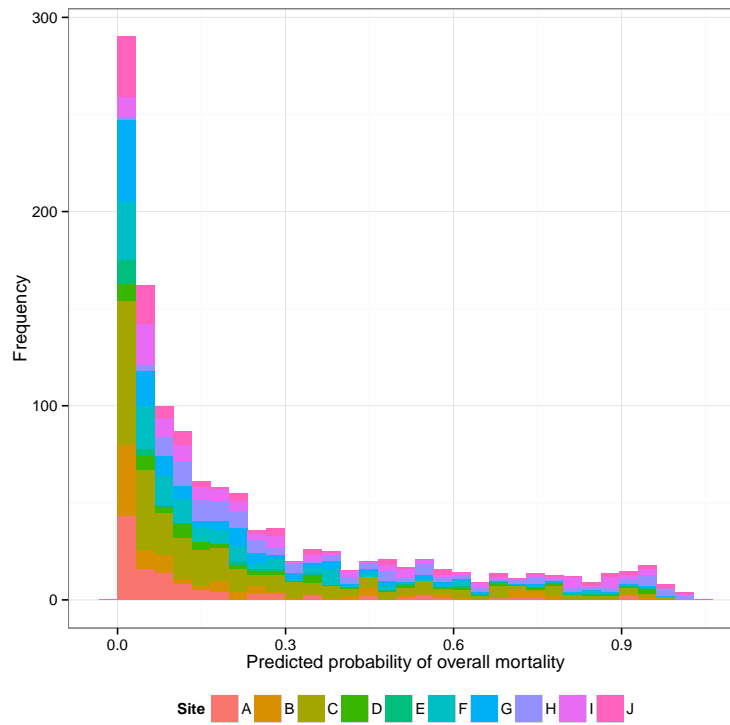


Figure 4.3: Histogram of individuals across the range of the predicted probabilities of mortality

CHAPTER 4. QUALITY OF CARE COMPARISON

Next, in order to characterize the individuals whose 24-hour mortality would have been most affected by a site type change, we compared patients on either side of the point at which the loess curves started to diverge from 0, an area highlighted in blue in 4.4. This identified 146 individuals with negative residuals, (survived beyond 24 hours) and 52 individuals with positive residuals (died before 24 hours) who would be most affected by a site change. The pairwise comparisons shown in Table 4.3 highlight the differences between these patients and those who would be less affected by a site change. Patients who were more affected by the site change had lower penetrating injury rates but higher injury severity scores and heart rates. They also had significantly longer partial thromboplastin times than those with small residuals. Additionally, their Glasgow coma scores, base deficit, and INR were all lower than the patients who were more affected by a site change. They were more likely to experience bleeding events and to be massively transfused and also had a higher probability of mortality at every time point, as shown in Table 4.4.

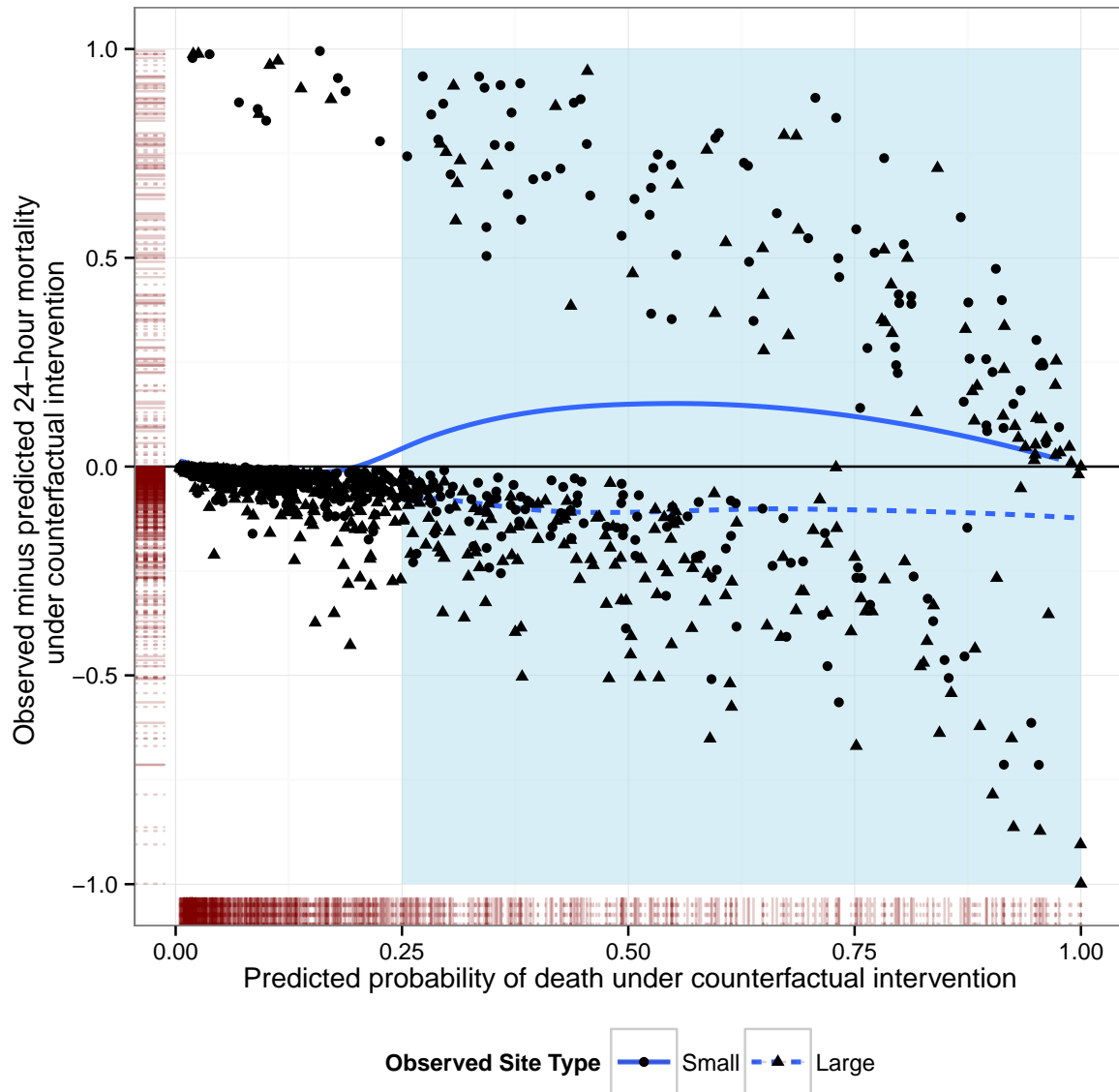


Figure 4.4: Residuals for 24-hour mortality versus predicted probabilities of overall mortality with area for subset analysis marked in light blue

CHAPTER 4. QUALITY OF CARE COMPARISON

Variable	Mean in small residuals (SD)	Mean in large resid- uals (SD)	Missing	Pvalue
Age	37.33 (17.24)	41.44 (19.58)	1	<0.001
Base deficit	-5.72 (4.77)	-8.68 (6.03)	94	<0.001
Black race	0.19 (0.39)	0.24 (0.43)	0	<0.001
BMI	27.15 (5.98)	28.03 (7.3)	177	<0.001
GCS	11.76 (4.91)	7 (5.11)	41	<0.001
Heart rate	106.32 (25.47)	106.15 (33.06)	9	<0.001
Hemoglobin	12.08 (2.15)	12.08 (2.38)	29	<0.001
Hispanic	0.24 (0.43)	0.17 (0.37)	33	<0.001
INR	1.3 (0.55)	1.98 (2.45)	55	<0.001
ISS	21.8 (12.62)	34.56 (14.32)	0	<0.001
Male	0.75 (0.43)	0.76 (0.43)	0	<0.001
Penetrating	0.45 (0.5)	0.32 (0.47)	0	<0.001
Platelets	256.17 (73.4)	220.41 (92.32)	35	<0.001
PTT	27.79 (6.8)	40.53 (25.4)	75	<0.001
Systolic BP	105.88 (29.52)	103.77 (37.36)	15	<0.001
White race	0.6 (0.49)	0.54 (0.5)	0	<0.001
Asian/Pacific Islander race	0.04 (0.19)	0.05 (0.21)	0	0.003
Anticoagulants	0.1 (0.3)	0.08 (0.27)	109	0.005
Unknown race	0.02 (0.15)	0.03 (0.17)	0	0.020

Table 4.3: Covariate comparisons in subset analysis

CHAPTER 4. QUALITY OF CARE COMPARISON

Variable	Mean in small residuals (SD)	Mean in large residuals (SD)	Pvalue
2-hour mortality	0 (0.05)	0.08 (0.27)	<0.001
6-hour mortality	0.02 (0.13)	0.19 (0.39)	<0.001
Massive transfusion (data)	0.2 (0.4)	0.38 (0.49)	<0.001
Massive transfusion (reported)	0.2 (0.4)	0.39 (0.49)	<0.001
Overall mortality	0.06 (0.25)	0.46 (0.5)	<0.001
Plasma units at 24h	6.03 (7.32)	11.41 (12.44)	<0.001
Platelet units at 24h	0.47 (1.09)	1.06 (1.75)	<0.001
RBC units at 24h	6.57 (7.24)	12.31 (13.89)	<0.001
Substantial bleeding	0.25 (0.43)	0.49 (0.5)	<0.001
Complications	0.03 (0.17)	0.05 (0.22)	0.004
Multiple organ failure	0 (0.05)	0.03 (0.16)	0.001

Table 4.4: Outcome comparisons in subset analysis

4.3.2 Propensity score matching follow-up

In addition to the subset analysis, we explored the matched data set generated by the propensity score procedure to determine whether the procedure achieved balance in the covariates. The heatmap in Figure 4.5 is a visualization of the entire matched data set where the values have been scaled to be between 0 and 1. We performed hierarchical clustering of the individuals based on their covariate values, summarized in the dendrogram on the left side of the heatmap. The color bar between the dendrogram and the heatmap indicates the site size where each individual was treated (purple corresponds small-volume sites and blue to large-volume sites). If the individuals in the matched data set were able to be clustered based on their covariates, that is, if the covariates were not balanced between the two site sizes, we would see larger blocks of color in the color bar. While there were some small clusters where patients belonged to the same site size, overall there was not much clustering of the individuals by their covariates, suggesting that we were able to achieve balance in the covariates via propensity score matching.

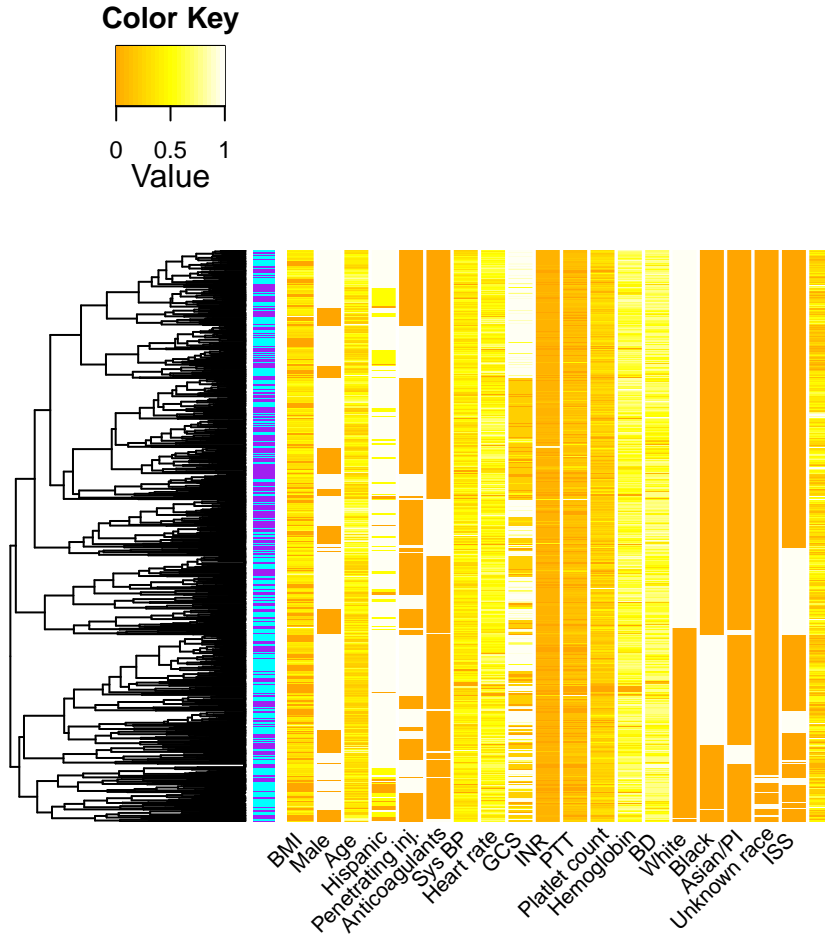


Figure 4.5: Heatmap of matched data set checking for balance in the covariates

4.3.3 Center differences effect on outcomes

Additionally, we wanted to explore which sites in particular were driving the differences in outcomes we were seeing among the large-volume patients. To do so, we built a SuperLearner predictor on every site and then predicted outcomes for all individuals using that model and subtracted the observed outcomes from those predictions. If the mean of this difference was positive, that suggested that the observed outcome was higher than expected based on the model for a particular site. If the difference was negative, that suggested that the observed outcome was lower than expected based on the model for a particular site. A heatmap of the results is shown in Figure 7, where each row is the site used to build the SuperLearner and each column corresponds to an outcome (the three blood product columns were scaled to be between 0 and 1). The bar on the left side

CHAPTER 4. QUALITY OF CARE COMPARISON

of the plot shows indicates the small-volume sites in purple and the large-volume sites in orange. Using the small-volume site J, overall there is a much larger predicted probability of mortality overall, at 6 hours, and 24 hours than was observed, suggesting that this site is partially responsible for the magnitude of the ETT for 24-hour mortality (ETT = -0.05). While sites G, F, and A also have higher predicted probabilities of mortality, the difference is most striking for site J. The large-volume site models all predicted higher probabilities of massive transfusion while most of the small-volume site models predicted lower probabilities of massive transfusion with the exception of sites C and J, suggesting that sites H, D, G, F, and A were driving the magnitude of the ETT for the massive transfusion outcomes.

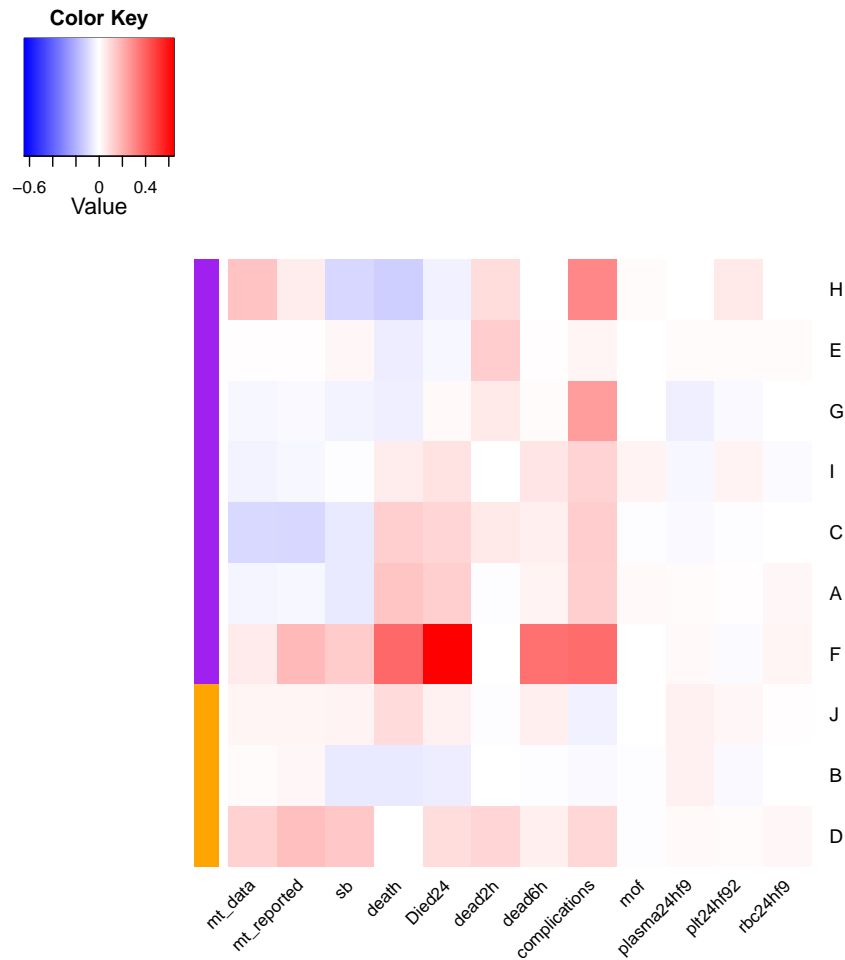


Figure 4.6: Site-specific results

4.4 Discussion

We have provided a general method for the objective comparison of the quality of care at different clinical sites using a parameter motivated by the causal inference literature and making use of machine-learning techniques for semiparametric estimation. Our analysis showed that there were significant differences in outcomes based on the size of the site at which a patient was treated. Our analysis suggests a mortality benefit for those who were treated at large-volume centers. Further analysis suggests that this difference arose primarily in patients in the middle of the injury severity scale in whom treatment has the greatest propensity to alter outcome. For the moderately injured group of patients, had they been treated at low-volume sites, (as opposed to large-volume sites) they would have had higher probabilities of mortality and incumbent lower probabilities of being massively transfused. In our subset analysis, the large-volume patients who would have been most affected by being treated at a small-volume site were indeed clotting less, bleeding more, and generally worse off than their comparison group. Uncontrolled bleeding remains the single largest contributor to preventable 24-hour mortality, making the rapid identification and treatment of hemorrhage after trauma critically important[15]. It is plausible that the centers with higher volumes more judiciously identified those who were in need of early initiation of massive transfusion, which has been shown in prior studies to improve mortality in patients with potentially survivable injuries [16-18]. In contrast, for those with nearly fatal or universally fatal injury were expected to die no matter the treatment they received.

The patients who were less affected by the site switch generally had higher Glasgow coma scores, that is, they were more conscious, had the shortest prothrombin time, highest platelets, and best base deficit. They also had lower probabilities of negative outcomes and generally required less transfusion of blood products. This data supports the findings that those who are mildly injured are far less likely to be bleeding and therefore, will do well no matter what center they go to. Additionally large-volume sites appear to be more effective at handling patients who are bleeding substantially. By estimating the ETT and adjusting for differences in patients across sites as aggressively as possible, we were able to examine the potential outcomes for patients at large-volume trauma centers, had they been treated at low-volume centers. This process highlighted the utility of the combination of machine learning and causal inference modeling in clinical research and allowed for a comparison that would have been infeasible in practice. The estimation of this comparison could be carried out using several methods. We utilized machine learning approaches in both propensity score matching estimates, as well as those based on the outcome regression models in order to avoid relying on unnecessary modeling assumptions. Overall, the direction of estimated effects from each of the four estimators was the same and the magnitudes were comparable, suggesting that each approach was

CHAPTER 4. QUALITY OF CARE COMPARISON

capturing some underlying differences between the centers. The unadjusted comparisons demonstrated how a naive approach to comparing the centers that does not adjust for patient characteristics could miss some important outcomes that did in fact differ across the sites. The prediction-based estimator that applied the prediction model built on the small sites to the large sites offered an intuitive approach to comparing the centers but did not provide any means of obtaining statistical inference. Propensity score matching is another intuitive approach that can be used to generate counterfactual outcomes for individuals by matching them to patients from other sites and we showed that, in fact, balance was achieved in the covariates with a multivariate analysis of the matched data set. These more commonly used approaches provided similar results to the TMLE, which is an estimator designed to achieve the optimal bias-variance tradeoff and that has other desirable statistical properties [12,19]. Thus, we believe that it offers an important alternative method for estimation and advocate for its use in the future. In addition, we were able to perform follow-up comparisons that reinforced the TMLE findings with empirical evidence.

Our results suggested that even among the top-rated trauma centers involved in PROMMTT, there is variability in the quality of care with respect to a variety of clinical outcomes. Such comparisons have the potential to be a useful tool for identifying areas of improvement for individual centers and standardizing care across centers. We acknowledge that these estimators are more complex than other approaches, but they allow for the estimation of clinically meaningful parameters of interest that have causal interpretations under some assumptions. This analysis purposely did not utilize treatment information to avoid bias as a result of sicker patients receiving more treatment (treatment by indication). Thus, we cannot address questions of treatment efficacy. Additionally, it is possible that we did not include all confounders of the effect of site size on the outcomes of interest. However, this approach allows for adjustment by a large covariate set, so expansion of the confounders is possible. Another possible expansion of this work would be to include other centers in the comparison and use this method to validate current trauma center rankings. This is by no means the only approach to compare the quality of trauma care at different hospitals. Indeed, additional comparisons could examine more closely what factors in particular are responsible for the differences in site effectiveness. Given the proliferation of interest in comparative effectiveness, familiarity with these methods will allow for a better understanding of factors influencing trauma patient care and provide directions towards other areas for improvement.

Conclusions

We have explored three large questions related to the analysis of critical care data. First, we examined the semiparametric prediction of clinical outcomes of interest using covariates measured in the emergency department. Not only can we predict these outcomes well, but we also showed the potential for improving upon current scoring systems used to identify high-risk patients. Second, we used causal inference to motivate a variable importance measure with a clinically meaningful interpretation and applied it to gene expression data in order to examine how important genes varied within and across time in trauma patients. Finally, we derived an objective measure to compare the quality of care at different hospitals which suggests that even at the top tier of current trauma center rankings, there is room for improvement.

These analyses highlight the importance of a principled approach when answering questions of interest in order to maintain transparency regarding required assumptions, derive statistical parameters that address the question of interest, and make the most efficient use of the observed data. Our approach to the prediction problem utilized SuperLearning to allow for model flexibility, generalizability, and honest comparisons of predictive performance. We showed that causal inference can motivate parameters that may not have a causal interpretation but are still interesting measures of variable importance in the analysis of the gene expression data. We also made clear the assumptions required to give a causal interpretation to the comparison of quality of care at different hospitals and motivated a clinically meaningful parameter. These approaches improve upon current practice in the analysis of critical care data and opened up many areas for further research. The improvement of individualized patient care as well as the standardization of care quality across hospitals can benefit from a clearer understanding of the mechanisms underlying response to injury as well as the efficacy of treatment.

Bibliography

- [1] Trunkey DD (2000) History and development of trauma care in the united states. *Clinical orthopaedics and related research* 374: 36–46.
- [2] Wangensteen OH, Wangensteen SD (1978) The rise of surgery: from empiric craft to scientific discipline. Dawson.
- [3] Mullins RJ (1999) A historical perspective of trauma system development in the united states. *Journal of Trauma-Injury, Infection, and Critical Care* 47: S8–S14.
- [4] Demetriades D, Martin M, Salim A, Rhee P, Brown C, et al. (2006) Relationship between american college of surgeons trauma center designation and mortality in patients with severe trauma (injury severity score \geq 15). *Journal of the American College of Surgeons* 202: 212–215.
- [5] Baker SP, o’Neill B, Haddon Jr W, Long WB (1974) The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma-Injury, Infection, and Critical Care* 14: 187–196.
- [6] Palmer C (2007) Major trauma and the injury severity score-where should we set the bar? In: *Annual Proceedings/Association for the Advancement of Automotive Medicine*. Association for the Advancement of Automotive Medicine, volume 51, p. 13.
- [7] Nunez TC, Voskresensky IV, Dossett LA, Shinall R, Dutton WD, et al. (2009) Early prediction of massive transfusion in trauma: simple as abc (assessment of blood consumption)? *Journal of Trauma-Injury, Infection, and Critical Care* 66: 346–352.
- [8] Pearl J (2000) *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press.
- [9] van der Laan M, Rose S (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer.

BIBLIOGRAPHY

- [10] Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688.
- [11] Holcomb JB, Fox EE, Wade CE, Group PS, et al. (2013) The prospective observational multicenter major trauma transfusion (prommtt) study. *Journal of Trauma-Injury, Infection, and Critical Care* 75: S1–S2.
- [12] Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, et al. (2005) Application of genome-wide expression analysis to human health and disease. *Proceedings of the National Academy of Sciences of the United States of America* 102: 4801–4806.
- [13] Holcomb JB, del Junco DJ, Fox EE, Wade CE, Cohen MJ, et al. (2013) The prospective, observational, multicenter, major trauma transfusion (prommtt) study: comparative effectiveness of a time-varying treatment with competing risks. *JAMA surgery* 148: 127–136.
- [14] Eastman AB (2010) Wherever the dart lands: toward the ideal trauma system. *Journal of the American College of Surgeons* 211: 153–168.
- [15] Shackford SR, Mackersie RC, Holbrook TL, Davis JW, Hollingsworth-Fridlund P, et al. (1993) The epidemiology of traumatic death: a population-based analysis. *Archives of Surgery* 128: 571–575.
- [16] Counts R, Haisch C, Simon T, Maxwell N, Heimbach D, et al. (1979) Hemostasis in massively transfused trauma patients. *Annals of surgery* 190: 91.
- [17] Ledgerwood AM, Lucas CE (2003) A review of studies on the effects of hemorrhagic shock and resuscitation on the coagulation profile. *Journal of Trauma-Injury, Infection, and Critical Care* 54: S68–S74.
- [18] Holcomb JB, Jenkins D, Rhee P, Johannigman J, Mahoney P, et al. (2007) Damage control resuscitation: directly addressing the early coagulopathy of trauma. *Journal of Trauma-Injury, Infection, and Critical Care* 62: 307–310.
- [19] Newell MA, Bard MR, Goettler CE, Toschlog EA, Schenarts PJ, et al. (2007) Body mass index and outcomes in critically injured blunt trauma patients: weighing the impact. *Journal of the American College of Surgeons* 204: 1056–1061.
- [20] Hirschberg R, Weiss D, Zafonte R (2008) Traumatic brain injury and gender: what is known and what is not. *Future Medicine* .
- [21] Champion HR, Copes WS, Sacco WJ, Lawnick MM, Keast SL, et al. (1990) The major trauma outcome study: establishing national norms for trauma care. *Journal of Trauma-Injury, Infection, and Critical Care* 30: 1356–1365.

BIBLIOGRAPHY

- [22] Egede LE, Dismuke C, Echols C (2012) Racial/ethnic disparities in mortality risk among us veterans with traumatic brain injury. *American Journal of Public Health* 102: S266–S271.
- [23] Glance LG, Osler TM, Mukamel DB, Meredith JW, Li Y, et al. (2013) Trends in racial disparities for injured patients admitted to trauma centers. *Health Services Research* 48: 1684–1703.
- [24] Britt L (2012) *Acute Care Surgery*. Ph.D. thesis, Springer.
- [25] Di Bartolomeo S, Marino M, Valent F (2014) Effects of anticoagulant and antiplatelet drugs on the risk for hospital admission for traumatic injuries: A case-control and population-based study. *Journal of Trauma and Acute Care Surgery* : 437–442.
- [26] Teasdale G, Jennett B (1974) Assessment of coma and impaired consciousness: a practical scale. *The Lancet* 304: 81–84.
- [27] Schreiber MA, Perkins J, Kiraly L, Underwood S, Wade C, et al. (2007) Early predictors of massive transfusion in combat casualties. *Journal of the American College of Surgeons* 205: 541–545.
- [28] Robert A, Chazouilleres O (1996) Prothrombin time in liver failure: time, ratio, activity percentage, or international normalized ratio. *Hepatology* 24: 1392–1394.
- [29] Giles C (1981) The platelet count and mean platelet volume. *British Journal of Haematology* 48: 31–37.
- [30] Nijboer JM, van der Horst IC, Hendriks HG, ten Duis HJ, Nijsten MW (2007) Myth or reality: hematocrit and hemoglobin differ in trauma. *Journal of Trauma-Injury, Infection, and Critical Care* 62: 1310–1312.
- [31] Davis JW, Shackford SR, Mackerse rC, Hoyt DB (1988) Base deficit as a guide to volume resuscitation. *Journal of Trauma-Injury, Infection, and Critical Care* 28: 1464–1467.
- [32] Davis JW, Parks SN, Kaups KL, Gladen HE, O’Donnell-Nicol S (1996) Admission base deficit predicts transfusion requirements and risk of complications. *Journal of Trauma-Injury, Infection, and Critical Care* 41: 769–774.
- [33] Heller K, Reardon R, Joing S (2007) Ultrasound use in trauma: the fast exam. *Academic Emergency Medicine* 14: 525–525.
- [34] Eiseman B, Beart R, Norton L, et al. (1977) Multiple organ failure. *Surgery, gynecology & obstetrics* 144: 323.

BIBLIOGRAPHY

- [35] Lenz A, Franklin GA, Cheadle WG (2007) Systemic inflammation after trauma. *Injury* 38: 1336–1345.
- [36] Klein M, Silver G, Gamelli R, Gibran N, Herndon D, et al. (2006) An overview of the multicenter study of the genomic and proteomic response to burn injury. *Journal of Burn Care and Research* 27: 448–451.
- [37] Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, et al. (2005) Application of genome-wide expression analysis to human health and disease. *Proceedings of the National Academy of Sciences of the United States of America* 102: 4801–4806.
- [38] Brownstein B, Logvinenko T, Lederer J, Cobb J, Hubbard W, et al. (2006) Commonality and differences in leukocyte gene expression patterns among three models of inflammation and injury. *Physiological Genomics* 24: 298–309.
- [39] Warren H, Elson CM, L HD, Schoenfeld DA, Cobb JP, et al. (2009) A genomic score prognostic of outcome in trauma patients. *Molecular Medicine* 15: 220–227.
- [40] Xiao W, Mindrinos MN, Seok J, Cuschieri J, Cuenca AG, et al. (2011) A genomic storm in critically injured humans. *The Journal of experimental medicine* 208: 2581–2590.
- [41] Hastie T, Tibshirani R, Friedman JJH (2001) *The elements of statistical learning*, volume 1. Springer New York.
- [42] Hawkins DM (2004) The problem of overfitting. *Journal of Chemical Information and Computer Sciences* 44: 1–12.
- [43] Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys* 4: 40–79.
- [44] Van Der Laan MJ, Dudoit S (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *bepress* .
- [45] van der Laan MJ, Dudoit S, Keles S (2003) Asymptotic optimality of likelihood based cross-validation. *bepress* .
- [46] Vaart AWvd, Dudoit S, Laan MJvd (2006) Oracle inequalities for multi-fold cross validation. *Statistics & Decisions* 24: 351–371.
- [47] Breiman L (2001) Random forests. *Machine Learning* 45: 5–32.
- [48] Polley EC, van der Laan MJ (2010) Super learner in prediction. UC Berkeley Division of Biostatistics Working Paper Series 266.

BIBLIOGRAPHY

- [49] van der Laan M, Polley E, Hubbard A (2007) Superlearner. *Statistical Applications in Genetics and Molecular Biology* 6: 1-21.
- [50] Horton NJ, Kleinman KP (2007) Much ado about nothing. *The American Statistician* 61.
- [51] Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* : 1360–1383.
- [52] Friedman JH (1991) Multivariate adaptive regression splines. *The annals of statistics* : 1–67.
- [53] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and regression trees*. Wadsworth & Brooks.
- [54] Breiman L (2001) Random forests. *Machine learning* 45: 5–32.
- [55] Ripley B (1996) *Pattern Recognition and Neural Networks*. Cambridge.
- [56] Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3: e161.
- [57] Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97: 77–87.
- [58] Bradley AP (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition* 30: 1145–1159.
- [59] Ledell E, L PM, van der Laan MJ (2012) Computational efficient confidence intervals for cross-validated area under the roc curve estimates. UC Berkeley Division of Biostatistics Working Paper Series 304.
- [60] Olden J, Jackson D (2004) Illuminating ?the black box?: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modeling* 154: 135–150.
- [61] Olden J, Joy M, Death R (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modeling* 178: 389–397.
- [62] Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources, and a solution. *BMC Bioinformatics* 8.

BIBLIOGRAPHY

- [63] Olden JD, Jackson DA (2002) Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154: 135–150.
- [64] Dudoit S, J F (2003) Classification in microarray experiments. In: *Statistical Analysis of Gene Expression Microarray Data*, Harvard School of Public Health.
- [65] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* 344: 539–548.
- [66] Petersen ML, Porter K, Gruber S, Wang Y, van der Laan MJ (2010) Diagnosing and responding to violations in the positivity assumption. UC Berkeley Division of Biostatistics Working Paper Series 269.
- [67] van der Laan MJ (2010) Targeted maximum likelihood based causal inference: Part i. *The International Journal of Biostatistics* 6.
- [68] Van der Laan MJ, Robins JM (2003) *Unified methods for censored longitudinal data and causality*. Springer.
- [69] van der Laan MJ, Gruber S (2010) Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics* 6.
- [70] Hampel FR (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69: 383–393.
- [71] J van der Laan M, Pollard KS (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference* 117: 275–303.
- [72] Brohi K, Cohen MJ, Davenport RA (2007) Acute coagulopathy of trauma: mechanism, identification and effect. *Current opinion in critical care* 13: 680–685.
- [73] van der Laan MJ, Gruber S (2011) Targeted minimum loss based estimation of an intervention specific mean outcome. University of California, Berkeley, Division of Biostatistics Working Paper Series 290.
- [74] Hurtuk M, R Lawrence Reed I, Esposito TJ, Davis KA, Luchette FA (2006) Trauma surgeons practice what they preach: the ntdb story on solid organ injury management. *Journal of Trauma-Injury, Infection, and Critical Care* 61: 243–255.
- [75] Rogers FB, Shackford SR, Hoyt DB, Camp L, Osler TM, et al. (1997) Trauma deaths in a mature urban vs rural trauma system: a comparison. *Archives of Surgery* 132: 376.

BIBLIOGRAPHY

- [76] Callcut RA, Cotton BA, Muskat P, Fox EE, Wade CE, et al. (2013) Defining when to initiate massive transfusion [mt]: A validation study of individual massive transfusion triggers in promtt patients. *The Journal of Trauma and Acute Care Surgery* 74: 59.
- [77] Malone DL, Hess JR, Fingerhut A (2006) Massive transfusion practices around the globe and a suggestion for a common massive transfusion protocol. *Journal of Trauma-Injury, Infection, and Critical Care* 60: S91–S96.
- [78] Nathens AB, Xiong W, Shafi S (2008) Ranking of trauma center performance: the bare essentials. *Journal of Trauma-Injury, Infection, and Critical Care* 65: 628–635.
- [79] Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling* 7: 1393–1512.
- [80] Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software* 42: 1–52.
- [81] Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.
- [82] Hastie TJ, Tibshirani RJ (1990) *Generalized additive models*, volume 43. CRC Press.