Yet Another Local Learner (YALL): a localized machine learning algorithm with applications to precision medicine

by

Sara Elizabeth Moore

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 John M. Colford

Fall 2023

Yet Another Local Learner (YALL): a localized machine learning algorithm with applications to precision medicine

Abstract

Yet Another Local Learner (YALL): a localized machine learning algorithm with applications to precision medicine

by

Sara Elizabeth Moore

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Chair

We sought to develop an improved decision algorithm for applications in precision medicine. This algorithm should perform as well or better than existing prognostic scores, be interpretable and parsimonious in the predictors it uses for an individual patient, and be able to perform competitively against best-in-class existing ensemble learning algorithms. The algorithm must be flexible enough to predict well in heterogeneous patient populations, where we might expect the covariates that are related to the outcome of interest to be different for different types of patients. To this end, a new supervised classification method is proposed which performs both dimension and instance reduction data-adaptively to hone in on only the most relevant information for a given patient.

To Jenny

For your encouragement, understanding, and unrelenting (but not unending) patience.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my sincere, deep gratitude to my advisor and dissertation chair, Alan Hubbard, without whom this project could not have come to fruition. I am grateful for his guidance and hold great respect for his humility, patience, and ability to demystify even the most complex concepts.

I am grateful to Mark van der Laan for his insightful feedback on this work as well as his commitment to challenging both his students and the statistical status quo. Many thanks also to Jack Colford for his review of this manuscript and Lexin Li for his feedback during my qualifying exam. In addition, I would like to recognize the great impact Maureen Lahiff's mentorship and compassion had on me during my time at UC Berkeley.

I also owe a debt of gratitude to the UCSF PROPPR research team, whose subject-matter knowledge and fellowship were invaluable during this project. This team included Lucy Kornblith, Amanda Conroy, Mary Nelson, Rachael Callcut, Ben Howard, and Mitchell Cohen.

The camaraderie of Berkeley Biostat students made this experience infinitely more enjoyable and fulfilling. I have the utmost respect and eternal gratitude for so many people I met during this program, including Anna Decker, Erin LeDell, Marla Johnson, Molly Davies, Sam Lendle, Oleg Sofrygin, Monika Izano, Robin Mejia, Lucia Petito, Jeremy Coyle, Laura Balzer, Alex Luedtke, Luca Pozzi, Irene Headen, and Mary Combs.

Special thanks go to Mike Elashoff, whose eleventh hour encouragement, empathy, and offers to review this manuscript were instrumental in its completion.

Thank you to my Dad for always believing I could do anything and never letting me forget it.

Last but never least, my eternal gratitude goes to Jenny Trull, my long-suffering partner who encouraged and supported this endeavor from our very first date onward.

Chapter 1

Introduction & Literature Review

Trauma Scoring

Traumatic Injury

Since 1976, the United States has spent more on healthcare per capita than any other country in the world [35]. This trend is not subsiding; year after year, the U.S. consistently outspends nearly all other countries in this arena. For example, in 2017, the U.S. was second only to the small island nation of Tuvalu in healthcare spending as a proportion of gross domestic product (GDP) [1]. Trauma-related disorders comprised the eighth largest medical expenditure in the U.S. in that same year, totaling over $87 billion. Among the subset of patients aged 18 to 44 years, trauma-related medical expenditures ranked third, surpassed only by ("normal") preganancy/birth and mental disorders (ranked first and second, respectively) [19].

Also in 2017, traumatic injury was the primary cause of 243,003 deaths in the U.S. [48]. U.S. mortality data from 2017 provided by the CDC [48], broken down into seven broad causes and fifteen ranges of age at death, is displayed in Figure 1.1. The width of each bar indicates the relative proportion of individuals who died within the respective age range. (Individuals for whom age was 'not stated' – 91 of 2,818,862 – are not included in this visualization.) Traumatic injury is defined broadly here as 'external causes' of death, which includes such specific categories as 'motor vehicle accidents,' 'suicide,' and 'homicide,' but also includes causes not typically thought of as injuries, such as 'poisoning.'

Both the wide-angle view in Figure 1.1a and the zoomed-in view in Figure 1.1b demonstrate that children and adults under the age of 45 are disproportionately susceptible to death due to traumatic injury. In 2017, as in many years before it, traumatic injury was the leading individual cause of death in Americans between

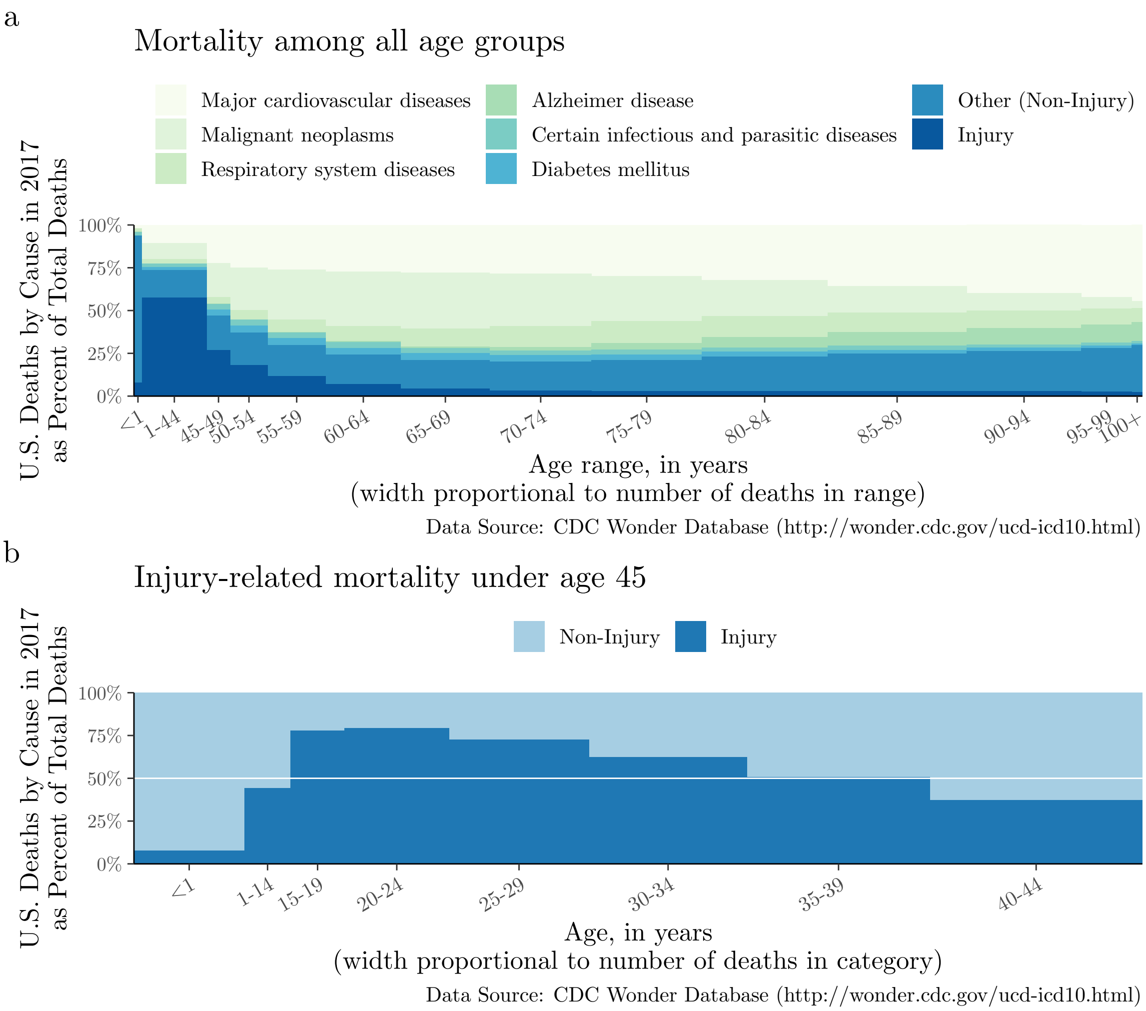

Figure 1.1: Causes of Death in the U.S., 2017

1 and 44 years of age, and was responsible for more deaths of individuals in the U.S. between the ages of 15 and 34 than all types of natural causes, combined (see Figure 1.1b). This threat is particularly pronounced for those between ages 15 and 24, for whom 79% of deaths were caused by traumatic injury [48]. For comparison, unintentional injuries were the third-leading cause of death for Americans ages 45 to 64 in that same year [49].

Despite the higher cost of healthcare in the United States and resources available to U.S. trauma centers not available elsewhere in the world, the traumatic injury mortality rate in the U.S. is no better than the worldwide average. Taking all classes of fatal injury into consideration, whether accidental, self-inflicted, or due to assault, injury was responsible for 9% of deaths worldwide in 2014 [2] and 9% in the U.S. three years later [48].

However, there is evidence to suggest that some of these deaths could be prevented via improved trauma care. Severely injured patients tend to have a better chance of surviving their injuries when treated at a level I trauma center than at non-trauma center hospitals, even after adjustment for observed confounders [73]. This pattern holds when comparing survival rates of the severely injured outside the U.S. at trauma centers versus non-trauma center hospitals [18]. Assuming a level I trauma center in an urban area of an industrialized nation is the standard-bearer of modern trauma care, patients who succumb to severe injury at these centers could be characterized to better elucidate situations where modern medical interventions could not have changed the outcome. Conversely, using data to better understand which traumatically injured patients' deaths are preventable would inform what improvements in care would be necessary to decrease injury-related mortality rates.

For example, in 1998, one study recommended that, since low-income developing nations without basic emergency medical services (EMS) experience relatively high mortality rates of seriously injured adults prior to reaching a hospital, the most cost-effective reductions in mortality could be realized by improving telecommunications infrastructure and providing basic first aid training to existing emergency response organizations. Overall trauma mortality rates are lower in middle-income developing nations with basic EMS, but still higher than the rates observed in industrialized nations. Relative to poorer nations, these middle-income countries see fewer prehospital deaths but more deaths in the emergency department (ED), shifting the realm in which care must be improved [77].

In mature trauma systems in developed nations, evidence indicates that the quality of care could still be improved. Elderly patients, those who suffer from dementia or are taking anticoagulants have been shown to be at higher risk of undertriage and thereby mortality after major trauma in the U.S. [99, 10]. The American College of Surgeons – Committee on Trauma (ACS-COT) recommends an undertriage (or false negative) rate of no higher than 5% [28] but, in practice, the overall rate may be much higher. One recent estimate places the rate at 34% nationally [121].

Overtriage under high patient load, or, more generally, an overwhelmed trauma system – one exceeding surge capacity where the short-term demand outstrips the local resources, typically as a result of a mass casualty incident – can also lead to higher mortality rates in severely injured patients [5]. Precise overtriage (false positive) rates for recent mass casualty events are in short supply, but the 2005 London Bombings were estimated to have a field overtriage rate of 64% [7]. In contrast, ACS-COT suggests that an acceptable overtriage rate falls between 25% and 35% [28].

Montmany et al. [79] reviewed 115 trauma fatalities in Spain over nearly nine years and deemed 16.5% "preventable or potentially preventable." They concluded that the majority of these deaths were due to human "rule-based" error where medical professionals failed "to observe the established instructions or protocols." Recent studies from the U.S. using the same standard to classify errors in treatment came to mixed conclusions. In one, rule-based errors were again the largest contributor [54], while in another, the majority of errors made in preventable trauma deaths were due to lack of medical knowledge [117].

How many deaths could be prevented via improved trauma care? By improving trauma care at only "low-performing hospitals" to that of average performers, a crude estimate puts the number of civilian lives that could be saved at over 4,000 annually in the U.S. alone. If all low or average performing U.S. hospitals could improve to reach the performance level of high performers, the same estimation method predicts that nearly 20,000 lives could be saved each year [46].

These figures do not account for potentially preventable casualties internationally, civilian or military. For example, of 4,596 U.S. military combat fatalities during Operations Iraqi Freedom and Enduring Freedom, 976 (or approximately 21% of) deaths which occurred over that nearly ten year period were potentially survivable [34]. Moreover, preventable trauma mortality rates are not the only figures by which trauma care quality can or should be judged – a patient whose life was saved may still suffer a permanent disability after a traumatic injury. Death or disability can result in lost wages for individuals or families, another hidden cost of traumatic injury. Improved care could yield lower the rate of lifelong disability resulting from critical injuries.

Regardless of the maturity of the trauma system at play, the appropriate and efficient allocation of resources can save lives. This is particularly salient for resource-limited settings and mass casualty incidents, but novel improvements even have the potential to further improve the current "gold standard" trauma mortality rate in prosperous nations.

Precision medicine

In recent years, many novel improvements in patient care have fallen under the umbrella of *precision medicine* or *personalized medicine*, both of which rapidly became buzzwords early in the twenty-first century. These concepts serve as major motivators for the work presented in this manuscript, warranting a concise definition of both terms (provided below).

In 2011, the National Research Council (NRC) produced a report called "Toward Precision Medicine," and in it, they defined *precision medicine* as "the tailoring of medical treatment to the individual characteristics of each patient." The authors went on to state the following:

> "Precision medicine does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not." [29]

The report clarified that the term *personalized medicine* is also used to convey the same meaning as *precision medicine*, but *personalized medicine* can be misinterpreted as implying that unique treatments will be designed for each individual patient. Therefore, they recommend use of the phrase *precision medicine*. They also note that "precision" should not be interpreted strictly in the statistical sense (i.e. the inverse of the variability of a measurement or estimator), but is instead intended to imply both accuracy and precision (i.e. both unbiasedness and reliability).

From a statistical standpoint, the implementation guidance derivable from this definition is that the development of a *precision medicine* treatment would involve separating individuals with known medical outcomes into small subpopulations based on their clinical history and other relevant attributes. After exploring which treatment regimens led to the best clinical outcomes for patients in each group, the treatment for a new patient could be informed by using her characteristics to associate her with the group to which she is most similar. In this way, a patient's treatment plan is customized, broadly, to her individual clinically-relevant attributes. In addition, when guided by this *precision medicine* approach, clinicians may be able to utilize resources more efficiently than they otherwise would have. In a trauma care setting, those resources can include blood products, medications, labs, imaging, trauma surgery, and, perhaps most importantly, the treating physician's time spent monitoring and developing a treatment plan for the patient.

Existing Trauma Scoring Systems

Physicians who attend to critically injured patients often have limited time and information to make life-saving treatment decisions. Efficiency, both in determining the small subpopulation into which a patient 'fits' and in the allocation of appropriate resources to that patient, is therefore of particular importance in emergency and trauma care. These decisions are often based upon only a handful of variables which can either be chosen based on clinical experience and combined qualitatively, or chosen via and summarized using a regression-based trauma severity score formula. Many modern trauma scoring systems have chosen the latter approach [122, 24, 120, 105, 115, 91]. Because these summary scores are simple by design, they must ignore the majority of (potentially relevant) information that would likely be available for the typical trauma patient and instead use the same information for every patient. In other words, use of these scoring systems in practice rests on the assumption that sufficient meaningful variability and predictive power can be captured by the same few variables across all patients. Given that each new trauma score emerging from the literature purports competitive performance via utilization of a novel combination of covariates, this may be a dubious assumption unless serious redundancy exists between the many available predictors.

Clinical Outcomes

Of the dozens of trauma, intensive care unit (ICU), organ dysfunction, and organ- or disease-specific prognostic scores in the literature, many – but not all – were designed to predict mortality risk, whether by 24 hours, 30 days, or over some other fixed time period pre- or post-discharge. However, even scoring systems which don't target mortality are used to that end, a practice which can result in diminished predictive performance for this clinical outcome. For example, the Glasgow Coma Scale (GCS) was designed to quantify a patient's level of consciousness, and although it may successfully serve as an indicator of physiological well-being, was never statistically "targeted" for any outcome (i.e. with coefficient estimates) [111]. However, the GCS has been used in comparison studies as an independent predictor of in-hospital mortality. For instance, in Guzzo et al. [43], it performed respectably but not competitively with multi-variable, mortality-targeted scoring systems.

Although trauma scores tend to target mortality risk more than any other clinical outcome, death is not the only outcome of interest in trauma care. Other commonly used scoring systems such as Assessment of Blood Consumption (ABC) [93] and Trauma-Associated Severe Hemorrhage (TASH) [123] were designed to predict, with high sensitivity, a patient's need for massive transfusion (MT) (of blood products).

Other popular scoring systems predict risk of a specific illness [116], assess injury to or predict failure of a specific organ [86, 84, 80, 81, 85, 82, 83], or assess injury to a particular part of the body [27], while still other scores are designed to predict risk of general organ failure [74, 40]. Ideally, additional clinical outcomes could be targeted easily as needed, but in practice, development of a new scoring system is lengthy process.

Specific Scores

Table 1.1 outlines some of the more popular modern trauma and ICU scoring systems used for mortality prediction in the general adult trauma patient population. Scores designed to assess specific illnesses or organ failure (such as the GCS [111], Multiple Organ Dysfunction Score (MODS) [74], Logistic Organ Dysfunction System (LODS) [40], and Sepsis-related or Sequential Organ Failure Assessment (SOFA) [116]) are omitted from this comparison.

Table 1.1: A Comparison of Common Modern Scoring Systems used to Predict Mortality in Trauma Patients.

Abbreviation	Name	Authors	First Published	Injury Scoring*	Weights
ISS	Injury Severity Score	Baker et al.[9,8]	1974	AIS	No
TRISS	Trauma ISS	Champion et al.[23]	1981	AIS; mech.	Yes
APACHE II	Acute Physiology and Chronic Health Evaluation II	Knaus et al.[59]	1985		No
RTS/T-RTS	(Triage -) Revised Trauma Score	Champion et al.[22,20]	1989		Yes/No
ASCOT	A Severity Characterization of Trauma	Champion et al.[21]	1990	AIS; mech.	Yes
SIRS	Systemic Inflammatory Response Syndrome score	Bone et al.[14,92]	1992		No
SAPS II	Simplified Acute Physiology Score II	Gall et al.[38]	1993		No†
ICISS	International Classification of Disease (ICD)-based ISS	Osler et al.[96]	1996	ICD	Yes
NISS	New ISS	Osler et al.[94]	1997	AIS	No
HARM	Harborview Assessment for Risk of Mortality	West et al.[118]	2000	ICD	Yes
KTS	Kampala Trauma Score	Kobusingye and Lett[61]	2000	count	No
PTS	Physiologic Trauma Score	Kuhls et al.[64]	2002		Yes
SAPS 3	Simplified Acute Physiology Score 3	Moreno et al.[90]	2005		No†
TMPM	Trauma Mortality Prediction Model	Osler et al.; Glance et al.[95,42]	2008	AIS/ICD	Yes
RISC/RISC II	Revised Injury Severity Classification (II)	Lefering; Lefering et al.[68,69]	2009	AIS	Yes
MGAP	Mechanism, GCS, Age, and Pressure	Sartorius et al.[103]	2010	mech.	Yes
GAP	GCS, Age, and Pressure	Kondo et al.[62]	2011		Yes
mREMS	modified REMS	Miller et al.[76]	2017		No
NTS/T-NTS	(Triage -) New Trauma Score	Jeong et al.[55]	2017		Yes/No

* AIS = Abbreviated Injury Scale; ICD = International Classification of Disease Clinical Modification codes; count = Number of injuries (none, one, or multiple); mech. = Mechanism of Injury (blunt or penetrating)

† Regression coefficient estimates required to use score to estimate probability of death

Also omitted are some of the many iterations on a few of the more popular scores. For example, only one of the four variations of the Acute Physiology and Chronic Health Evaluation (APACHE) score and two of the three varieties of the Simplified Acute Physiology Score (SAPS) are included. These omissions are made for the sake of both brevity and viability – APACHE II requires substantially fewer variables than other APACHE scores (only 17 [59] versus 26 [60], 34 [58], or 142 [125]), making its use in the ED more practical, while the original SAPS [39], shown to have significantly worse predictive performance than its successor [38], has largely fallen out of use in comparison studies.

Predictors

Typical covariates employed in trauma scores are anatomical, physiological, and/or demographic characteristics of the patient. For example, anatomical descriptors could include the mechanism of injury (whether the injury was blunt or penetrating), number of injuries, AIS scores, or summary of ICD codes. Temperature, heart rate, respiratory rate, blood pressure, GCS (or its three component scores), and base deficit/excess are some of the more common physiological features used. Age and/or gender are among the most commonly included demographic characteristics of the patient, if such information is available.

The components of most scores are chosen subjectively via "clinical knowledge," though some scores' development began with large lists of predictors which were reduced via stepwise regression (as in HARM and PTS) or more complex regression-based techniques (as in TMPM). Selected variables are nearly always discretized, with cutoffs often also chosen via "clinical knowledge" but occasionally via exploratory statistical techniques such as Locally Weighted Scatterplot Smoothing (LOWESS). The typical argument posed for discretization is ease of use in the field – an argument that some researchers bolster by reporting the absence of a statistically significant difference in predictive performance between models using continuous versus binned versions of the same covariates.

However, discretization of continuous predictors as a practice is based upon untenable assumptions, uses additional degrees of freedom, results in residual confounding, and induces loss of information, power, and precision [45, 101]. Under certain conditions, this approach can increase both bias in the coefficient estimates and the probability of erroneously rejecting the null hypothesis for tests of statistical significance performed on coefficient estimates resulting from multiple regression [6, 101]. This is important in the current context because discretized predictors are sometimes tested for statistical significance during scoring system development and included in (or excluded from) the score based on the results of those tests.

In addition, cutpoints are often uncertain and inconsistent across studies regardless of whether they are set a priori or informed by the data at hand. As an alternative to discretizing or binning, it has been suggested that analysts utilize more flexible methods that can accommodate non-linear relationships between predictor(s) and outcomes [45, 101].

Score Development/Statistical Considerations

Scores for which Table 1.1 indicates "Yes" in the "Weights" column require exact coefficient estimates to be used in their calculation. Most scores requiring weights were developed using logistic regression, typically with an outcome of in-hospital mortality over some fixed time period. The "ease of use" argument made for discretization of variables in this context is particularly puzzling, since a calculator would still be required to use regression coefficients and apply a logit function to the result. In contrast, scores without weights were intentionally developed as such for ease of use in the field. There are also hybrid approaches; SAPS II and 3 assign "points" to (ordinally) discretized "bins" of each variable and were developed using regression (where points for each bin were based on rounded regression coefficient estimates). Notably, a separate regression equation – using exact, not rounded, coefficient estimates – is necessary to "convert" a SAPS II or 3 score into an estimated probability of death [38, 90].

Generally speaking, the more recent the scoring system, the more complex the statistical modeling technique used to formulate it. The exceptions to this rule are newer scores which build upon older scores, of which there are many examples. An existing scoring system can be published as a retooled, "new" scoring system after only minor modification. For instance, the New Trauma and Injury Severity Score (NTRISS) is the result of replacing the ISS with the NISS in the TRISS [31]. The Base Excess Injury Severity Scale (BISS) is estimated by switching out the RTS in the TRISS with "the absolute difference of the base deficit from its normal range" (-2 to 2) [63]. BISS was followed by BISSGCS, which added (binned) GCS to the BISS regression equation [57], and the Base Excess New Injury Severity Scale (BNISS), wherein the ISS was ousted from the equation in favor of the NISS [66].

Another common situation where new scores build upon old are simple updates to coefficient estimates of existing scoring systems using a different patient database, be it newer, larger, or the result of sampling from a different population. For example, the seminal paper establishing covariate weights for the TRISS used patient data from the Major Trauma Outcome Study (MTOS) for coefficient estimation [15]. In the thirty years since, many versions of covariate weights for the TRISS have been published, each based on a slightly different data source or structure. Perhaps the most widely used set of "alternative" TRISS weights are those estimated in 2009 using the National Trauma Data Bank (NTDB) [104]. However, the dissension surrounding TRISS weights is perhaps best exemplified by a single issue of The Journal of Trauma published in 1995, wherein Champion, Sacco, and Copes [20] provided updated TRISS weights based on AIS-90 injury coding (whereas the original TRISS weights were based on injury data scored using the older AIS-85 system), Hannan et al. [44] updated TRISS weights using the Institute for Trauma and Emergency Care (ITEC) patient database, and Jones, Redmond, and Templeton [56] provided an argument against use of the original MTOS-derived TRISS weights in a dissimilar sample of trauma patients.

Mutable covariate weights, then, allow the TRISS (and other scoring systems) to act as data-adaptive procedures requiring calibration, rather than as deterministic equations. However, the implication of this flexibility is that any specific set of weights – and perhaps even chosen sets of covariates – are not expected to be broadly applicable and are unlikely to stay constant in the population of interest over time or space. One scenario under which this can occur is given by Champion et al. [23]: "as improvements in trauma care over time result in decreased mortality… coefficients can be expected to change."

Some scores are not updated, but are instead completely revised every few years, with the resulting new score proposed as a replacement for the old. SAPS and APACHE are representative of this strategy; for example, the "points" assigned to various categories in scores such as SAPS II and 3 are not designed to be re-estimated.

Even when weights are periodically re-estimated for a given population of patients, the motivation appears to be future reuse of those weights so that re-estimation is not necessary for every new sample for which prediction is desired. One major purpose of this "fixed weight" approach is to provide a consistent basis for comparing outcomes at different trauma centers for quality assurance or academic research purposes – perhaps the most common use of trauma scoring systems outside of triage.

One way to conceptualize this "do-ahead" approach of trauma scoring systems from a machine learning or artificial intelligence perspective is as an "eager learning" algorithm where the computationally intensive portion of the individual patient's estimate is completed well ahead of time. Specifically, covariates are selected and weights estimated (regression coefficients, point systems, or, in the case of ICISS, survival risk ratios (SRRs): empirical probabilities of survival given a particular injury [96]) prior to the time when predictions will be needed, using clinical knowledge and/or via a data-driven procedure performed on a large patient database. Unfortunately, prediction accuracy can suffer when an eager learning approach is utilized if the prediction function is too generalized or the training data used are not drawn from the same population as the sample on which a prediction is desired.

Limitations

Some trauma scoring systems are subject to additional limitations. Of the scores listed in Table 1.1, APACHE II [59] and SAPS II [38] aren't useful for triage or prediction of early mortality, as the scores are designed to be based on the worst measurements over the first 24 hours spent in the ICU (though at least one study has demonstrated that calculating such scores using admission values may not result in significantly decreased discrimination performance [51]). Some of their brethren, not listed in Table 1.1, have the same requirement [39, 58, 60, 125].

Most importantly for the purposes of triage and early outcome prediction, many trauma scores are typically not available until days or weeks after the scores could be utilized to inform treatment. This includes any scoring system which depends on AIS or ICD coding (see the "Injury Scoring" column in Table 1.1). The reason for this delay is that, to correctly code injuries, some of the required details aren't knowable until treatment is well underway or concluded. Assigning AIS ratings based only on information available in the ED would result in underestimation of any trauma score which depends on those values, particularly for patients with penetrating injuries [72, 122, 105].

In addition, translating medical records into standardized injury or medical codes is expensive and time-consuming, requiring extensive specialized training for medical coders. AIS coding is typically too costly for smaller hospitals, non-trauma centers, or medical facilities in developing nations to implement, and is therefore primarily utilized at large, urban trauma centers with trauma registries [105]. Post-operative AIS coding by trauma surgeons using a specialized mobile health application may provide a more affordable alternative [109]. On the other hand, ICD coding is available at most U.S. hospitals as a byproduct of the billing process, but since the ICD standard was not designed with injury severity assessment in mind, its capacity for predicting clinical outcomes may be limited [102].

A Trauma Scoring System for the 21st Century

Machine Learning

Over roughly the past twenty-five years, trauma researchers have increasingly utilized advanced machine learning methodology in scoring system development. Liu and Salinas [71] identified 65 published studies pertaining to the use of machine learning to predict clinical outcomes in trauma patients. Although 1993 was the earliest publication date of the articles retained by their search, just over half (50.8%) of the selected studies were published in the seven years prior to the article's publication (2010 to 2016, inclusive).

While the majority of the research catalogued by Liu and Salinas [71] focused on the prediction of patient mortality, other outcomes of interest included shock, hemorrhage, hospital admission, and length of hospital stay. Many covariates were common across studies, particularly patient age, physiological measurements (blood pressure, RR, and HR), and trauma severity scores (especially GCS and ISS). Popular prediction algorithms included artificial neural networks, support vector machines, decision trees, Bayesian methods (such as naive Bayes classifiers or Bayesian belief networks), clustering/nearest neighbor methods, random forests, and other ensemble classifiers (such as SuperLearner [65]). The great majority of these studies observed improvements in prediction over existing trauma scores when using machine learning methods, a gain that was perhaps due to the former's adaptability to nonlinear relationships or their objectivity in weighting or selecting predictive features. However, performance varied greatly across studies.

Computerized Clinical Decision Support Systems

Aside from post hoc trauma scoring, a promising and related mechanism by which machine learning could be better utilized is via clinical decision support (CDS) during triage or treatment of traumatically injured patients. Berner [12] defines CDS systems (CDSS) as "computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made." Such systems are not intended to replace physicians but rather to encourage standardization of care and offload a portion of the memory-intensive, easily automatable, and/or error-prone tasks associated with patient care. In fact, accurate clinician input is critical for a CDSS to provide reliable output, and eliciting accurate CDSS input from a physical exam of a patient in the ED has been shown to be highly dependent upon the treating clinician's level of expertise [50].

Computerized CDSS can improve drug prescribing and dosing, reducing drug-drug interactions and adverse drug events [119]. CDSS also have been shown to improve adherence to protocols for the prevention of diseases for which traumatically injured patients are at higher risk, such as venous thromboembolism, thereby conferring a lower risk of morbidity [119, 78]

Historically, CDS has been implemented as combinations of deterministic rules, developed and specified by subject-matter experts. A rule- or "knowledge"-based system such as this can provide suggestions based on information available in a patient's electronic health record (EHR) but cannot "learn" from clinical data. In contrast, a CDSS which employs machine learning techniques can identify novel patterns in data, requires less direct supervision (since specification of a priori rules isn't required), and can improve its performance as clinical data are added [12].

Machine learning algorithms are also more robust to missing information than rule-based systems and hold the potential to yield more accurate predictions when many variables influence the outcome [3].

With medical practitioners' increasing use of handheld devices by the bedside and increased availability and affordability of high-performance computing resources, including machine learning in CDS is more feasible than ever before. However, to be useful in the ED or for mobile triage, predictions must be available in real-time. Past work on real-time prediction for CDS includes ensembles of simple classifiers for early prediction of adverse outcomes in trauma patients [25] and patient-specific calibration of bedside monitor alarms in the ICU [124]. Perhaps the most critical hurdle for the successful integration of machine learning into CDS, though, is gaining the trust of clinicians. To achieve this goal, machine learning algorithms must be accurate, their methodology must not be enveloped in an "opaque box," and their results must be easily interpreted by medical experts [3].

A Novel Approach

With the above requirements and constraints in mind, we sought to develop an improved decision algorithm for precision trauma medicine.

The ideal prognostic tool would be more accurate than existing scoring systems and would perform competitively against established ensemble machine learning algorithms. This algorithm should be parsimonious in the predictors it uses for an individual, and comprehensible to end users, clinicians, who may be wary of so-called "opaque box" machine learning algorithms. It should utilize all available information and make use of the knowledge of what information isn't available. For example, there may be value in utilizing a patient's baseline respiratory rate for prediction of a particular outcome, but knowing that a patient's baseline respiratory rate was not observed may also be informative.

Our goal is to provide a patient-specific, personalized prognosis of a future outcome at any timepoint during care. However, the algorithm used must be flexible enough to predict well in heterogeneous patient populations, where a physician might expect the covariates that are related to the outcome of interest to be different for different subpopulations of patients.

Physicians and trauma surgeons are skilled at synthesizing information from the many patients they treat and case studies they read about and subsequently mentally grouping similar patients together. Partly via this practice, some physicians are able to develop an "instinct" for patient care based on the symptoms, treatment, and outcomes they've seen in patients past. Motivated by this learned instinct, we'd like to know if we can accomplish something analogous by synthesizing patients' medical records using some type of statistical modeling.

This idea of mimicking the physician's learned instinct or personal "knowledge base" built by observing many patients inspired us to employ what can be referred to as a local or localized learner, a concept introduced in earlier chapters in this manuscript. These are algorithms which don't assume a global model but instead adapt to the characteristics of an observation for which a prediction is desired. A toy example of this concept in a trauma setting is looking in a database to find past patients who are "most similar" to the patient for whom a prediction is desired. The new patient's prediction could then simply be an average (or similarity-weighted average) of the observed outcomes for the similar patients who were identified.

Prognoses from such an algorithm – using only the covariates available during treatment – would ideally be made directly available to the treating physician via a mobile app or bedside biomedical device. This means that the prediction algorithm would need to be fast without sacrificing predictive power. One tactic for accomplishing that goal would be to complete much of the time-consuming computational work in advance.

Local Learning

Definition

'Local' or 'localized' machine learning has previously been described as follows:

> "Instead of globally modeling data, **local learning** is more task-oriented. It does not aim to estimate a density from data as in global learning… it also does not intend to build an accurate model to fit the observations of data globally…. it only extracts useful information from data and directly optimizes the learning goal." [53]

More generally, local learners are non- or semiparametric algorithms which don't assume a global model, and instead, fit 'locally' in that they adapt to the characteristics of an observation for which a prediction is desired by using a subset of information from similar observations seen during model training. More specifically, in this manuscript, the phrase 'local learner' is intended to describe a machine learning algorithm which is designed to be capable of using *observation-specific* a) reduced feature spaces **and/or** b) reduced data spaces (in practice, reduced sets of training observations) in its prediction of outcomes of interest.

Local learning spans a spectrum of algorithmic designs, with some algorithms more 'local' than others. Some related categories of machine learning algorithms include memory-based, instance-based, case-based, or distance-based learners [113]. As alluded to in the paragraph above, it is perhaps easiest to define local learners relatively; in other words, local learning algorithms are best defined by what they are **not**: global.

To provide a concrete comparison point, a classic example of a 'global' learner is simple linear regression. It is considered a global learner because a single set of parameters (α and β) are estimated using a single predictor to predict an outcome of interest. When these parameters are estimated, all observations in the training set are used, and when this function is used to generate predictions, it does so using the same single predictor and same estimated parameters ($\hat{\alpha}$ and $\hat{\beta}$) across any and all available observations.

In contrast, a simple example of a local learning algorithm is a decision tree. Each branch of the tree can be thought of as containing a 'local' model that was estimated using a subset of the data space and potentially a subset of the feature space. For example, consider a decision tree with depth two where two branches are grown out of the root node and subsequently two branches are grown out of each of the two first-level nodes, producing a total of four terminal nodes at the second level. Imagine further that the first split that creates the two branches between the root node and the first-level nodes is predicated on the value of X_1. The split that grows two branches out of the 'left' first-level node is based on the value of X_2, while the split that grows two branches out of the 'right' first-level node is based on the value of X_3. This simple tree can then be thought of as representing two prediction functions, where the function applied to a given observation is dependent on that observation's value for X_1. Each of these two prediction functions uses a subset of the feature space: X_1 and X_2 on the left side of the tree, and X_1 and X_3 on the right side of the tree. Additionally, after each split, the left and right branches will each use only a subset of the data space to both grow additional branches and to predict the outcome of interest, with the observations eligible to contribute being filtered by any prior splits in the tree.

Taking the local concept to its extreme in the 'data space' dimension, another example of a local learner could involve selecting from a training set a single record that is most 'similar' to the observation for which a prediction is desired (more commonly referred to as its 'nearest neighbor'). The prediction, in that case, would simply be the observed outcome for that single most similar observation. In this case, only a subset of the data space ($n = 1$) is used to predict the outcome for the observation of interest, selected via some quantification of similarity (e.g. an inverse Euclidean distance) involving some or all of the feature space.

Feature Space Reduction Methods

We will first provide some concrete examples of algorithms that help to define 'localness' in the feature dimension. This is by no means an exhaustive list of feature selection or reduction methods, but is intended to highlight some commonly-used algorithms that possess desirable qualities.

Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator, or LASSO [112], is a particular form of regularized regression. Broadly speaking, application of regularization to a machine learning algorithm constrains or shrinks the magnitude of coefficient estimates [33]. In the case of the LASSO, the shrinkage method utilized happens to have the additional benefit of setting some coefficient estimates to exactly zero, effectively eliminating the influence of the corresponding features from the final model. This makes the LASSO a useful algorithm for feature selection.

The LASSO achieves this feature elimination via the addition of the product of a penalty term and a regularization parameter to the least squares regression equation where the optimal set of coefficient estimates still minimizes the residual sum of squares. Specifically, the penalty term used is the sum of absolute value of the coefficient estimates, or the L_1 norm of the vector $\hat{\beta}$, $\|\hat{\beta}\|_1$. For this reason, the LASSO is also known as L_1 regularization. The regularization parameter λ, where $\lambda \geq 0$, indirectly controls the degree of shrinkage, with a larger λ forcing the magnitude of coefficient estimates toward zero and $\lambda = 0$ being equivalent to no regularization (i.e. ordinary least squares).

In practice, λ serves as a tuning parameter and a search across many potential values of λ is performed, with each λ_i potentially resulting in a different set of 'optimal' coefficient estimates $\hat{\beta}_i$. The selection of the 'optimal' value of λ for a given feature selection problem is typically performed via k-fold cross-validation. This yields a single set of 'optimal' coefficient estimates $\hat{\beta}$. In the feature selection context, the retained features are those with non-zero $\hat{\beta}$ at this 'optimal' λ.

Random Forests

The family of decision tree algorithms, introduced above as an example of a broad class of local learners, forms the basis of random forests (RF) [17]. As the name implies, random forests are collections or ensembles of many trees, each of which is fit on a random subset of features and bootstrap-sampled observations from the overall training set. A predicted outcome for the i^{th} observation is estimated by aggregating across the predictions for i from each tree in which i was omitted from the bootstrap sample [47], where the aggregation method depends on the prediction task (e.g. averaging for regression; majority vote for classification).

More generally, the technique of aggregating across B model fits, each of which used one of B bootstrap samples as its training set, is known as 'bagging' [16]. An estimate that utilizes, for each observation, only those predictions where the observation was omitted from the bootstrap sample, is referred to as an 'out-of-bag' estimate. Thus, predicted outcomes generated by random forests are out-of-bag predictions.

As is the case with many algorithms used for variable selection, including the LASSO described above, random forests are primarily supervised learners that happen to double as highly effective feature selectors. In the case of random forests, variable importance can be estimated via either Breiman's original method or one of a number of alternative methods [108]. Subsequently, using a tuning parameter specifying a cutoff, the continuous variable importance measure can be dichotomized into a variable selector.

Specifically, Breiman's original variable importance method for random forests uses the same model fit but a series of slightly modified datasets to generate P alternative sets of predictions for each of P features. For each of the features in turn, the values for the out-of-bag observations are scrambled or permuted to simulate the loss of any predictive information from that feature. Then P new accuracy estimates are compared with the accuracy estimated from the original predictions generated without permutation [17]. The percent decrease in accuracy under permutation can then be used as a measure of variable 'importance,' where a larger value indicates that the feature was more informative for accurate prediction of the outcome and a smaller value indicates that the feature led to relatively little improvement in the accuracy of the predictions.

The value of the tuning parameter used to dichotomize the variable importance measure into a variable selection indicator will typically be slightly greater than zero. However, the 'optimal' setting may depend on the objective: a small value very close to zero may be used to maximize prediction accuracy, whereas a larger value further from zero may be better to maximize interpretability [41].

Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) [37] predict an outcome vector by optimally combining several piecewise linear functions, or splines. A spline is estimated for a feature by iteratively choosing a point in the training set to serve as a 'hinge' or 'knot.' Each knot serves as a breakpoint, and the placement of k knots allows a formerly linear function to be broken into $k + 1$ segments. In order to be chosen as the next knot during any given iteration, the spline formed using that knot must minimize the prediction error relative to the set of splines that could be formed by placing the knot at any other point (of the points available in the training set). To avoid overfitting, this knot selection process can be controlled by one or more tuning parameters.

In MARS, the optimal set or ensemble of splines is selected via a two-step process where, during the first (forward) step, many splines are added, and during the second (backward) step, a subset of those splines are removed. The backward step performs this deletion or 'pruning' by iteratively choosing the term that confers the least reduction in error and removing it until the stopping criterion is reached. This stopping criterion or 'optimal' fit is typically one which minimizes a generalized cross-validation criterion balancing prediction error and model complexity. The features contributing to the splines remaining in the model after the backward pass completes can be thought of as 'selected' features, and in this way, MARS can be used as a feature selection algorithm.

Data space reduction methods

k-Nearest Neighbors

The k-nearest-neighbors (kNN) algorithm is quintessential example of a non-parametric supervised learning algorithm which uses only a subset of the data space for each prediction. In kNN, for each observation for which a prediction is desired, a 'neighborhood' relative to the training set observation is defined using a distance metric and k, or the number of neighbors to consider in estimation of the predicted value [30]. A smaller distance between a pair of observations implies they are more similar and therefore closer neighbors, whereas a larger distance implies dissimilarity and so the two are more distant neighbors. Equivalently, the closest neighbors of an observation can be thought of as those maximizing similarity as measured by the pairwise *inverse* distance.

This distance is a function of the covariate space. The canonical example (in the case of continuous covariates) is the application of the Euclidean distance function to the standardized covariates or a particular subset thereof. Subsequently, the mean (or majority vote, in a classification context) is estimated across the k neighbors in the training set closest to the test set observation, and that mean (or majority class) serves as the prediction for that test set observation. Specifically, the k-nearest neighbor outcome estimate $\hat{Y}$ in the continuous case is defined as

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

where $N_k(x)$ is the neighborhood of x defined by the closest points x_i in the training set [47]. The tuning parameter k can range from 1 (where the predicted value is the observed value in the closest neighbor) to n (where the predicted value is the mean across all training set observations). Any value of $k < n$ therefore serves to reduce the data space considered for any given prediction.

Holding the sizes of the training set and covariate space constant, as k increases, the bias of predictions increases but their variance decreases. Cross-validation can be used to choose the optimal k that balances bias and variance while minimizing mean squared error.

Weighted Nearest Neighbors

There are many extensions to the simple kNN framework described above that, taken together, form a family of nearest neighbor methods. One salient extension, for example, extends the kNN algorithm beyond the case where a prediction is the mean of or majority vote among the nearest neighbors to the more general case where the prediction for the observation of interest is estimated via a *weighted* combination of its nearest neighbors [32]. From this perspective, unweighted kNN can be thought of as a special case where a constant weight $\frac{1}{k}$ is applied to each observed value for all observations in the neighborhood of size k.

To predict a value for the observation of interest in the weighted case, the values of closer neighbors are upweighted and those further away are downweighted. Options for estimation of weight vectors include functions of the distance (such as the standardized inverse distance) or of the covariates themselves.

A method that could be used in either case is kernel density smoothing or kernel density estimation (KDE). In the latter case, a multivariate density could be utilized, using observed values of covariates as input. An example of the former case is a kernel function K input d, a vector of estimated distances. The function K also requires a tuning parameter: a bandwidth, h. If a K with bounded support is selected, its bandwidth h determines the extent of that support. This leads to the downweighting of distant neighbors to zero, while closer neighbors have larger, non-zero weights. In this way, the bandwidth h subsumes the role of the kNN tuning parameter k by limiting the number of neighbors and thereby defining each 'neighborhood' of training set observations. Simultaneously, the bandwidth h, in combination with the chosen function K, produces a vector of weights to be used in the prediction of the value for the observation of interest e.g. via a weighted mean (after standardizing the weight vector w such that $\sum w = 1$).

Advantages of weighted kNN over the unweighted approach include better-than-random tie-breaking in the classification context and an error rate in either context that is less sensitive to the choice of an unduly large k [32].

A Brief Review of Existing Local Learning Algorithms

Scatterplot Smoothers

"Scatterplot smoothers" are local regression algorithms that derive their colloquial name from their simplest, two-dimensional form [26]. Locally estimated scatterplot smoothing (LOESS) and the related locally weighted scatterplot smoothing (LOWESS) are the two most commonly used techniques in this category. These methods are popular nonparametric alternatives to linear regression, particularly for exploratory data analysis. They are in the category of local regression algorithms because they are made up of multiple locally-estimated regression estimates where each estimate uses only a subset of the observed data.

In basic terms, the LOESS is fit as follows. First, for each observation, distances to every other observation are computed. These distances are based on some set of one or more observed covariates (W) and are typically the Euclidean distance in the multivariate case. Then, these distances are used to calculate weights based on some weight function or kernel. Assuming a sample size of n observations, n locally weighted polynomial regressions are then fit for each of the n observations (for each $\{w_i, y_i\}$, a $\hat{y}_i$ is estimated). These may be of degree 0 (implying a running mean), they may be locally linear (degree $D = 1$), or they could be locally quadratic ($D = 2$). A larger D leads to decreased bias but increased variance [47] and is more likely to produce an overfit estimate.

In the typical implementation, estimation is performed iteratively to increase robustness to outliers. At a minimum, the following tuning parameters are needed: span/window width h, distance function d, weight/kernel function g, and polynomial degree $D \in \{0, 1, 2\}$. The optimal span (h) is often selected via cross-validation, while other necessary tuning parameters are typically fixed by the particular implementation or chosen by the analyst [47]. Although the continuous, nonlinear features of scatterplot smoothers make them valuable exploratory analysis tools, the lack of explanation provided for individual predictions ($\hat{y}_i$) limits the interpretability and usefulness of this class of algorithms.

LIME

The Local Interpretable Model-agnostic Explanation (LIME) [98] algorithm is another type of local learner, one which is focused on filling in the interpretability gap between any so-called "black box" model and one of its predictions. A small library of interpretable prediction algorithms (including linear regression and decision trees) are utilized by LIME to approximate and explain any model fit locally, where 'local' is with regard to the data space (within some tolerance around an individual observation of interest).

LIME's first focus is on model interpretability for the intended audience. Its second focus is that the "local surrogate model" and subsequent explanation produced are a faithful approximation of the original model, but only with regard to a small neighborhood centered around the individual observation of interest. Specifically, LIME minimizes the sum of the local infidelity of the explanatory model with regard to the original model and the complexity of the explanation (a proxy for ease of interpretation). In this way, regularization is encouraged and therefore a LIME estimate is often 'local' with regard to the feature space, as well.

As an extension to the core LIME algorithm, local (observation-specific) explanations can be compiled using a targeted subset of training set observations to produce a "global" model interpretation. This process, resulting in a global approximation of the original predictive model's feature weights, is referred to as "submodular pick." This "submodular" method of choosing a subset of local explanations to include in the global explanation has been shown to yield more useful and representative feature weights than random selection would [98].

Notably, the LIME algorithm does not directly optimize for prediction accuracy with regard to the observed outcomes. In addition, the predictions and complex behaviors of some "black box" model fits are not guaranteed to be faithfully representable via LIME or members of its library of interpretable prediction algorithms.

Conditional SuperLearner

The Conditional Super Learner (CSL) is a local learner developed concurrently with the algorithm to be presented in this manuscript [114]. This prediction algorithm, also like the algorithm to be presented in this manuscript, builds on a powerful ensemble learner, the Super Learner (SL) [65]. The SL algorithm splits a dataset into V cross-validation folds, and within each fold, fits a library of learning algorithms to the training set and evaluates those fits on the respective test set. The predictions from each learner in the library can be combined via non-negative least squares or other similar methods to yield weights, which are subsequently applied to each learner's predictions. The library of learners is then refit on the entire dataset, the predicted outcomes from which are combined according to the weights from the previous step, yielding a vector of predictions at least as good as the best learner in the library (asymptotically).

Instead of a global model estimate, the CSL algorithm adds a classification 'metalearner' layer (referred to as the 'oracle'), producing K localized SL fits (referred to as 'experts'). The cross-validated, two-step iterative algorithm first utilizes a subset of the covariate set to estimate the 'oracle,' dividing the covariate space into K regions. Subsequently, within each of the K regions, an 'expert' is fit and outcomes are estimated using each cross-validation training set that falls within that region. This iterative process is repeated T times and, finally, each observation's predicted outcome is the one produced by the final 'expert' for the region in which the observation's covariate values place it.

CSL's hierarchical approach allows for more interpretable models to be fit at both the global 'oracle' level and the local 'expert' level than would be possible if only a single global model were estimated instead. The authors demonstrate that the accuracy of predictions estimated via CSL are, in real-world datasets, nearly always no worse than the global model in which the 'expert' learners are stacked. However, the stability and performance of CSL are highly sensitive to the choice of K [114], and an automated method for choosing the optimal K has not been proposed. Additionally, this method does not target the best 'expert' model fit for an individual observation, but instead targets the best model fit for some subregion of the full covariate space.

Advantages of Local Learning

Through the review above, some desirable qualities of a local learning algorithm can be identified. The ideal local learning algorithm should be flexible in the true underlying data generating distributions it accommodates, allowing for complex and/or non-linear patterns in its estimation. Simultaneously, these algorithms should emphasize interpretability and explainable predictions. To that end, the model fit for a given observation should be parsimonious, but without being limited in the covariates it can choose to utilize in its prediction estimates. This behavior should enable accurate, personalized predictions, but without overfitting.

With the above in mind, the remainder of this manuscript proposes and validates a local learning algorithm that fulfills the above requirements. First, the specifics of the method will be presented, and subsequently, its potential use in clinical decision support as an alternative to existing trauma scoring systems will be demonstrated.

Chapter 2

Methodology & Simulation Results

Prediction Framework

Data and target parameter

The solution proposed in this manuscript applies to a data structure shared by many prediction problems. This data structure contains an outcome Y, which will be assumed, for simplicity, to be binary. It also contains covariates W, which could include variables both directly observed and calculated from observed, such as indicators of missing data or derived features. These covariates are assumed or known to be related to the outcome and therefore informative for its prediction. This framework yields an observed data structure of $O = (W, Y) \sim P_0$ where P_0 is the true – but in practice unknown – data-generating distribution.

The parameter of interest in the estimation problem presented here is the expected value of some outcome Y given the covariates W, for some specific combination of covariate values in the covariate space. More specifically, this target parameter is $\psi_0 = \Psi(P_0) = E_0(Y \mid W = w)$, or the conditional mean (or probability) of the (binary) outcome Y evaluated at a specific stratum $w \in \mathcal{W}$. For example, an estimate of the probability of death by 24 hours after hospital admission for a particular patient profile may be desired, where a 'patient profile' implies specific values of relevant clinical variables.

Algorithm

To estimate this target parameter, we propose a supervised semiparametric prediction algorithm which is 'localized' such that only the *most relevant* information for a given observation is utilized for that observation's prediction. Here, *relevant* information includes both observations and covariates. The *most relevant observations* are those for which observed covariate values are most *similar* to the values in the stratum of interest, where *similarity* is defined in the "Distances and Neighborhoods" section below. The *most relevant covariates* are those which are determined, via feature selection or variable importance algorithms, to be more closely related to, or predictive of, the outcome of interest than other covariates. General outlines of the algorithms used in this determination are detailed below under " 'Global' Feature Selection" and " 'Local' Feature Selection."

Data Preprocessing

If the observed data for which a prediction is desired contain missing observations within the covariates, any covariates not fully observed should be preprocessed to eliminate these missing values before applying the algorithm described below. One approach, referred to as the "missingness indicator" method, is to recode missing observations to zero (or another suitable value, such as the empirical mean or median) and create an indicator of whether the covariate was measured for each observation. This strategy, or one like it, implies that the missingness may be informative in itself for prediction of the outcome of interest. In fact, it has been argued that this approach to handling missingness may be superior to other approaches when the primary goal is prediction (as opposed to parameter estimation), the missingness is potentially predictive, and particularly when it is difficult or impossible to achieve consistency in the approach taken to handle missingness between model training and model deployment [110].

Note that the selection of imputation strategy could be considered a tuning parameter (explained in greater detail below) and an alternative *a priori* choice could be made by the analyst at this juncture. Since missing data imputation was not the focus of this project, for the purpose of simplicity, the "missingness indicator" method was utilized wherever necessary for all analyses presented herein.

Algorithm Overview

After any preprocessing steps are complete, the proposed prediction algorithm first reduces the size of the feature set. This reduced set of covariates is then used to define a neighborhood of observations that are similar to the observation for which a prediction is desired, reducing the observed data to be utilized. Finally, within this small neighborhood, a smaller, neighborhood-specific set of covariates is chosen, further reducing the dimensionality of the feature space. Simple parametric prediction functions are then used to predict the observation's outcome. In short, this method performs both dimension and instance reduction data-adaptively to hone in on only the most relevant information for a given observation. The details of this process will be expanded upon in the subsections that follow, and a high-level summary of the algorithm can be reviewed in Algorithm 1. Additionally, an R package implementing this algorithm is available on GitHub [89].

Algorithm 1: Local Learning Algorithm

```
within the tuning set, divide records into k folds;
foreach unique combination of tuning parameters do
    for i = 1 to k do
        using training set observations for which fold ≠ i, perform global
          variable selection W' ⊆ W via CV Super Learner [65];
        foreach validation set observation j for which fold == i do
            estimate dist. D_j from j to each training set obs. using W',
              centered on W'_j and scaled by a selected norm;
            input D_j into kernel function(s) to define j-specific kernel
              weighted neighborhood(s) of training set obs.;
            estimate Ŷ_j via simple parametric algorithms on weighted
              training set obs., utilizing neighborhood-specific W'' ⊆ W';
select the tuning parameter combination optimizing a measure of model
  performance in the tuning set;
foreach test set observation m do
    predict Ŷ_m using the selected tuning parameter combination set for a
      model trained on the tuning set;
```

Training Procedure

Each step of this procedure is data-adaptive and therefore utilizes *tuning parameters* (also known as *hyperparameters*) that must be tuned or optimized for the data generating mechanism at hand. The process of selecting optimal tuning parameters is also sometimes referred to in machine learning as *model selection*. This is accomplished by first "training" the prediction algorithm on a representative data set, which will be referred to here as a *tuning set*, assessing prediction performance via cross-validation under a variety of settings for each tuning parameter. These performance assessments can be used to inform the selection of an appropriate value for each tuning parameter – selections which must be made prior to predicting outcomes for new observations of interest.

To retain some data on which to fairly evaluate the prediction performance of the algorithm with final chosen tuning parameter settings, the data should be split. For example, 25% of a data set could be 'held out' for use as a *test set*, and the remaining 75% would therefore be utilized as the *tuning set* to select tuning parameter settings. Alternatively, an entire data set with observed outcomes could be utilized as a *tuning set*, and the selected tuning parameter settings would subsequently be used to predict outcomes for incoming new observations, provided outcomes could be ascertained for those new observations to estimate the performance of the prediction algorithm.

The phrase *tuning set* is used rather than the more customary *training set* in order to signify that these *tuning set* observations will be further split, during the tuning parameter selection process, into V cross-validation folds, resulting in V partitions of the data into *training* and *validation sets*.

The purpose of this cross-validation procedure is to choose the optimal value for each tuning parameter data-adaptively. A combination of tuning parameter settings is deemed "optimal" for a given data set when it maximizes (or minimizes) some objective function. For example, the goal may be to find the combination of tuning parameter settings which maximizes the mean cross-validated area under the receiver operating characteristic (ROC) curve (cvAUC) [67] in the *tuning set*. The search for "optimal" *tuning parameter* settings may be performed exhaustively over a smaller space of possible parameter settings (i.e. via grid search) or randomly over a larger space of possible parameter settings (e.g. via random search) [11].

Specific tuning parameters for this algorithm will be discussed below. Generally, these are parameters which can impact the results of any of the steps of the local learning algorithm: feature selection, neighborhood selection, and prediction.

Once identified, optimal tuning parameter settings are subsequently used in combination with the entire *tuning set* to build a model with which to predict the outcomes $\hat{Y}$ in the *test set*. It is this predictive performance by which the algorithm is ultimately evaluated. Prediction performance in the *test set* could be compared with, for example, the Super Learner's [65] predictions in the *test* set (after being trained on the *tuning* set). This performance could also be compared against existing field-specific deterministic prognostic scoring systems, such as those used in trauma care.

Initial Covariate Space

A prediction problem of interest is assumed to begin with an initial set of features or covariates, which could be restricted by those available or by a priori hypotheses regarding covariates relevant for targeting the parameter and outcome of interest.

As a toy example, consider the following simulation of 200 tuning set observations and a single test set observation arising from a simple data structure of three covariates (W) and a binary outcome (Y). The initial covariate space in this example is three-dimensional: $\{W_1, W_2, W_3\}$. The first of these covariates is Gaussian noise and so is irrelevant to the prediction problem by design, as it does not influence the outcome or other covariates in any way. The remaining two covariates $\{W_2, W_3\}$ are related to one another and to the outcome Y.

$$
\begin{aligned}
W_1 &\sim \mathcal{N}(\mu = 1,\, \sigma^2 = 1) \\
W_2 &\sim \ln \mathcal{N}(\mu = 3.56,\, \sigma^2 = 0.44^2) \\
W_3 &\sim \ln \mathcal{N}\left(\mu = 0.08 \times \ln\left[W_2\right] + 3.01,\, \sigma^2 = 0.21^2\right) \\
Y &\sim \mathrm{Bern}(\mathrm{logit}^{-1}[0.015 \times W_2 - 0.013 \times W_3 - 2.35]) \\
&\qquad \text{where} \quad \mathrm{logit}^{-1}(\alpha) = \frac{\exp(\alpha)}{\exp(\alpha) + 1}
\end{aligned}
\tag{2.1}
$$

Equation Set 2.1 specifies the true data-generating distribution of the simulated data and therefore the true relationship between these variables. Figure 2.1 depicts, graphically, the relationship between the covariates W and outcome Y in the form of a directed acyclic graph (DAG) [107]. Figure 2.2 visualizes the empirical relationships between the three covariates and the outcome in the 200 simulated tuning set observations. In each panel of this plot matrix, visualizations are stratified by the binary outcome with the two outcome groups distinguished from one another by color. Gaussian kernel density estimates of each covariate appear along the diagonal, for which the bandwidths are estimated using Silverman's rule of thumb [106, p. 48]. In the panels on the lower left, scatterplots with simple linear regression lines overlaid describe the observed relationships between the continuous covariates. Due to sampling variability, spurious differences may appear to the naked eye where they do not exist in the data generating distribution.

'Global' Feature Selection

Particularly when the initial covariate space is high-dimensional, problems such as irrelevant features and data sparsity (either in the observed data or in the data-generating distribution) can hinder or even damage predictive performance [4]. Therefore, refining this initial feature space for the target parameter and population of interest may be desirable for improved predictive performance. High-dimensional spaces containing irrelevant features can be particularly troublesome when similarities or distances must be calculated [13], e.g. in a "localized" learning algorithm, as those estimates may indicate that two observations are "nearest" neighbors for

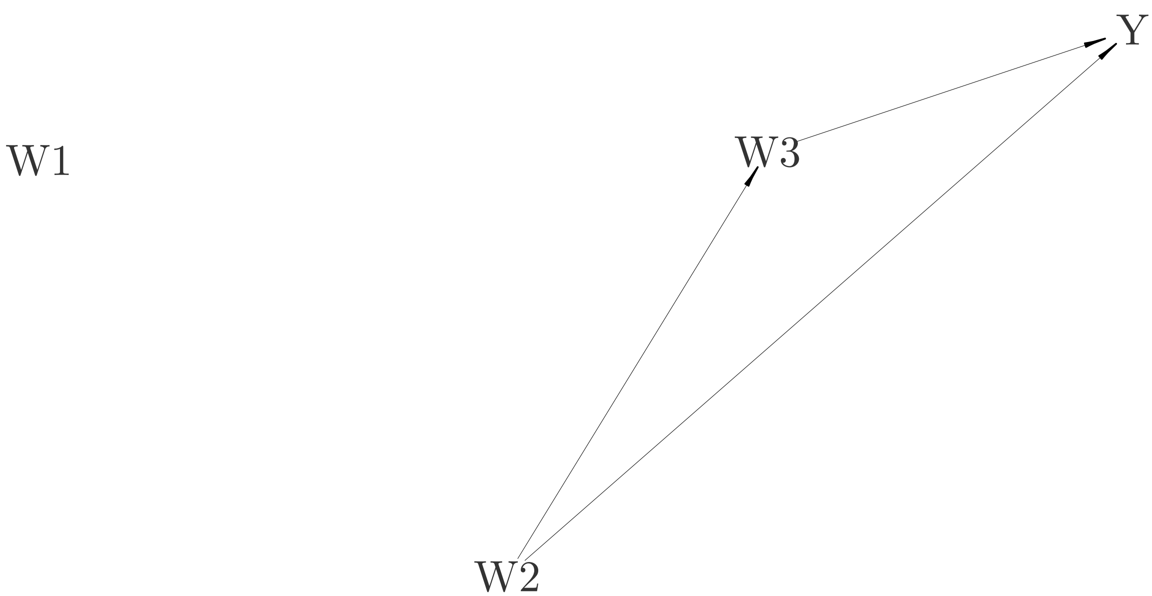

Figure 2.1: DAG representing relationships between variables in simulated example

reasons irrelevant to the outcome of interest. For that reason, feature selection or reduction is a critical first step in that context, since determining what is "similar" or "local" to a given observation is best determined by considering only relevant features.

Consequently, the first step in the local learning algorithm described here is to perform "global" feature selection. Figure 2.3 displays the result of having (hypothetically) applied feature selection globally – utilizing all tuning set observations – to the simulated data. The single simulated test set observation is overlaid in gray. In this case, the predictor known to be unrelated to the others via knowledge of the data-generating distribution, W_1, has been eliminated, while features known to be related to the outcome of interest Y have been retained. More generally, performing feature selection on the initial feature space W yields $W' \subseteq W$.

"Global" feature selection can be achieved in many ways, but in the preliminary results presented in the next section and chapter, feature selection was performed via a novel algorithm implemented in the SuperSelector R package [88]. This algorithm applies an ensemble learner – the cross-validated Super Learner (CVSL) [65] – within each of k cross-validation folds in the tuning set to select k sets of features to retain. The CVSL algorithm chooses the best combination of individual learners via minimization of cross-validated loss and subsequently reports the overall cross-validated loss of the entire Super Learner. Many ensemble loss functions for the Super Learner

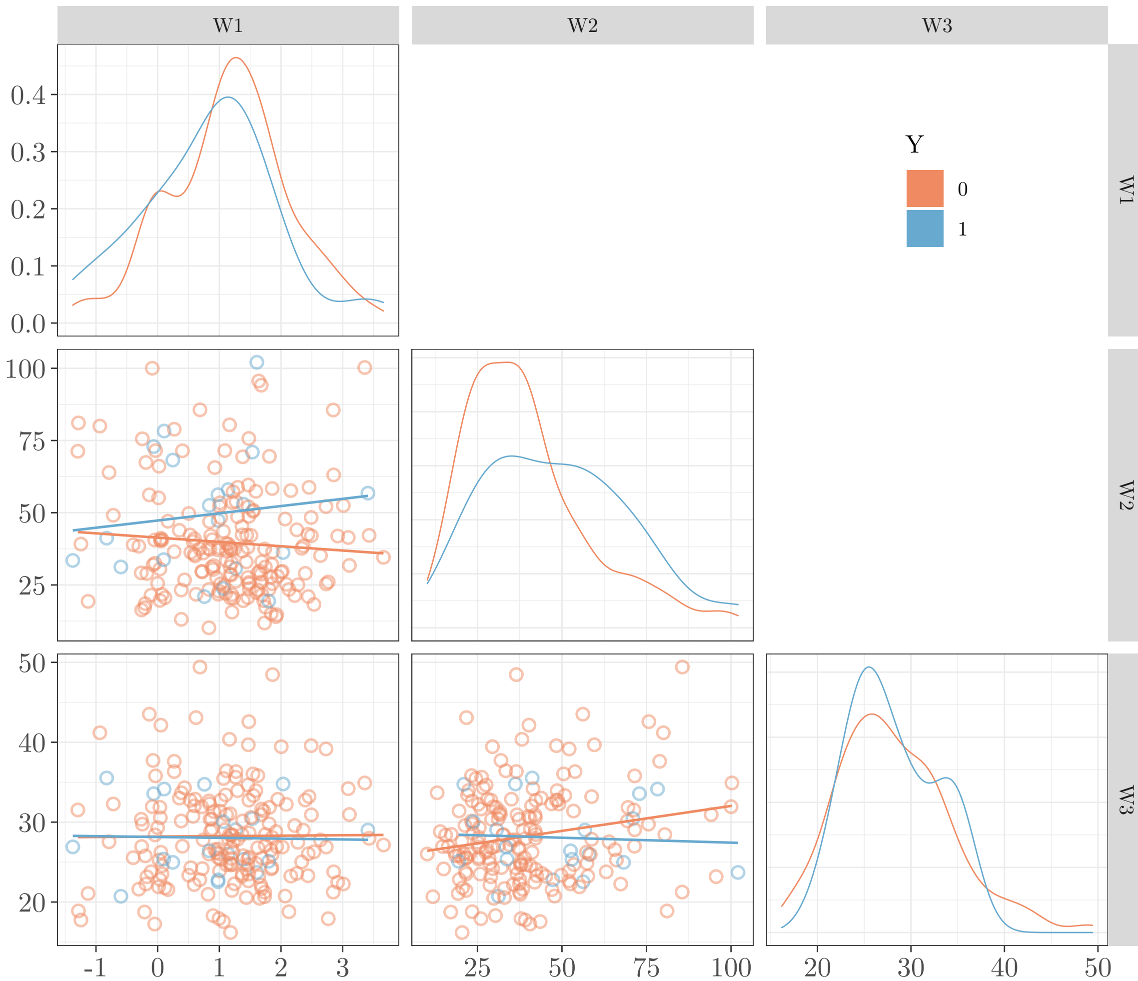

Figure 2.2: Initial covariate space

are available, but the two employed in the section that follows (included as potential settings for a tuning parameter, with one "optimal" setting chosen via grid search) were non-negative binomial likelihood maximization and non-negative least squares. Alternatives include AUC maximization. The loss function was one of a variety of tuning parameters for which the "best" setting was chosen via application of the algorithm to the tuning set, where this combination of settings would later be applied to the test set.

For the "global" variable selection process in this analysis, algorithms provided to 5-fold CVSL included LASSO [112, 36], random forest (RF) [17, 70], and multivariate adaptive regression splines (MARS) [37, 75]. Specifically, each algorithm performed

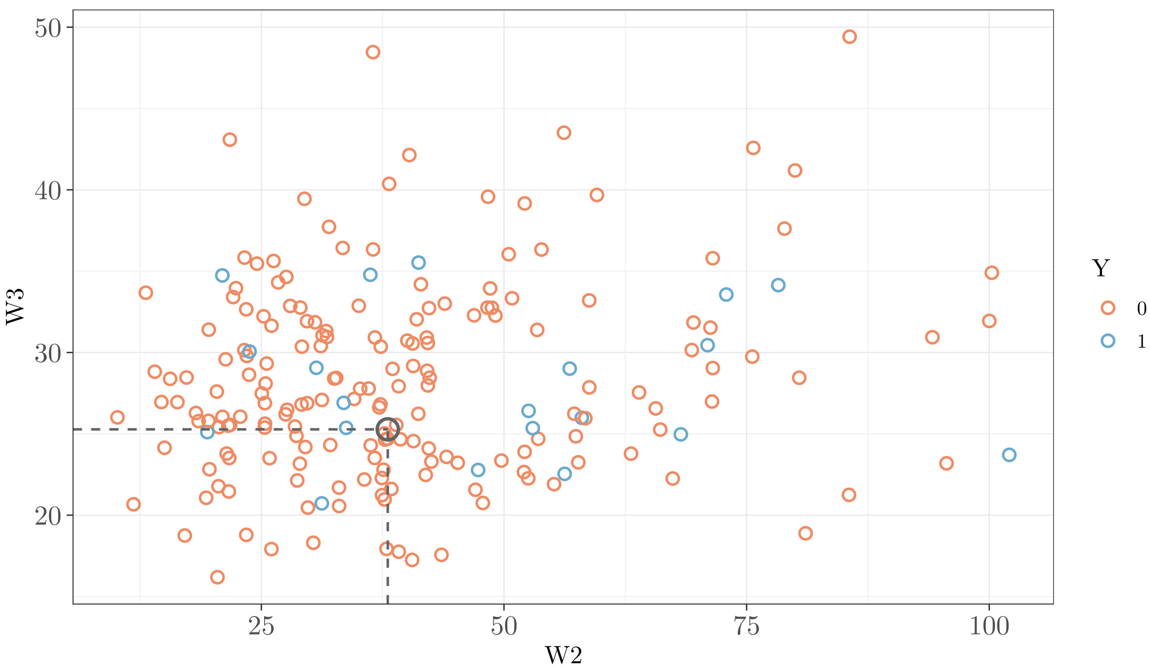

Figure 2.3: After 'Global' Feature Selection

variable selection as follows. The regularized GLM with LASSO penalty retained covariates with non-zero coefficient estimates at an optimal regularization parameter λ (where the optimal λ was one which minimized a 10-fold cross-validated binomial deviance loss function). RF chose covariates for which the mean decrease in accuracy under permutation exceeded a reasonable cutoff value (0.005). Covariates selected via MARS were those included in its final model. This was controlled by a parameter (n_k) limiting the number of terms, including the intercept, created by the forward pass. In Super Learner's MARS "wrapper function," the default setting for n_k is one plus two times the number of covariates, up to a maximum of 21. Alternatively, this can be controlled by specifying the maximum number of terms to keep from the backward or "pruning" pass (n_{prune}).

To select features globally for prediction of the test observation's outcome, the CVSL procedure was applied across the entire tuning set using the settings of the tuning parameters selected after cross-validation was performed within the tuning set. Overall feature importance for the test observation was determined by the mean proportion of CVSL folds across the tuning set in which the covariate was retained. This per-fold proportion could either be calculated as unweighted or weighted, where the weight would be the coefficient estimated for the respective algorithm by Super

Learner in that fold. The final tuning parameter choice needed for this process was the number of high-ranking features to keep, which could be a constant (such as top 20 or top 40, depending on the dimensionality of the initial covariate space) or a data-adaptive parameter. Ad hoc data-adaptive settings for this parameter include limiting the selected features to those with a higher ranking than the point where the maximum gap in importance scores (weighted or unweighted mean proportions) occurs or those for which the importance score was higher than some constant.

Alternatively, many other feature selection methods could instead, or additionally, be used at this junction. The local learning algorithm can treat the feature selection method as a tuning parameter and choose the "optimal" one via grid search.

Distances and Neighborhoods

Localized prediction algorithms require rules regarding which observations will be included in the prediction function of a given observation – and with what weights. These rules typically begin, implicitly or explicitly, with distance metrics $d(x_{test}, x_{tuning_i})$, functions of two vectors $\{x_{test}, x_{tuning_i}\}$ in the covariate space. Figure 2.4a demonstrates the Euclidean distances between the observation of interest x_{test} and all tuning set observations in the two-dimensional covariate space W'. Similarly, in Figure 2.4b, transparency is proportional to the Euclidean distance of tuning set observations (in two-dimensional covariate space) from the observation of interest x_{test}. The distances are calculated using centered and scaled features $W' \subseteq W$ (centered on W'_{test} and scaled by their L^1-norm). The distance metric (here, Euclidean or L^2) could be considered a tuning parameter.

These distances must, in turn, be transformed into a measure of similarity, wherein a larger value indicates that two observations are more "similar" to one another than a pair with a smaller value. The simplest such transformation is the inverse distance, or $\frac{1}{d(x_{test}, x_{tuning_i})}$. An alternative method and the one which is employed here is to use the distances as inputs to a smoothing kernel $K_h(x_{test}, x_{tuning_i})$. Implicit in this function K is the bandwidth or window width h. The resulting similarity measure can be used as a weight for each observation x_{tuning_i} relative to the observation for which a prediction is desired, x_{test}. One advantage of this method of distance transformation is that the similarities it yields are standardized to be between zero and one. Moreover, if the kernel has bounded support, some observations' weights will be explicitly zero, effectively creating a neighborhood $\mathcal{N}(x_{test})$ around the observation of interest, x_{test}. An example of this transformation with a tricube kernel is provided in Equation Set 2.2 below and assumes $d(x_{test}, \vec{x}_{tuning})$ is ordered from smallest to greatest and at least $1 - h\%$ of $d(x_{test}, \vec{x}_{tuning})$ are non-zero.

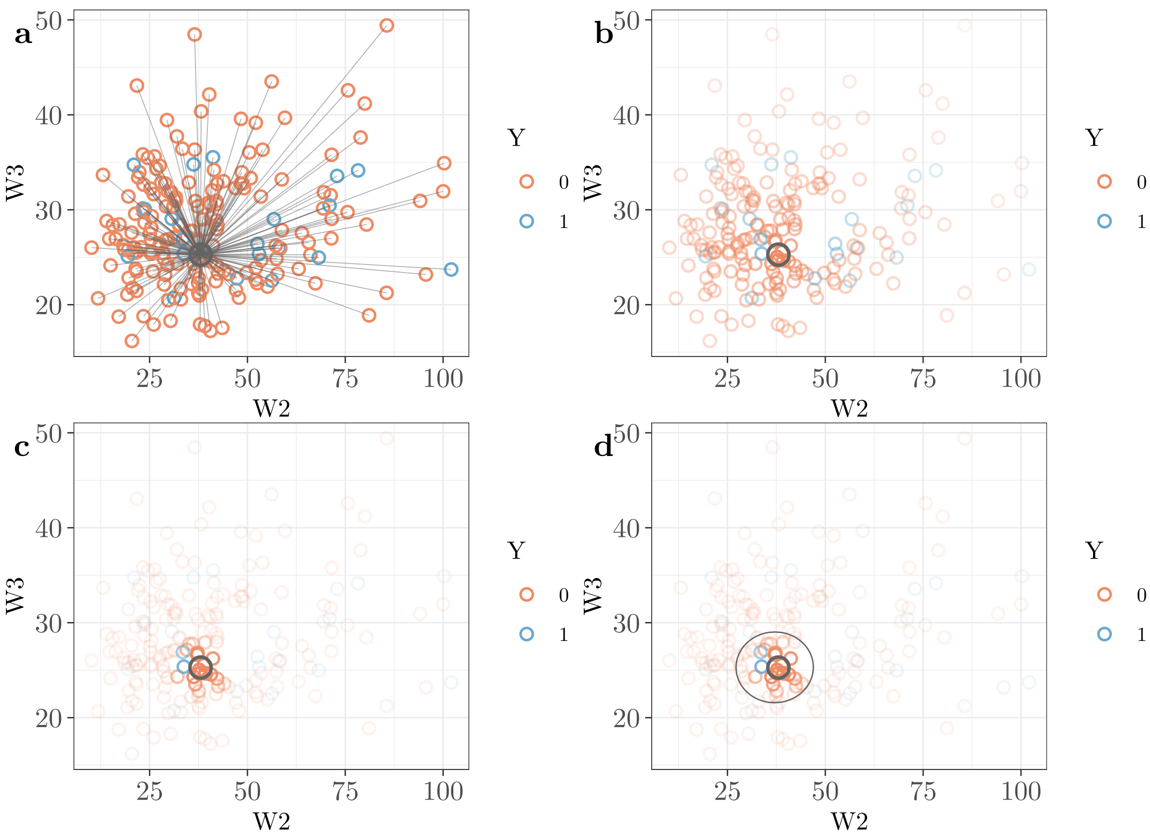

Figure 2.4: Distances and Neighborhoods

$$
\begin{aligned}
v &= d_{\lceil n_{tuning} \times h \rceil}(x_{test}, \vec{x}_{tuning}) \\
K_h(x_{test}, x_{tuning_i}) &= \begin{cases} \frac{70}{81} \times (1 - d(x_{test}, x_{tuning_i})^3)^3 & \text{if } d(x_{test}, x_{tuning_i}) \leq v \\ 0 & \text{if } d(x_{test}, x_{tuning_i}) > v \end{cases}
\end{aligned} \tag{2.2}
$$

In Figure 2.4c, the transparency of each point now is proportional to its weight, which (in this example) is the result of applying a tricube kernel like that in Equation Set 2.2 with a window width (h) of 20%. Figure 2.4d displays the same information, but an ellipse is now overlaid to delineate the covariate space containing tuning set observations with non-zero weight from that space outside the "neighborhood" where observations are assigned weights of zero.

The kernel K and window width h are also tuning parameters. Figure 2.5 contains a graphical depiction of the kernels included in this example: uniform, tricube, and Epanechnikov. These kernel functions are commonly used and share the prop-

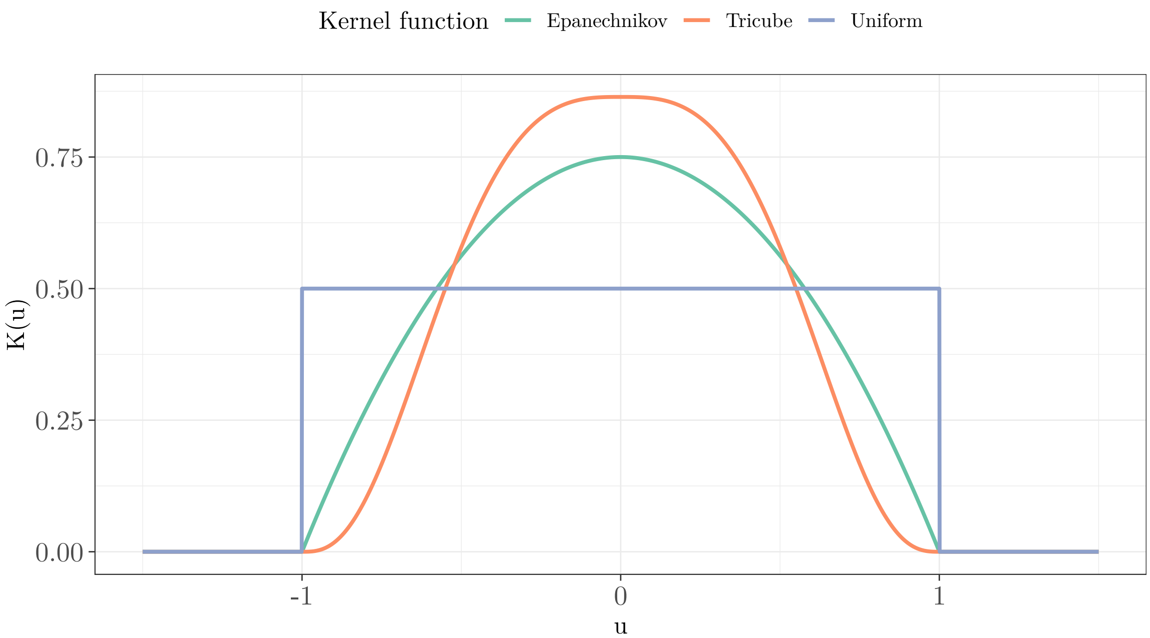

Figure 2.5: Demonstration of weighting functions for weighted regression: kernels used, in common coordinate system

erty of bounded support, meaning that some observations' weights will be explicitly zero. This leads to weighted 'neighborhoods' of observations, enabling a 'localized' prediction based on a subset of observations in the tuning set. Other kernel functions could be used instead. Window widths included in the examples which follow are expressed as a percentage of the number of observations in the tuning set (10%, 15%, 20%, and 25%) because these are also the sizes of the resulting "neighborhoods" around the test observation.

'Local' Feature Selection

Within the relatively small "neighborhood" determined above, we perform "local" feature selection. This process determines a further reduced set of covariates specific to the "neighborhood" of the test observation: $W''^{test} \subseteq W'$. For a given test observation, some quantity (e.g. five) of the top covariates in that observation's "neighborhood" could be used. This ranking could be determined by, for example, sorting on the magnitude of the weighted Pearson correlation coefficient estimates between the feature and the outcome of interest. Both the number of local covariates and the process by which they are selected can be additional tuning parameters.

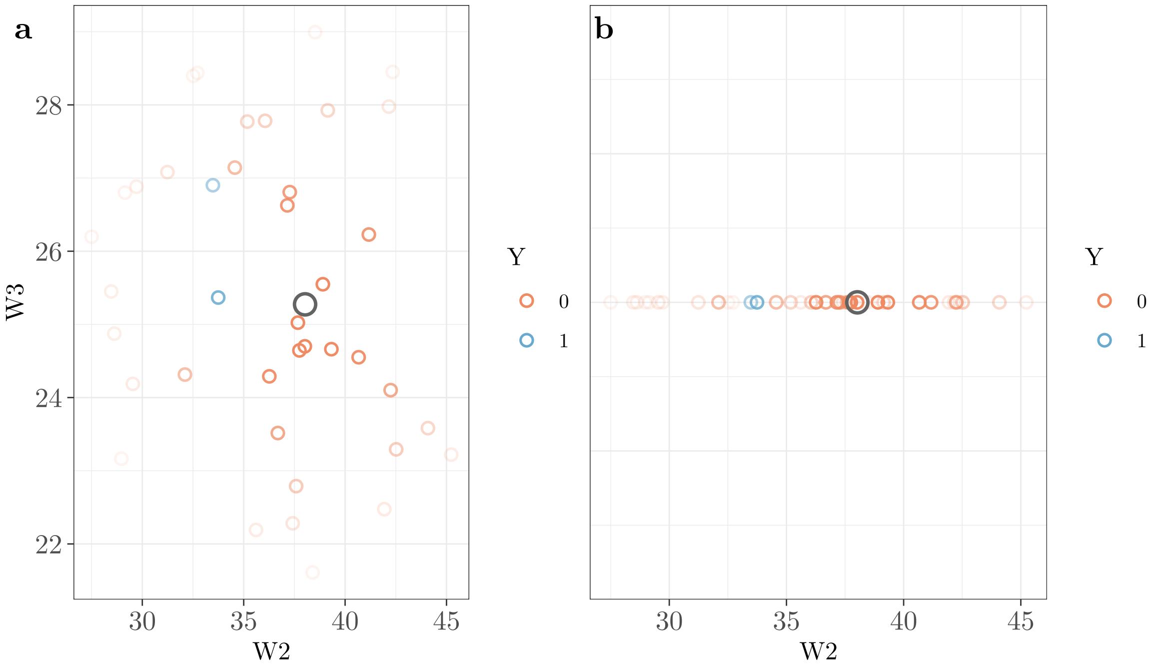

Figure 2.6: 'Local' feature selection

Figure 2.6a focuses on the reduced covariate space of $W' = \{W_2, W_3\}$ where the transparency of each point is proportional to the weight given to that tuning set observation (where weights are derived from the kernels and distances above). Figure 2.6b demonstrates the collapsing of the space of W' into a further reduced covariate space resulting from 'local' feature selection, $W''^{test} = \{W_2\}$.

Prediction

Finally, within this relatively small neighborhood, using a neighborhood-specific set of covariates, simple weighted parametric prediction functions are used to predict the outcome for the observation of interest. The list of prediction algorithms fit at this step would be considered a tuning parameter and could include, for example, the weighted mean, weighted GLM, weighted GLM with pairwise interactions, or weighted GLM with pairwise interactions and quadratic terms.

Figure 2.7 demonstrates this process in a simple scenario where only a single covariate was retained by the local feature selection process. The points are vertically jittered for visibility around $Y = 0$ and $Y = 1$. The transparency of each point again corresponds to that observation's weight, determined by its distance from the

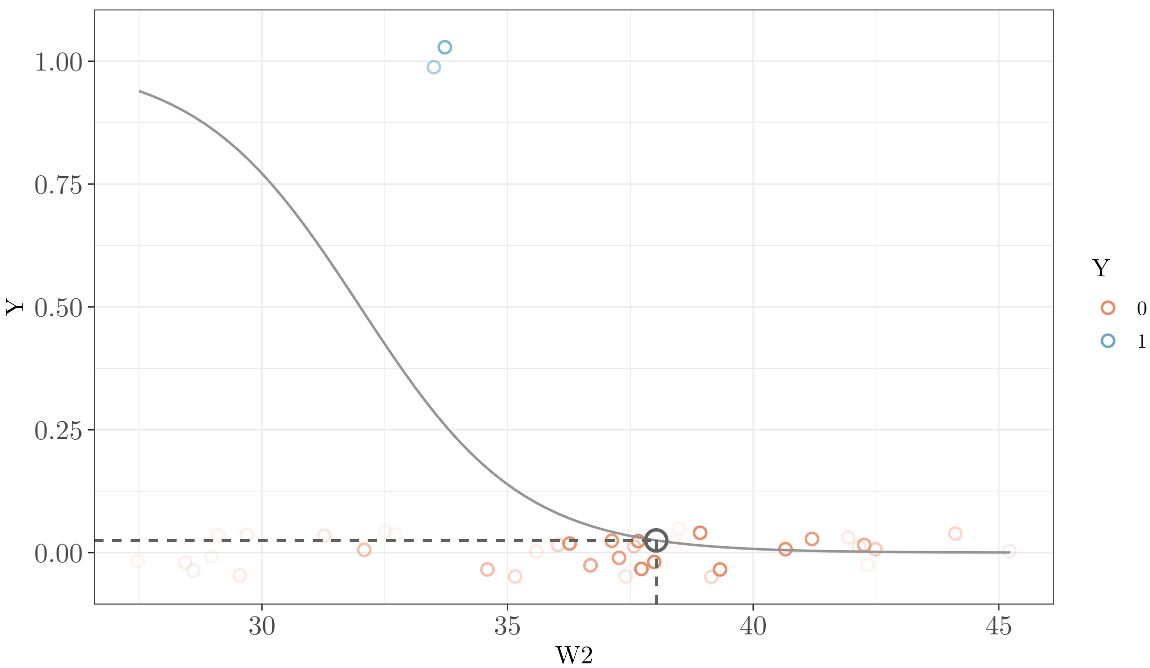

Figure 2.7: Prediction

observation of interest (overlaid in gray). A simple logistic regression line is fit to the tuning set data to obtain a predicted value for the test observation.

Summary

Several considerations went into designing this data-adaptive procedure. If we assume that the optimal prediction function is not static across the covariate space, then it might also stand to reason that the features which are important may change, too. Therefore, while global feature selection is performed, an additional localized feature selection step is also performed, with the goal of eliminating features irrelevant to a particular observation's prediction. Similarly, globally optimized predictions may not be sufficient if the optimal prediction function is not static across the covariate space, and so an appropriate loss function for local optimization must be used. Overfitting is combated via cross-validation, and the large number of tuning parameters are handled via a grid search on the tuning data.

Simulated Trauma Patient Dataset

To demonstrate the performance of the prediction algorithm, a simple trauma patient dataset composed of of 2,000 observations, each with nine 'observed' covariates and a single binary outcome Y, was simulated. Four of these covariates, $\{X_1, X_2, X_3, X_4\}$, were simulated as independent random normals and did not influence the outcome or other covariates in any way. The remaining five covariates are related to the outcome Y and, in some cases, one another.

Data-generating Process

$$
\begin{aligned}
[X_1, X_2, X_3, X_4] &\sim \mathcal{N}_4(\boldsymbol{\mu} = [-2, -1, 1, 2], \boldsymbol{\Sigma} = I_4) \\
\text{GCS} &\sim \begin{cases} \Pr(\text{GCS} = 3) = 0.32, \\ \Pr(4 \leq \text{GCS} \leq 14) = 0.28, \\ \Pr(\text{GCS} = 15) = 0.40. \end{cases} \\
\text{Age} &\sim \ln \mathcal{N}(\mu = 3.56,\ \sigma^2 = 0.44^2) \\
\text{Prehosp. OCY} &\sim \text{Exponential}(\lambda = 0.54) \\
\text{BMI} &\sim \ln \mathcal{N}\left(\mu = 0.08 \times \ln[\text{age}] + 3.01,\ \sigma^2 = 0.21^2\right) \\
\text{Sys. BP} &\sim \text{NB}(\mu = \exp[-0.002 \times \text{age} + 0.016 \times \text{prehosp ocy} + 4.7], \\
&\qquad \theta = 10.9) \\
&\qquad\qquad \text{where} \quad E[X] = \mu \quad \text{and} \quad Var[X] = \mu + \frac{\mu^2}{\theta}. \\
\text{24hr Mortality} &\sim \text{Bern}(\text{logit}^{-1}[1.34 \times \mathbb{1}\{gcs = 3\} - 0.45 \times \mathbb{1}\{gcs = 15\} \\
&\qquad + 0.00012 \times \text{sys. bp} - 0.013 \times \text{bmi} + 0.015 \times \text{age} \\
&\qquad - 0.16 \times \text{prehosp. ocy} - 2.68]) \\
&\qquad\qquad \text{where} \quad \text{logit}^{-1}(\alpha) = \frac{\exp(\alpha)}{\exp(\alpha) + 1}
\end{aligned} \tag{2.3}
$$

Equation Set 2.3 specifies the true data-generating distribution of the simulated data. A directed acyclic graph (DAG) [107] in Figure 2.8 illustrates, graphically, the relationships between the covariates and outcome Y.

In this simple simulated example, each observation can be thought of as a single trauma patient, with the binary outcome indicating survival (`0`) or death (`1`) at 24 hours post-admission. The probability of this event is directly influenced, in

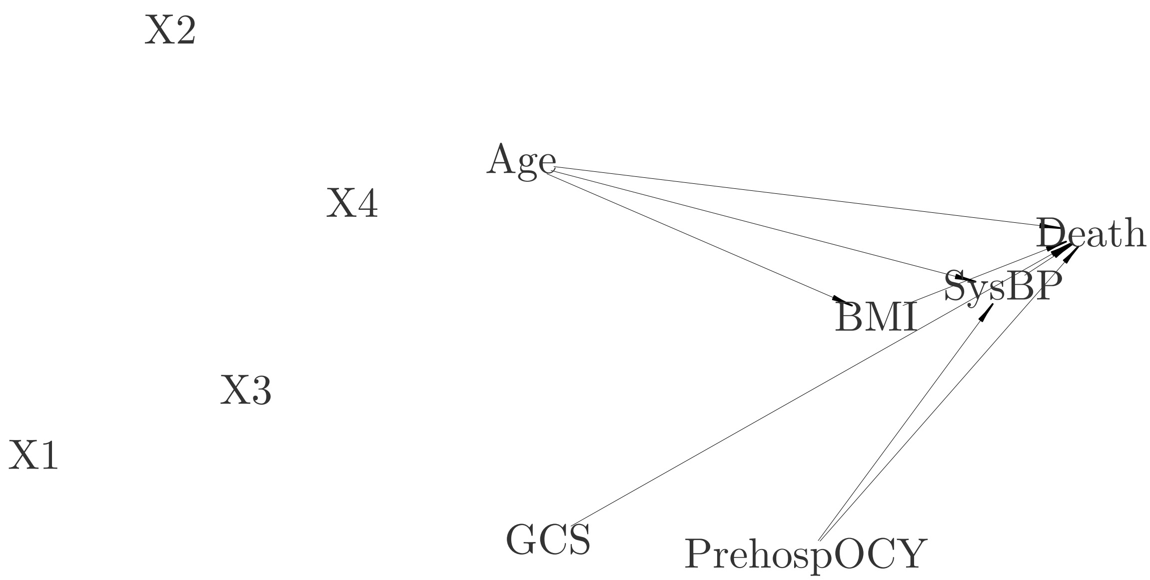

Figure 2.8: DAG representing relationships between variables in simulated trauma patient dataset

this simulation, by the patient's age, BMI, Glasgow Coma Scale (GCS) measurement (binned into three categories), systolic blood pressure (SysBP) at admission, as well as the quantity of crystalloids the patient received prior to hospital admission (PrehospOCY). Additionally, BMI is directly influenced by age, while systolic blood pressure at admission is directly influenced by both age and the quantity of crystalloids received prior to admission.

For example, within this data-generating process, all other covariates held constant, a patient with a GCS of 15 would be more likely to survive than a patient with a GCS between 4 and 14, whereas a patient with a GCS of 3 would be less likely to survive than patients with higher GCS measurements.

Simulated data

Figure 2.9 visualizes the empirical relationships between the outcome Y and the five related covariates in 400 randomly selected tuning set observations. Points and summaries in each panel of the plot matrix are stratified by the simulated binary outcome Y, indicated by color. This matrix of visualizations illustrates more concretely the underlying relationships described in Equation Set 2.3 and Figure 2.8 above. For example, in the scatterplot of BMI versus Age in the fourth row and second column, BMI is observed to increase with increasing age among both outcome groups. How-

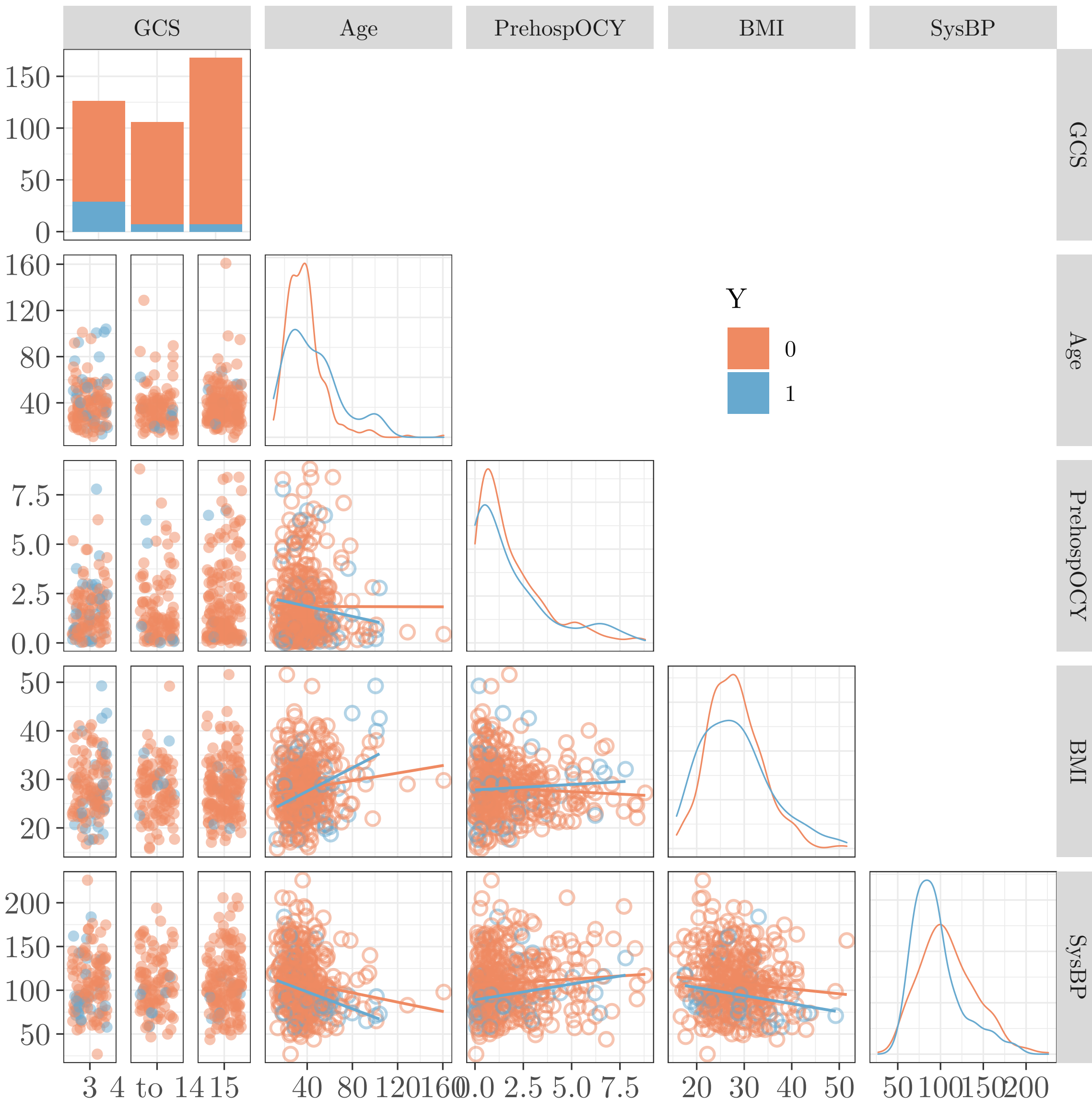

Figure 2.9: Bivariate and univariate simulated covariate visualizations stratified by simulated binary outcome

ever, on average, the older patients who did not survive had higher BMIs than those who did.

Fifteen hundred (75%) of the two thousand simulated observations were used for tuning the local learning model, and the remaining five hundred simulated observations were held out for inclusion in the test set. Subsequently, the model built using the tuning set was used to predict outcomes for observations in the test set. The overall prediction performance of the "best" localized prediction function in the test set was compared with predictions in the test set made by a logistic regression and by the ensemble learner Super Learner. The Super Learner and logistic regression fits were also estimated using the simulated tuning set only.

Results

'Global' Covariate Selection

Prior to ensembling the feature selection results to yield "globally" selected features, the results from each screening algorithm used by CVSL can be visualized separately. In Figure 2.10, the importance estimate from each screening algorithm for each simulated feature is represented as the percentage of CVSL folds in which that feature was retained. The variability across screening algorithms demonstrated in this visualization effectively illustrates the value of ensembling. For example, the random forest retains only GCS, and does so in every fold, but incorrectly discards several features known to influence the simulated outcome, also across all folds. At the other end of the spectrum, the LASSO screening algorithm retained most of the predictive covariates across all cross-validation folds (with the exception of BMI), but also retained two unrelated covariates in every fold, and retained every covariate in at least 40% of folds. Understanding the behavior of the individual chosen screening algorithms on simulated data can help us to tune each algorithm to improve sensitivity and specificity prior to applying the screening algorithm to real-world data.

Variable importance estimates are combined across screening algorithms in Figure 2.11 via the unweighted average of the proportion of folds within screening algorithm where the feature was retained. GCS stands out for having been retained relatively often, and generally (with the exception of BMI), the simulated covariates with a stronger true relationship with the outcome are retained by the ensemble of screening algorithms more frequently than the simulated random, unrelated covariates. As alluded to above, improved tuning of the screening algorithms could improve the ensembled variable importance in this simulation. Additionally, algorithm ensemble weights could be utilized to estimate variable importance via a weighted mean proportion.

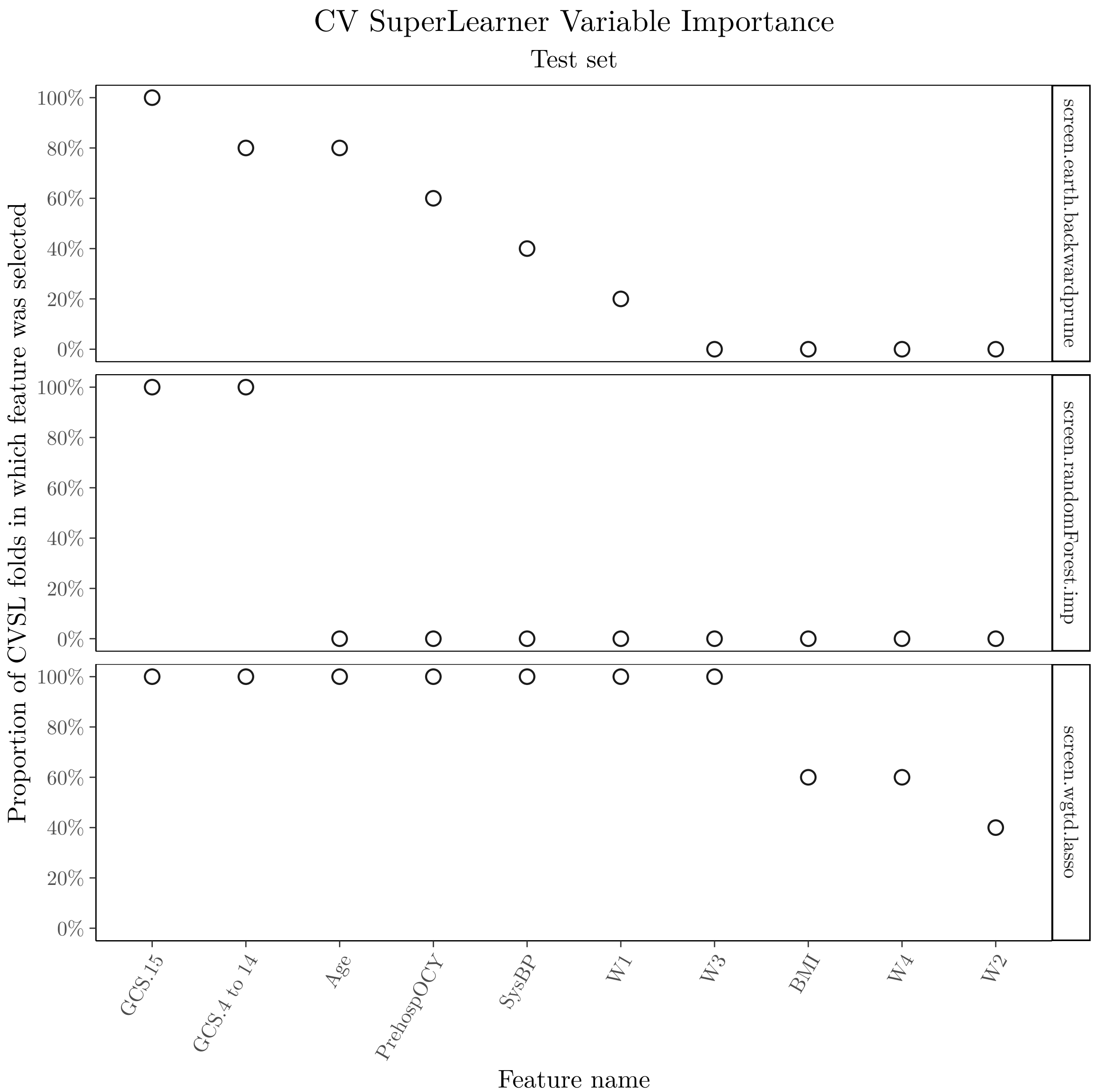

Figure 2.10: Feature selection by screening algorithm

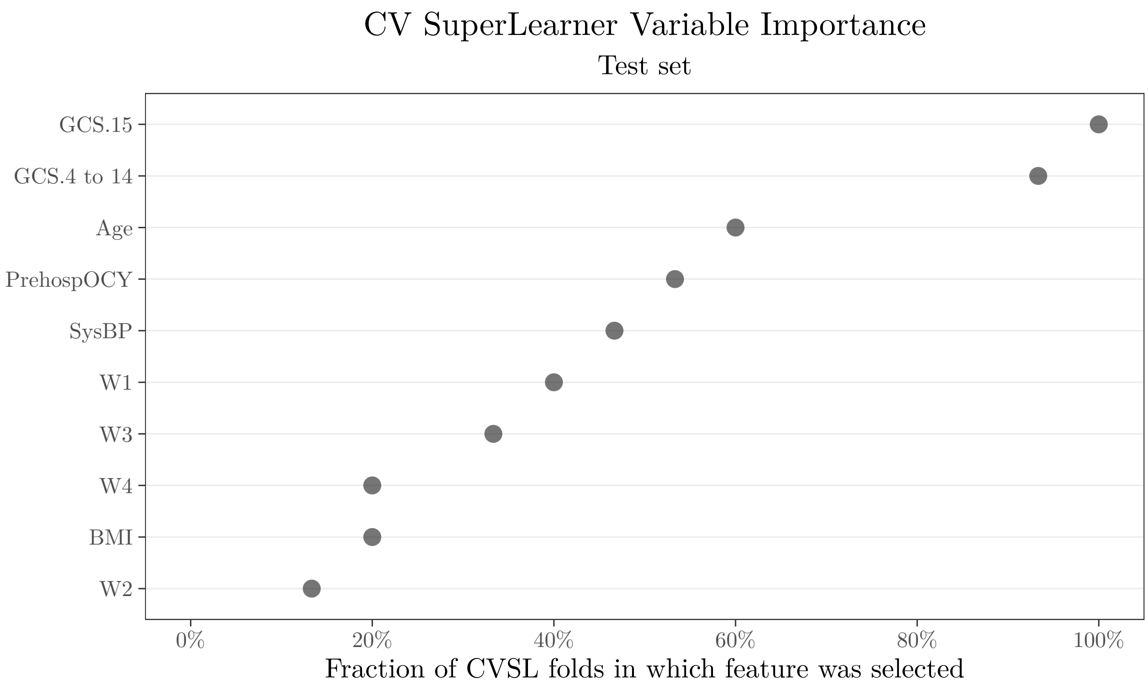

Figure 2.11: Feature selection across all screening algorithms

Performance comparison

The distribution of the predicted probabilities in the test set of the simulated outcome from each candidate prediction algorithm (the local learner, logistic regression, and Super Learner) are visualized in Figure 2.12. Perfect differentiation between outcomes by a prediction algorithm would manifest as predicted probabilities for the "Living" (blue) group stacking up on the left (low) end of the axis, while predicted probabilities for the "Deceased" (orange) group would accumulate on the right (high) end of the axis. Instead, while there are subtle differences between the distributions of the predicted probabilities among the three prediction methods – for example, logistic regression and Super Learner both correctly assigned high probabilities of death to several simulated patients, while the local learner did not – the overwhelming take-away from Figure 2.12 is that the three predictors all confidently and correctly predict low probabilities of death for the majority of the surviving patients, but the predicted probabilities for simulated deceased patients are more uniformly distributed across the range of predicted probabilities and are therefore less discriminating.

The accuracy of the local learner on the simulated test set over a range of cut-

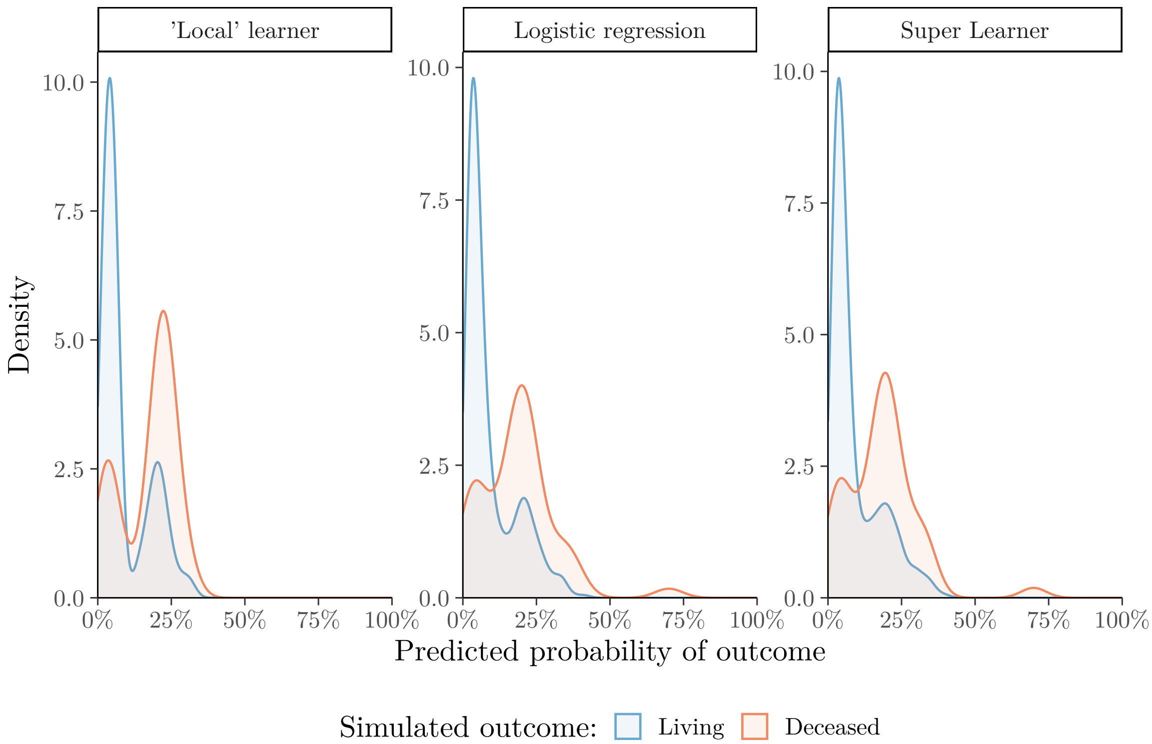

Figure 2.12: Prediction density comparison for simulated PROPPR data (test set only)

points applied to predicted probabilities is visualized in the green line of Figure 2.13. Instead of having a clear peak, the maximal accuracy in the test set (90.2%) is achieved at the maximum predicted probability, indicating that the best "overall" classifier performance is achieved when every simulated patient is predicted to have survived, such that the true negative rate (TNR) is 100% and the true positive rate (TPR) is 0%. However, in practice, and particularly in a real-time patient care setting, a different balance of true positives and true negatives is preferable. This balance can be selected via the orange (TNR) and purple (TPR) lines in Figure 2.13. If both are equally important to optimize, for example, one might choose a classification cutpoint near where the two lines cross (a predicted probability of approximately 13.3% in this simulated example), such that simulated patients with predictions above that threshold would not be predicted to survive, while simulated patients with predictions below it would.

Overall prediction performance in the simulated test set is compared in Figures 2.14 and 2.15 between the "best" localized prediction function (local learner), the

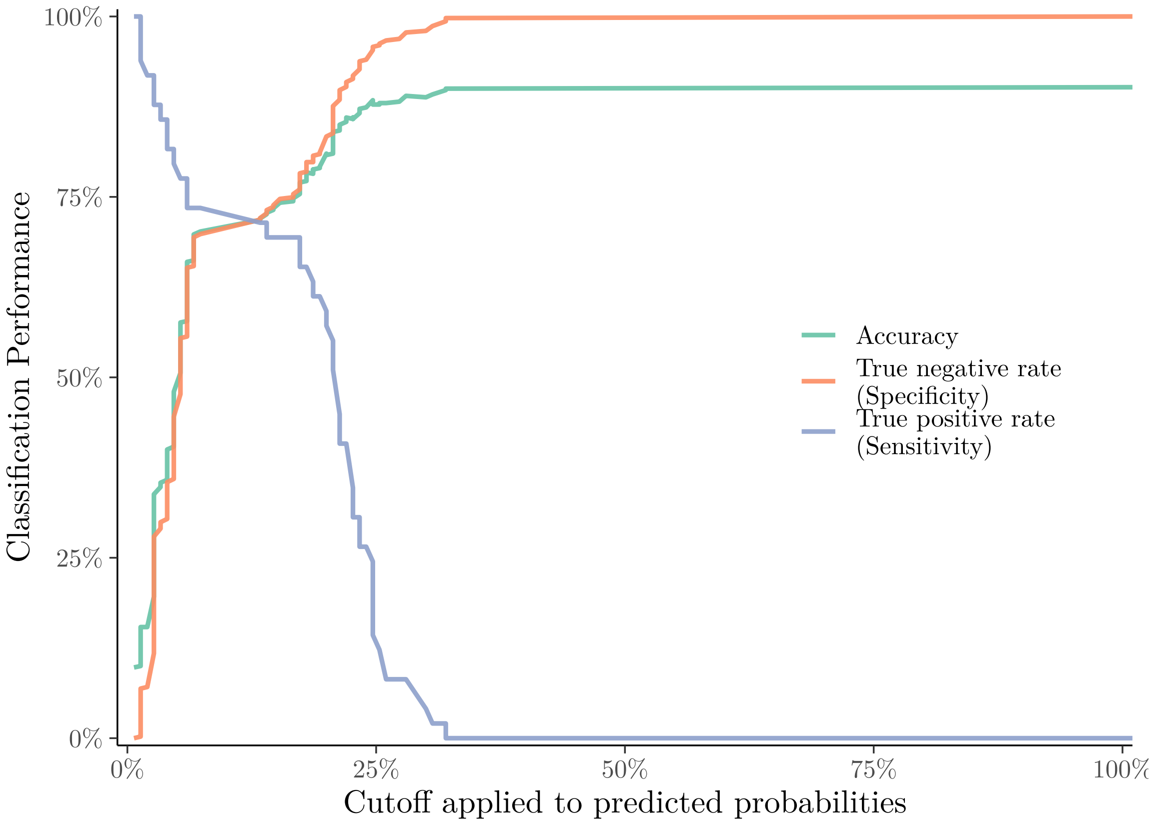

Figure 2.13: Local learner classification accuracy by predicted probability cutoff (test set only)

ensemble learner Super Learner, and a logistic regression. The estimates of area under the ROC curves in the test set are compared in Figure 2.14, the best of which belongs to the localized algorithm (AUC = 0.733). In particular, for a cutoff where survival is *correctly* predicted at a high rate (i.e. FPR $<$ 25%), the local learner correctly predicts death at a higher rate than the comparison algorithms. In comparison, at a cutoff where simulated surviving patients are often incorrectly predicted to have died (i.e. FPR $>$ 50%), the logistic regression and Super Learner more frequently correctly predict patient deaths than the local learning algorithm does.

However, the 95% confidence intervals of these AUC estimates indicate that the three modeling strategies don't significantly differ in their predictive performance in the simulated dataset. It is not unexpected that the local learning algorithm wouldn't dramatically outperform competing algorithms, given that the data gener-

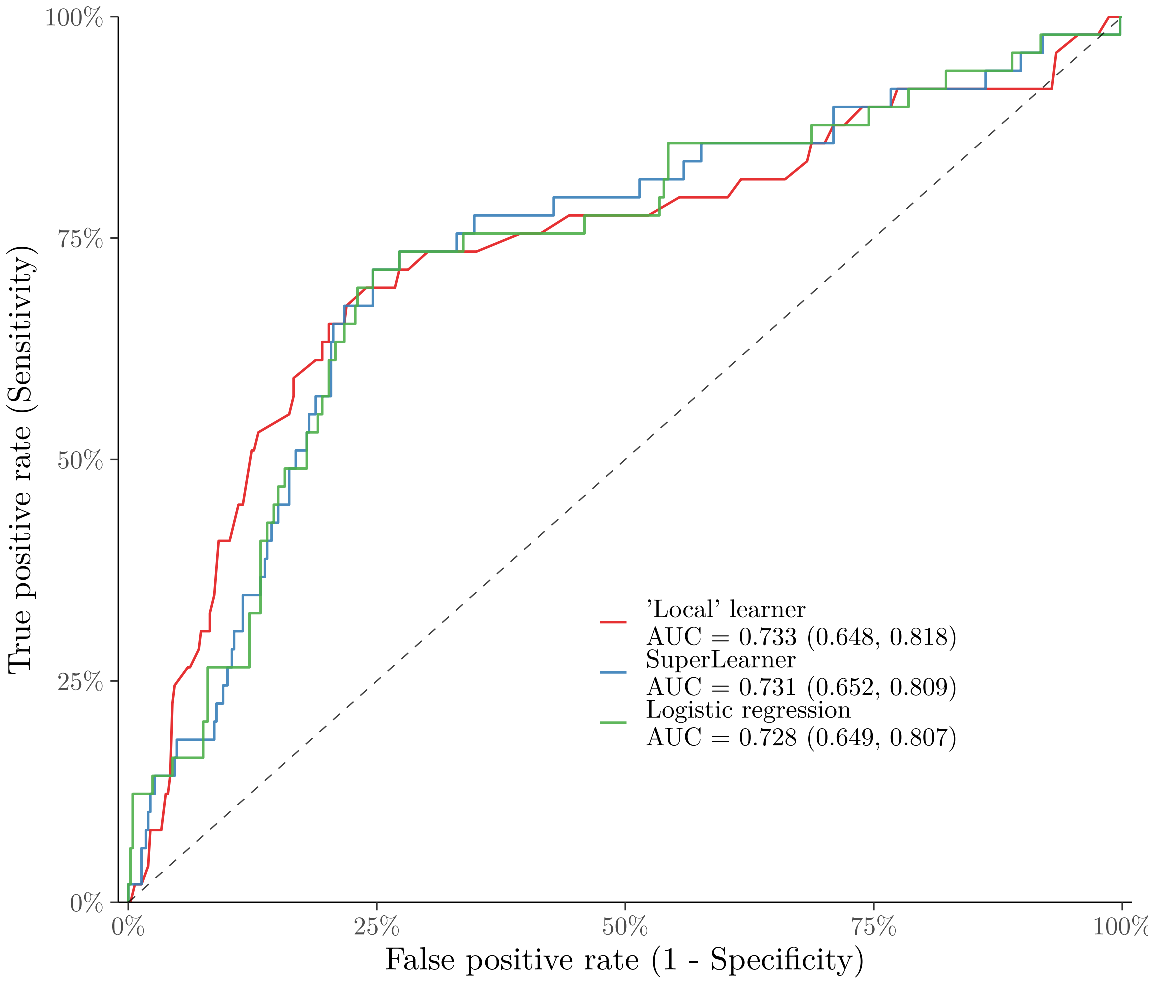

Figure 2.14: Receiver operating characteristic curve comparison (test set only)

ating functions are not designed to simulate a clustered dataset, the scenario under which the local learner would be expected to most excel.

The precision-recall curves in Figure 2.15 shed more light on each algorithm's performance and are particularly useful to highlight the minority class (patients who did not survive) in this simulation containing a somewhat imbalanced outcome. For example, the local learner falsely predicted many simulated patients would not survive in order to achieve even a relatively low rate of correct identification of simulated patients who did not survive. On the other hand, the logistic regression's area under the precision recall curve exceeds that of the local learner and the Super Learner, mostly due to it casting too small a net in predicting death (low probability

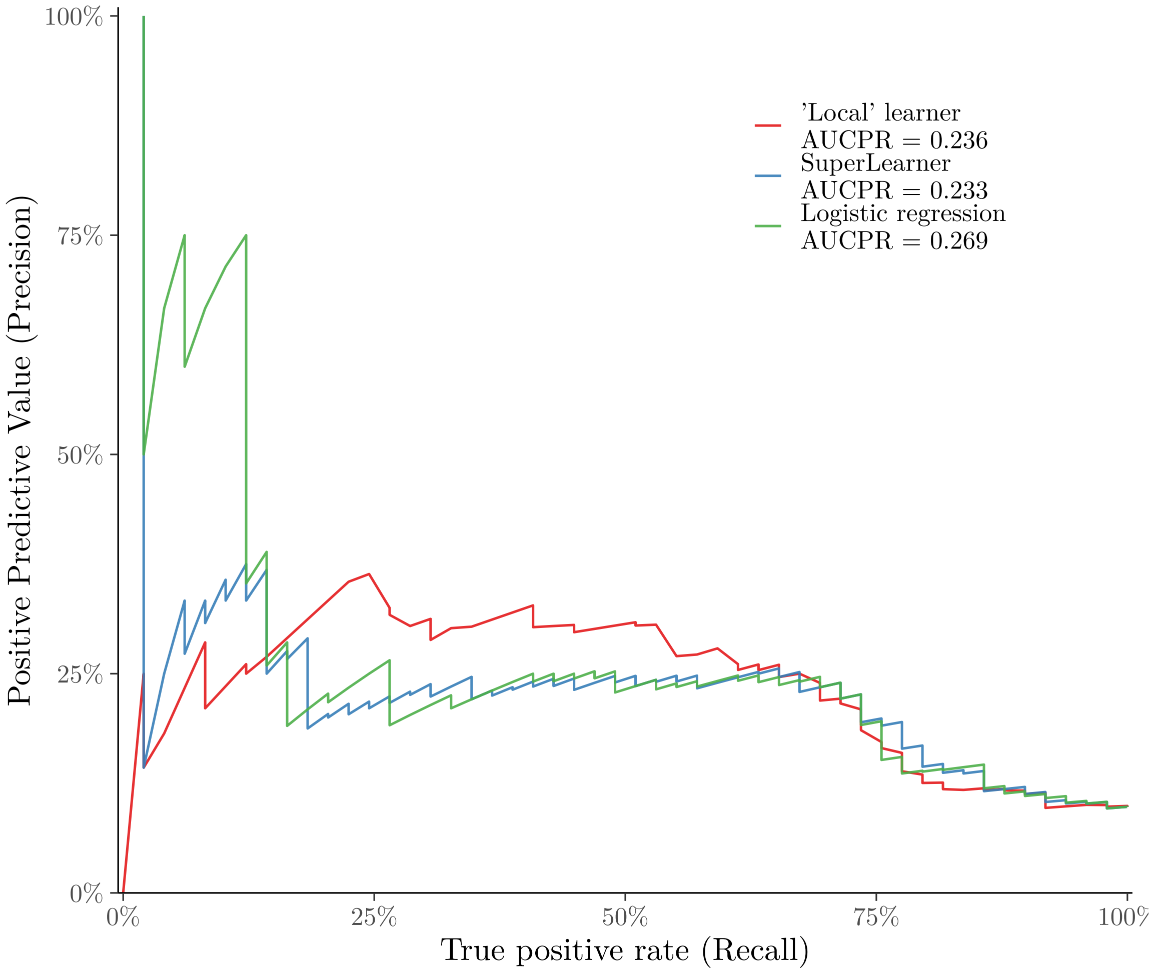

Figure 2.15: Precision-recall curve comparison (test set only)

the logisitic regression would predict death given the patient died) in order to achieve a higher proportion of patients who did not survive among those who were predicted to not survive. However, at classification cutoffs casting a wider net in flagging patients at risk for death (i.e. with a lower rate of misdiagnosis among those who didn't survive), as expected, the proportion of patients who did not survive among those predicted to not survive is much lower.

Conclusion

The simulation study described in the current chapter demonstrates that the local learner can perform competitively with existing prediction methods. However, real-world trauma patient data is significantly more complex than its simulated counterpart. This complexity is evidenced by past attempts at developing scoring systems and predictive models for this population, the details of which will be described in the first half of the upcoming chapter. Due to this complexity and the potential for, e.g., patient clusters and other patterns to arise (due to the true data generating process) that are not handled equally well by all modeling approaches, it is critical that any proposed new methodology also be compared with existing "gold standard" approaches utilizing real-world, highly complex patient data. This exercise will be the focus of the second half of the following chapter.

Chapter 3

Data study

Real-world Data

To test the real-world predictive performance of our algorithm, we utilized trauma patient data from the **Pragmatic, Randomized Optimal Platelet and Plasma Ratios** (PROPPR) study [52], a randomized clinical trial comparing two blood transfusion protocols. The study involved 680 patients at 12 U.S. Level 1 trauma centers. Patients qualified by both receiving blood products and triggering the highest trauma level activation. The ratio of plasma to platelets to red blood cells received in the two randomized groups was approximately 1:1:1 ("whole blood" or "balanced;" $n_{1:1:1} = 338$) or 1:1:2 ("unbalanced;" $n_{1:1:2} = 342$).

Outcome (Y)

In the example presented in this chapter, the clinical outcome of interest (Y) was all-cause death by 24 hours after randomization to treatment group. Eighty-five percent of patients enrolled in PROPPR remained alive 24 hours post-randomization. Other outcomes of interest in the PROPPR study that won't be points of focus in this text included death due to exsanguination or hemorrhagic shock at any time after randomization.

Covariates (W)

To predict the outcome above, we included measures established prior to randomization as well as results from labs for which blood was drawn very early, usually just minutes after beginning the treatment protocol. These covariates (W) included basic demographics, treatment information, injury type, primary vital signs, trauma scoring systems, lab measures from the initial blood draw, and pre-randomization fluid and blood products received.

Below is a full listing of the patient information utilized in the prediction of all-cause death by 24 hours post-randomization. Covariates in **bold** were fully observed. Any not in bold were included along with an additional (non-)missingness indicator. Values of categorical covariates displayed in *italics* served as the reference levels of those covariates.

- **Study Site**: Baltimore, Birmingham, Cincinnati, Houston, Los Angeles, Memphis, Milwaukee, Portland, *San Francisco*, Seattle, Toronto, or Tucson;
- **Age**; **Gender**; BMI; **Hispanic ethnicity**;
- **Racial Group**: African-American, Asian, Native American, Multiple Races, Pacific Islander, Unknown, or *White*;
- **ANY Blunt Injury**; **ANY Penetrating Injury**;
- Systolic Blood Pressure (SBP); Diastolic Blood Pressure (DBP); Pulse; Respiratory Rate (RR); Temperature;
- Results of **Focused Assessment with Sonography in Trauma (FAST) exam**: Positive, *Negative*, Indeterminate, or Not Done;
- **Glasgow Coma Score (GCS)**; **ABC Score**;
- Platelet count; White blood cell (WBC) count; Hematocrit (HCT); Hemoglobin; Base excess; pH;
- Thromboelastography (TEG) reaction time (R-Time); prothrombin time (PT); partial thromboplastin time (PTT); International Normalized Ratio (INR); Fibrinogen;
- **Pre-randomization volumes of: Red Blood Cells (RBCs), Plasma, Platelets, Crystalloids (OCY), Colloids (OCL), Other OCL/OCY**;
- Pre-randomization Plasma:RBC Ratio; Platelet:RBC Ratio;
- **Time to Randomization**; **Treatment Group**.

Notably, the Revised Trauma Score (RTS) was not employed as a predictor, as it served as a comparison score against which the algorithm's performance was judged. In practice, the RTS would likely be calculable for most trauma patients during assessment, as it only depends on its components (GCS, SBP, and RR) having been measured. Moreover, one or more of those components having *not* been measured could also aid in the prediction of certain clinical outcomes. Therefore, the RTS could feasibly be included as a predictor in a real-world setting and would have the potential to boost the performance of a highly parsimonious prediction model.

Methods

The sample of 680 patients was randomly partitioned into tuning (75%, $n_{tuning} = 510$) and test (25%, $n_{test} = 170$) sets, with outcome distributions approximately balanced in the two groups (mortality rates were 16% and 12% in tuning and test, respectively).

Five-fold cross-validation was utilized to avoid overfitting to the tuning set. Within each fold, a 'global' feature selector was first applied. Then, the local learning algorithm predicted the probability of the outcome for each observation, also within each fold, by first weighting observations by their similarity to that observation and subsequently using those weights in the application of a simple linear predictor. The details of each step are governed by one or more tuning parameters.

The local learning algorithm's tuning parameters were optimized by performing a grid search over many unique combinations of parameter settings while applying the algorithm to the tuning set. The 'optimal' combination of tuning parameter settings were defined as those resulting in the minimum cross-validated sum of squared prediction errors (cvSSE) within the tuning set. The cvSSE was assessed separately for the global feature selection algorithm and the local learning algorithm. These 'optimal' tuning parameter settings were used to generate a final local learning algorithm fit (without cross-validation) using the entire tuning set. This predictive model was subsequently applied to the test set to predict the probability of the outcome of interest for each of the n_{test} patients.

The local learner's prediction performance in the test set was compared with that of a logistic regression fit on the tuning set, along with that of a Super Learner fit on the tuning set, both likewise applied to the test set. The Super Learner predictor library included a selection of learners supplied with the SuperLearner R package [65]: logistic regression (with and without interaction terms), LASSO regression [112, 36], random forest [17, 70], and multivariate adaptive regression splines [37, 75].

The predictions from these machine learning methods were also compared with predictions for patients in the test set derived from the RTS trauma scoring method [22]. The RTS, as detailed above, is a function of the patient's SBP, RR, and GCS (a measure of how conscious a patient is). Champion et al. [22] provides, in Table V, a mapping from RTS values to survival probability estimates via a logistic function. Here, that logistic function was approximated using the provided information such that the RTS estimates presented below are one minus those survival probability estimates.

Seventy-three of 680 PROPPR patients were missing SBP and/or RR. In those cases, the RTS was calculated using all available information, and when information was missing, the RTS was imputed as the worst value possible given other observed information. Reasons for missing SBP and RR were that they couldn't be measured or that they weren't taken, both of which could be attributable to a patient being in critical condition when their treatment began. Therefore, the assumptions made by this imputation strategy seemed reasonable.

Results

'Global' Covariate Selection

Figures 3.1 and 3.2 visualize the variable importance estimates resulting from the application of the initial 'global' covariate selection step to the test set. Feature importance was estimated here via five-fold Cross-Validated SuperLearner (CVSL) [65] using three screening algorithms: observation-weighted LASSO [112, 36]; random forest [17, 70] using a data-adaptive variable selection cutoff [100]; and case-weighted multivariate adaptive regression splines [37, 75], using backward pruning and without cross-validation. These three screening algorithms are available in the `SLScreenExtra` R package [87]. Additional technical details on these screening algorithms are provided in the previous chapter.

Figure 3.1 conveys the importance estimate for each feature within each screening algorithm as the proportion of CVSL folds in which the feature was retained by that screener. Here, we observe a large amount of variability across screening algorithms, with the random forest retaining only pH in four of five folds and only BMI in the remaining fold, while the lasso and spline screeners retained many more features in each fold, but the retained features retained differ somewhat between the two screeners.

These screener-specific variable importance estimates are combined in Figure 3.2 by taking the unweighted average over screening algorithms of the proportion of folds within each where the feature was retained. GCS and pH stand out for having been retained relatively often, whereas many of the study site indicators along with some of the missingness indicators are rarely, if ever, observed to be retained by the screening algorithms.

For simplicity, only the unweighted average of screening algorithms' feature importance estimates was included in the tuning parameter grid search for this example. However, an alternative that could have been included in the grid search is the weighted average over screening algorithms of the proportion of folds within each where the feature was retained, where the weights would be derived from Super Learner's coefficient estimates for each combination of screening and prediction algo-

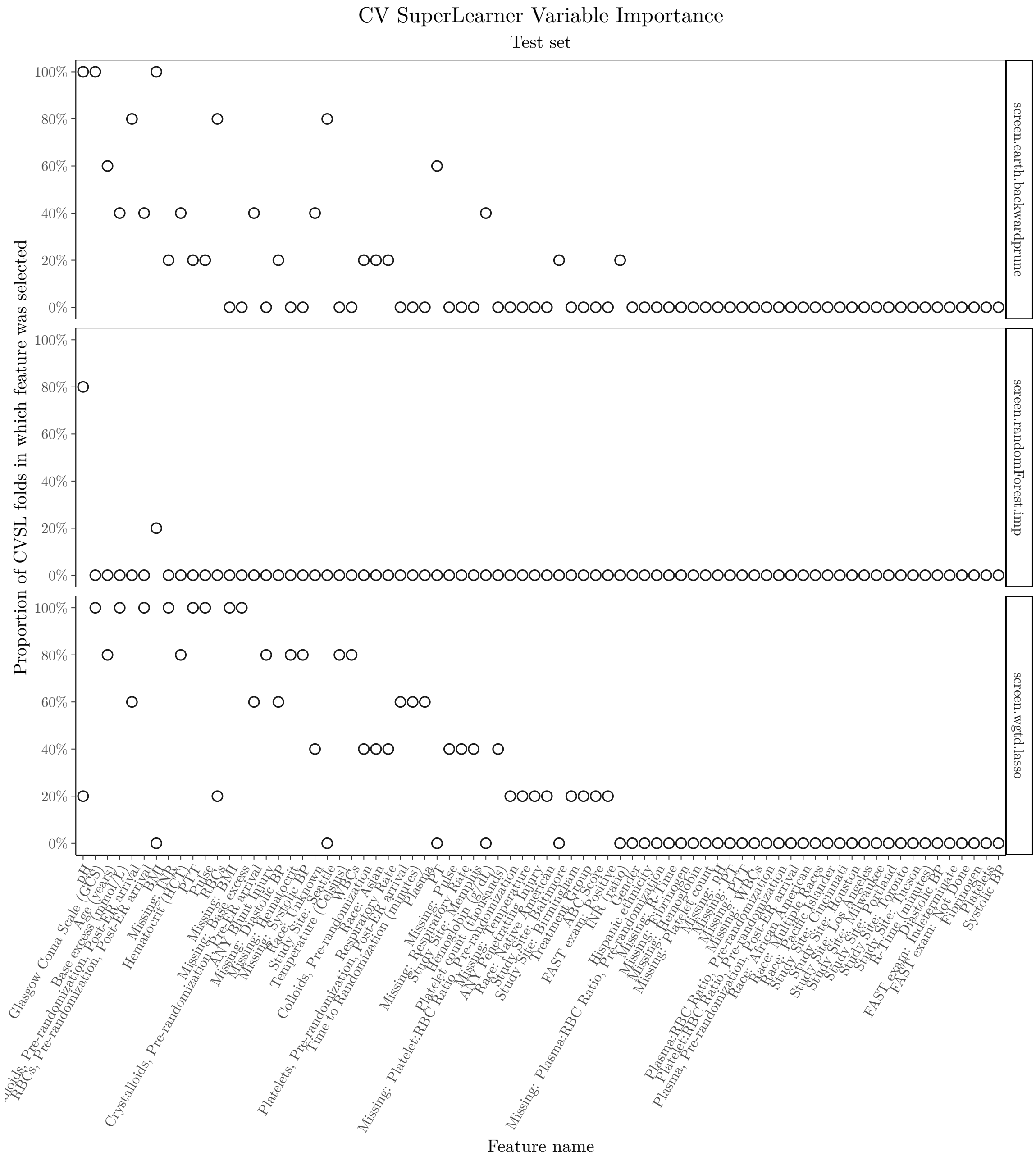

Figure 3.1: Feature selection by screening algorithm

rithms in the Super Learner library. Building on this, when including the weighted feature importance in the grid search, the method by which Super Learner coefficients are estimated (e.g. non-negative log-likelihood maximization, non-negative least squares minimization, or AUC maximization/rank loss minimization [97]) may also be included as a tuning parameter, the 'optimal' setting for which could then be selected via grid search.

The tuning parameters for 'global' feature selection which were searched over in this example were the metaselector parameters controlling what subset of the available features were selected by the 'global' feature selector, the options being:

- the 20 most highly ranked features,
- the 40 most highly ranked features,
- the features chosen in at least 5% of the [5 folds $\times$ 3 screening algorithms] (note that this is equivalent, in this case, to the feature being chosen at least once over all 5 folds $\times$ 3 screening algorithms), and
- the features which are separated from the others by the largest difference in estimated importance.

In the tuning set, the setting above which minimized the cvSSE was the final one: a subset of features was optimally chosen by determining which features were most separated from the others. Applying this selection method to the importance estimates in Figure 3.2, for this run of the CVSL feature importance estimation, only GCS and pH would be retained.

Local Learner

Features retained by the initial selection step are subsequently used to create a weighted 'neighborhood' of observations around each observation for which a prediction is desired. Several tuning parameters govern this process:

- A distance function: here, taxicab (L1) or Euclidean (L2);
- A kernel function: tricube, uniform, or Epanechnikov; and
- A window width or 'neighborhood' size: 10%, 15%, 20%, or 25% of the observations in the remaining sample.

Finally, prediction estimates are generated by combinations of screener and learner algorithms, pairings which are also treated as tuning parameters (where only one is selected, much like discrete Super Learner). The estimated weights from the previous step can be used by the screener and learner algorithms in this step. Three options were supplied to the grid search in this example:

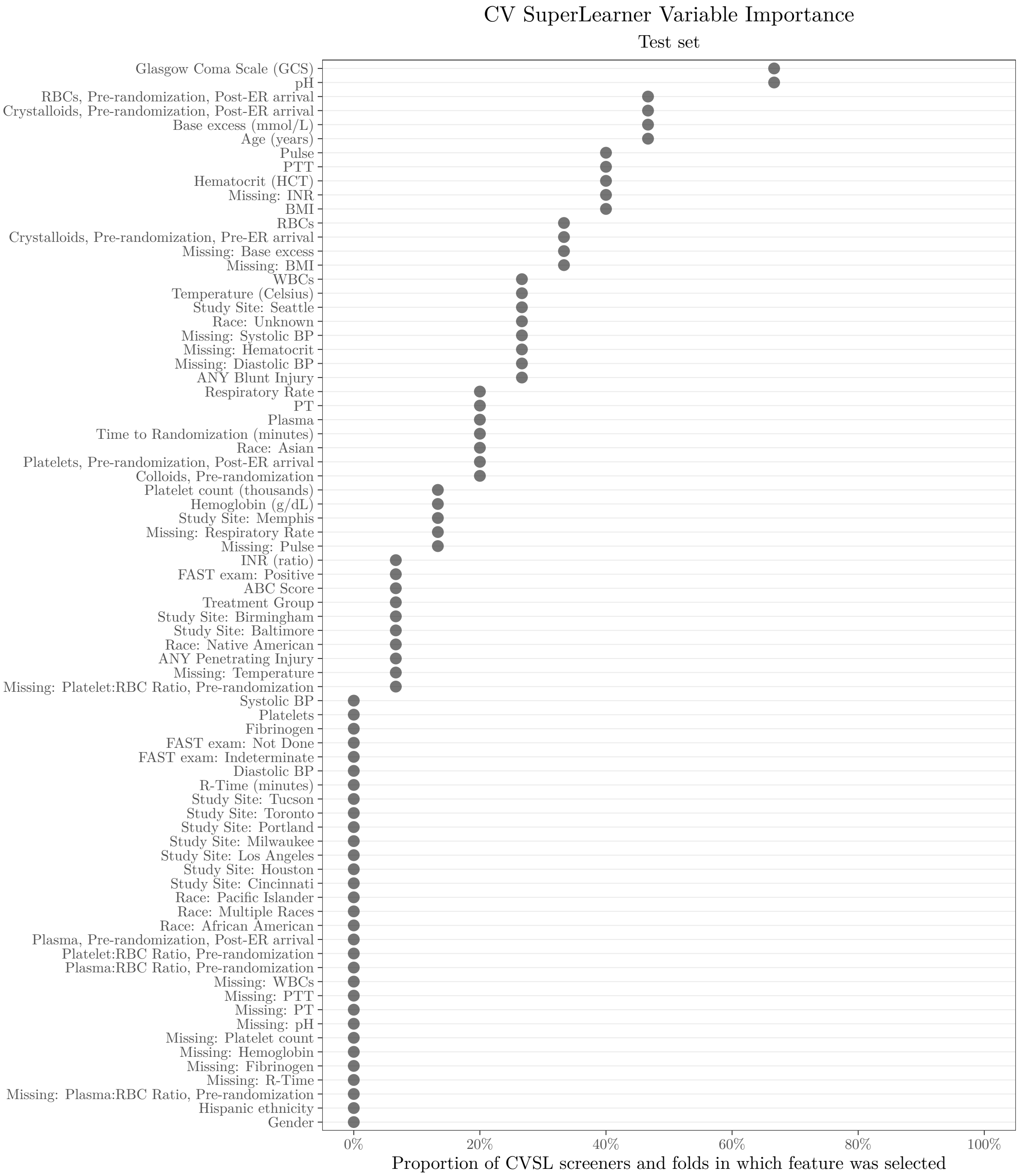

Figure 3.2: Feature selection across all screening algorithms

- weighted mean (screening algorithm immaterial),
- weighted GLM using, at most, the top 5 covariates in the 'neighborhood' (the ranking of which was determined by the p-values of Spearman's rank correlation coefficient estimates), and
- weighted GLM using, at most, the top 5 covariates in the 'neighborhood', plus pairwise interaction terms.

In the tuning set, the combination of tuning parameters minimizing the cvSSE were the L1 distance combined with the tricube kernel with a 10% window width and a weighted mean.

Performance comparison

Figure 3.3 visualizes the distribution of the predictions from each candidate prediction algorithm (or score). If a prediction algorithm had perfectly differentiated between the two groups of patients formed by their observed outcomes, the predicted probabilities for the "Living" (blue) group would be clustered on the left side of the plot and the predictions for the "Deceased" (orange) group would be clustered on the opposite end, signifying a high predicted probability of death. In contrast, Figure 3.3 indicates that all four predictors do a decent job of correctly assigning low probabilities of death to patients who survived – with the possible exception of the RTS – but that none of them does a particularly good job of correctly classifying the patients who did not survive.

Figure 3.4 more precisely identifies, for the local learning algorithm, the cutpoint for predicted probabilities at which the algorithm's classification accuracy is maximized. Setting the cutpoint higher than a predicted probability of approximately 30% would result in very little accuracy gained. This assessment echoes what is implied in the upper left panel of Figure 3.3 where a probability of $\leq 30\%$ was predicted for the great majority of the patients who were observed to survive.

In Figures 3.5 and 3.6, the overall prediction performance of the "best" localized prediction function in the test set was compared with the test set predictions of the ensemble learner Super Learner, a logistic regression, and the RTS. Super Learner and the logistic regression were fit using the tuning set only.

It is clear that, for the outcome of death in the first 24 hours after randomization, Super Learner performs best in the test set as measured by the area under the ROC curve displayed in Figure 3.5. The performance of the localized algorithm and the imputed RTS fall in the middle, and the performance of the logistic regression is the poorest of the three. However, these algorithms don't significantly differ in

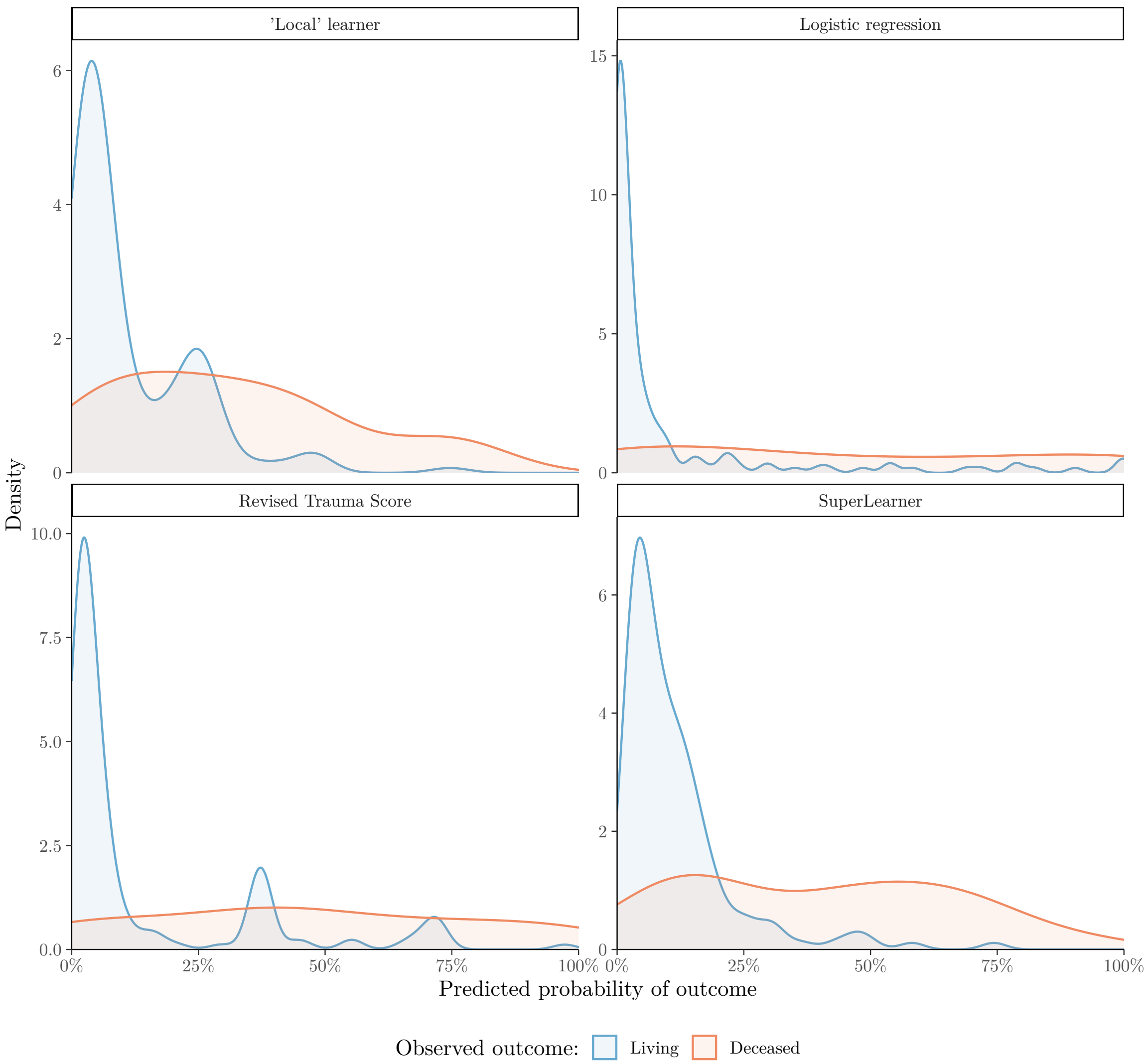

Figure 3.3: Prediction density comparison for death by 24h post-randomization among PROPPR patients (test set only)

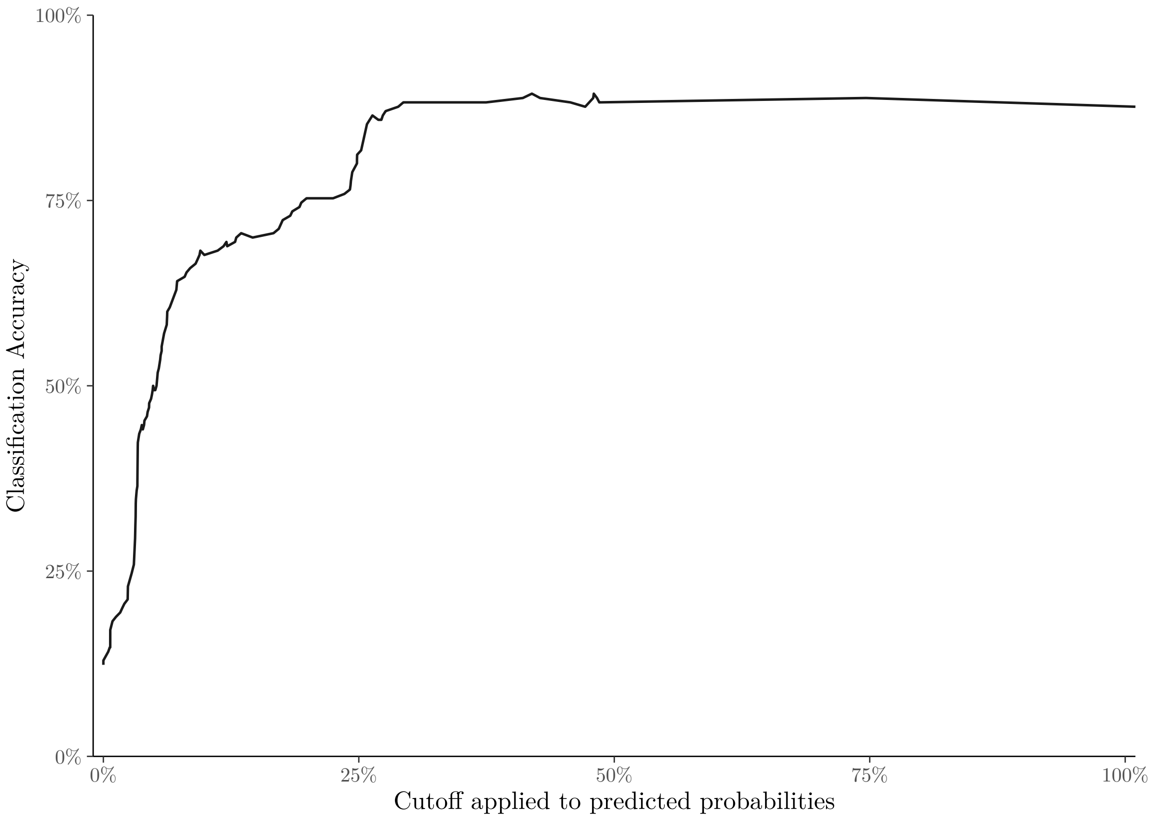

Figure 3.4: Classification accuracy by predicted probability cutoff (test set only)

their predictive performance according to the 95% confidence intervals of the AUC estimates.

If these predictions were utilized in practice to determine a patient's priority, false negative classifications might correspond to patients who would be undertriaged via this algorithmic approach, while patients who are false positives may have been overtriaged. According to Figure 3.5, a 35% overtriage (false positive) rate would correspond to a 19% undertriage rate (false negative rate, or 1 - sensitivity) for the local learning algorithm, while a more stringent 5% undertriage rate would correspond to a 62% overtriage rate.

In some cases where the binary outcome is imbalanced, the precision-recall curve may reveal performance issues that are hidden by the traditional ROC curve. However, here, the ranking established by the area under the ROC curves is echoed in Figure 3.6 when comparing the area under the precision-recall curves for the four prediction methods applied to the test set. Super Learner achieves the highest AUCPR,

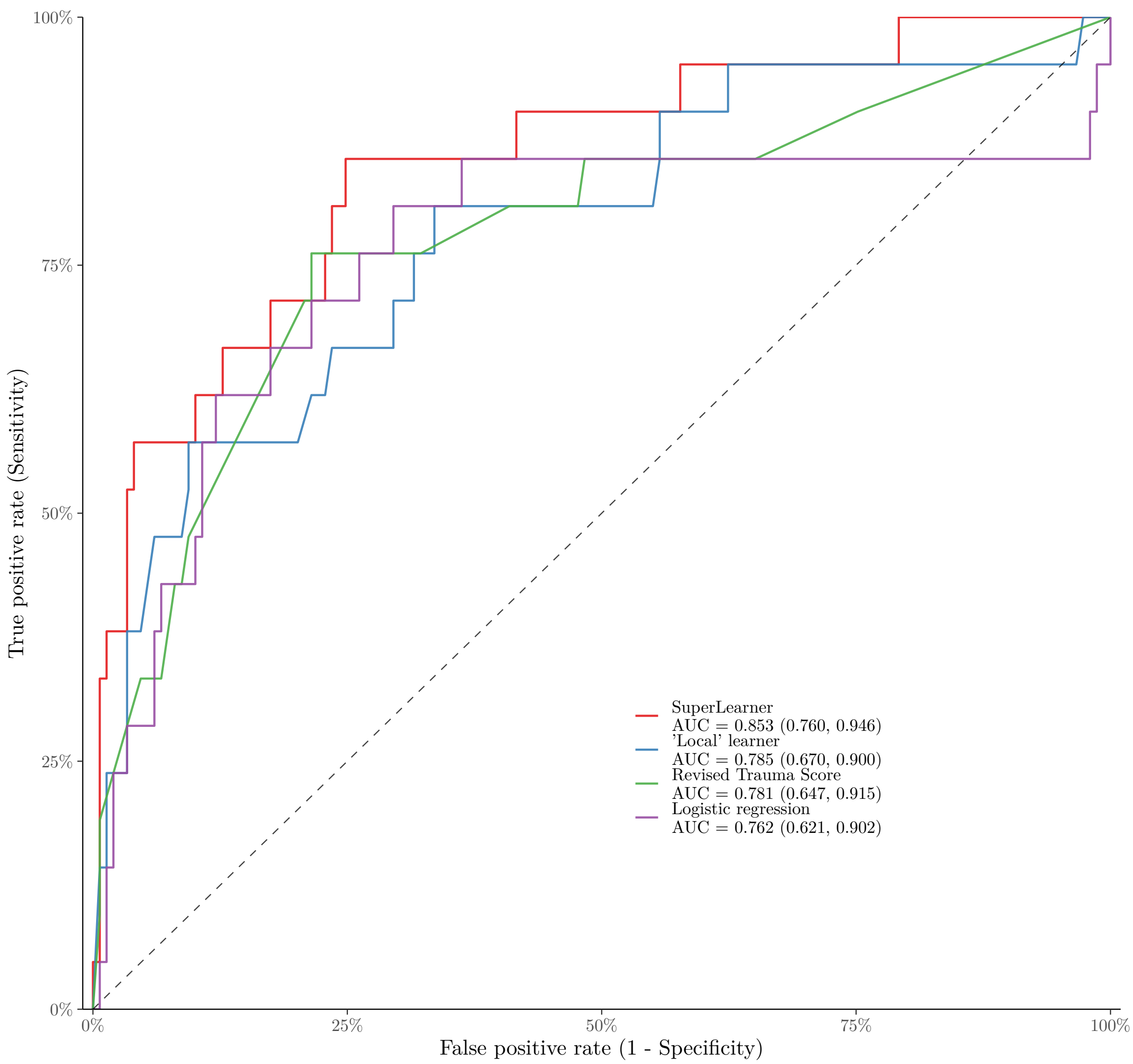

Figure 3.5: Receiver operating characteristic curve comparison (test set only)

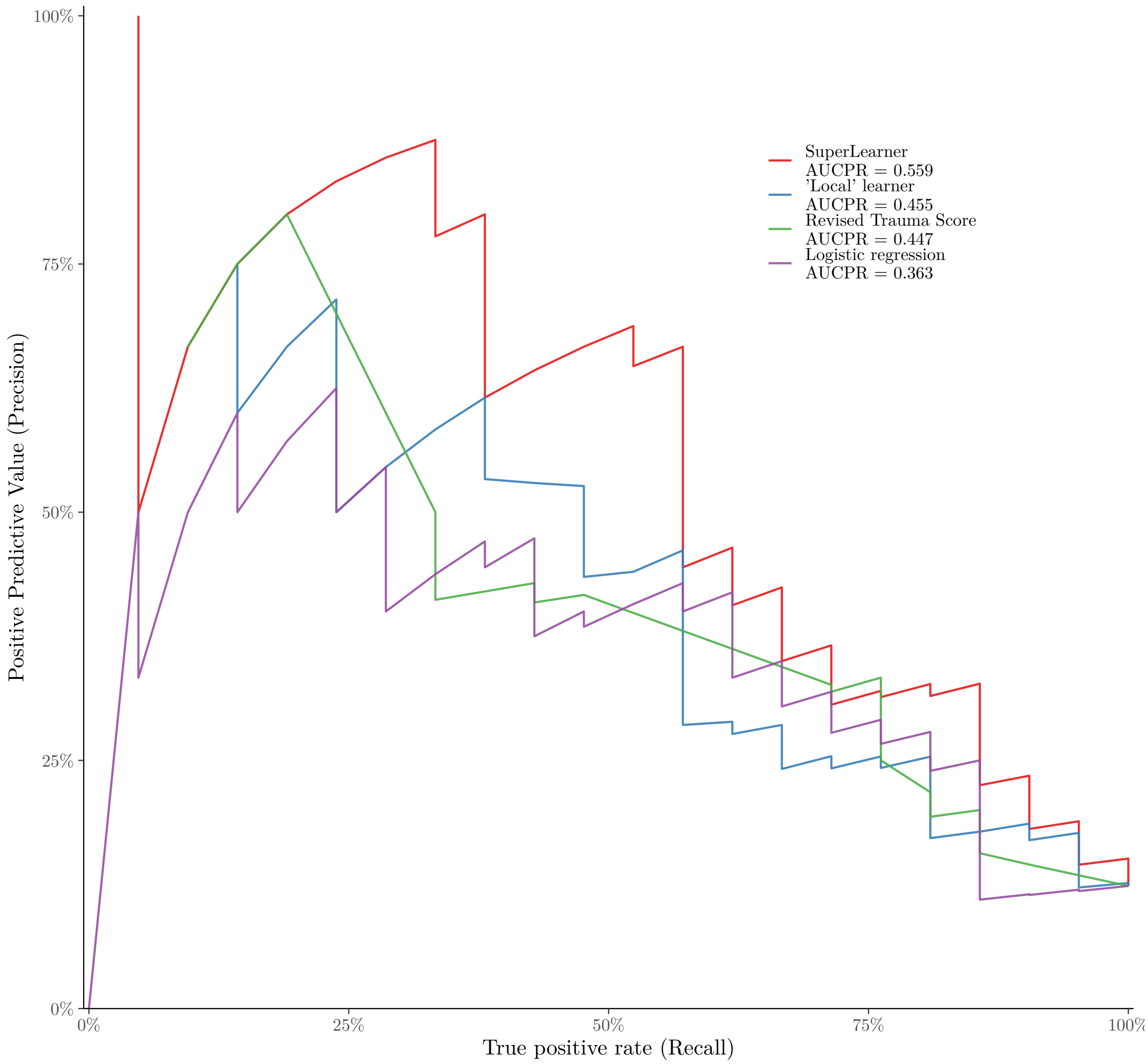

Figure 3.6: Precision-recall curve comparison (test set only)

followed by the local learner and the RTS, with logisitic regression lagging behind.

Table 3.1: Example Local Regression Model Coefficient Estimates

	Patient A			Patient B		
Characteristic	**OR**	**SE**	**95% CI**	**OR**	**SE**	**95% CI**
(Intercept)	90.4	1.99	3.35, 1.08e+04	2.67	35.0	0.00, 3.26e+27
pH	0.74	0.178	0.50, 1.03	0.39	4.66	0.00, 1.43e+04
Glasgow Coma Scale (GCS)	0.79	0.113	0.60, 0.96			
Crystalloids, Pre-randomization, Post-ER arrival	0.52	0.535	0.16, 1.42			
Missing: BMI	0.12	1.15	0.01, 0.96			
WBCs	0.97	0.112	0.77, 1.21	0.84	0.122	0.64, 1.05
PTT				1.01	0.013	0.98, 1.03
BMI				1.18	0.131	0.92, 1.59
Base excess (mmol/L)				0.91	0.126	0.67, 1.13

[1] OR = Odds Ratio, SE = Standard Error, CI = Confidence Interval

The local prediction function selected as a part of the 'optimal' tuning parameter settings for the local learner in the above example was the weighted mean. However, in order to better demonstrate the local learner's retention of interpretability, a 'runner-up' set of tuning parameter settings were applied, producing a final model in which a weighted GLM using the top 5 features in the patient's 'neighborhood' is chosen as the local prediction function. The local model fit for two example test set patients produced the coefficient estimates displayed in Table 3.1 and the partial regression plots (for one of the two test set patients only) displayed in Figure 3.7. Neither of the two example patients were alive at 24 hours post-randomization.

The exponentiated coefficient estimates in Table 3.1 demonstrate the differences and similarities between the top 5 features in each patient's 'neighborhood' – with WBC count and pH selected as important features for both – as well as the differences in each feature's relative contribution to the predicted outcome both across and within patients. The top 5 features for each were a subset of the 9 globally selected features, which included the 8 in Table 3.1, along with Age. Although most of the coefficient estimates are not statistically significant in these particular examples from the non-optimal local learner fit, the ease with which these estimates can be retrieved and examined is an advantage of the local learning approach more generally.

Using the partial regression visualization in Figure 3.7, it is clear not only which features influenced the prediction for this patient, but also where this patient's values for each predictor lined up against the values of the other patients in their 'neighborhood.' In this particular example, an extremely low base excess and pH, along with a fairly low WBC count and fairly high PTT, appear to be driving the predicted classification of this patient as one who would be at risk to not survive past 24 hours.

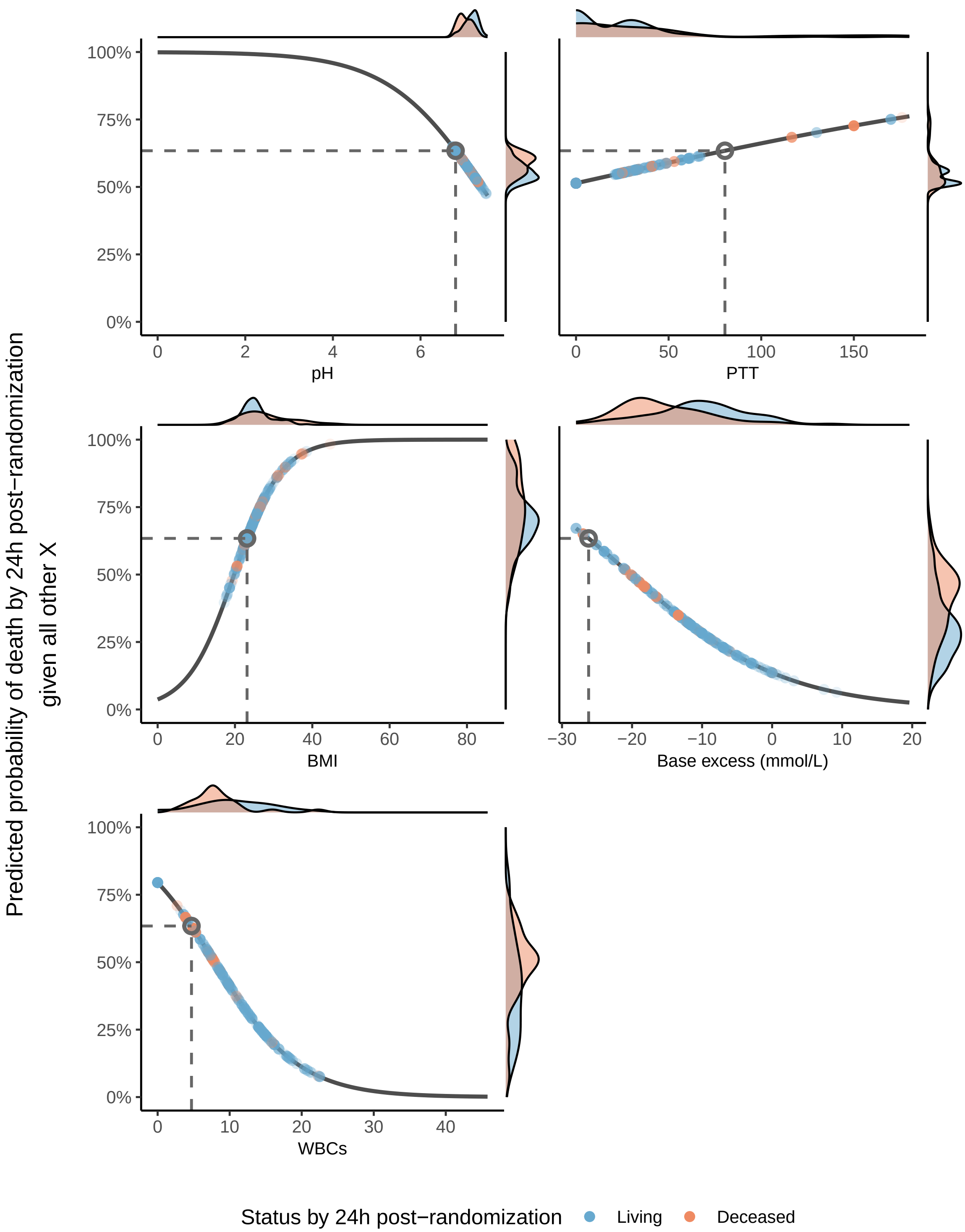

Figure 3.7: Partial regression plots

Chapter 4

Discussion

In this manuscript, we propose a novel local learning framework for prediction. This local learning algorithm accommodates complex and non-linear patterns between predictors and outcome, mitigating overfitting via cross-validation while simultaneously prioritizing interpretable results and explainable predictions.

Accuracy advantages over opaque and/or "global" learners are not guaranteed, nor did this work attempt to establish that data generating distributions exist for which the local learner would consistently and dramatically outperform any existing global learning algorithm. Theoretically, a local learner like the one proposed in this manuscript could have accuracy advantages when applied to e.g. real-world clinical data sampled from a mixture of several patient populations. However, the major advantage of the approach proposed in this manuscript – as demonstrated by simulations presented herein – is its retention of interpretability without significant sacrifices to predictive performance, particularly when compared against "global" learners. This interpretability is provided in the form of both local (observation-specific) variable selection and local parameter estimates.

The "training" of the local learner involves selection of a combination of tuning parameter settings optimizing a global criterion (e.g. maximizing AUC). After tuning parameters are selected, prediction is far less computationally intensive than model tuning. However, in the current implementation, the tuning process involves a grid search among many combinations of settings for the various tuning parameters and so is computationally intensive. This is mitigated somewhat via parallelization and vectorization, but future work should explore alternatives, such as explicit pre-specification of tuning parameters, tuning parameter selection via random search instead of grid search, and optimization of an objective criterion via iterating on tuning parameter settings until that criterion reaches a predefined threshold.

Some modeling approaches that could have improved predictive performance were eschewed in favor of retention of interpretability. One such approach is the centering of feature values for each local model around those of the observation for which a prediction is desired.

Still other extensions to this framework were beyond the scope of an initial implementation. These potential extensions include enabling an adaptive window width so that local models are not restricted to using a single, globally-imposed neighborhood size.

Bibliography

[1] World Health Organization (WHO). *Health expenditure, total (% of GDP)*. [Online; accessed 19-Oct-2020]. 2017. URL: http://apps.who.int/nha/database.

[2] World Health Organization (WHO). *Injuries and violence: the facts 2014*. Geneva: World Health Organization, 2014. URL: http://apps.who.int/iris/handle/10665/149798.

[3] María M. Abad-Grau, Claudio Cervino Jorge Ierache, and Paola Sebastiani. "Evolution and challenges in the design of computational systems for triage assistance". In: *Journal of Biomedical Informatics* 41.3 (2008), pp. 432–441. DOI: 10.1016/j.jbi.2008.01.007. URL: http://dx.doi.org/10.1016/j.jbi.2008.01.007.

[4] Naomi Altman and Martin Krzywinski. "The curse(s) of dimensionality". In: *Nature Methods* 15.6 (May 2018), pp. 399–400. DOI: 10.1038/s41592-018-0019-x. URL: https://doi.org/10.1038/s41592-018-0019-x.

[5] John H. Armstrong et al. "Is Overtriage Associated With Increased Mortality? The Evidence Says "Yes"". In: *Disaster Medicine and Public Health Preparedness* 2.1 (2008), pp. 4–5. DOI: 10.1097/dmp.0b013e31816476c0. URL: https://doi.org/10.1097/dmp.0b013e31816476c0.

[6] Peter C. Austin and Lawrence J. Brunner. "Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses". In: *Statistics in Medicine* 23.7 (2004), pp. 1159–1178. ISSN: 1097-0258. DOI: 10.1002/sim.1687. URL: http://dx.doi.org/10.1002/sim.1687.

[7] Christopher J Aylwin et al. "Reduction in critical mortality in urban mass casualty incidents: analysis of triage, surge, and resource use after the London bombings on July 7, 2005". In: *The Lancet* 368.9554 (2006), pp. 2219–2225. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(06)69896-6. URL: http://dx.doi.org/10.1016/S0140-6736(06)69896-6.

[8] Susan P. Baker and Brian O'Neill. "The Injury Severity Score: an Update". In: *Journal of Trauma and Acute Care Surgery* 16.11 (1976), pp. 882–885. DOI: 10.1097/00005373-197611000-00006. URL: http://dx.doi.org/10.1097/00005373-197611000-00006.

[9] Susan P. Baker et al. "The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care". In: *Journal of Trauma and Acute Care Surgery* 14.3 (1974), pp. 187–196. DOI: 10.1097/00005373-197403000-00001. URL: http://dx.doi.org/10.1097/00005373-197403000-00001.

[10] Chris Barsi et al. "Risk factors and mortality associated with undertriage at a level I safety-net trauma center: a retrospective study". In: *Open Access Emergency Medicine* 8 (2016), pp. 103–110. DOI: 10.2147/oaem.s117397. URL: https://doi.org/10.2147/oaem.s117397.

[11] James Bergstra and Yoshua Bengio. "Random Search for Hyper-Parameter Optimization". In: *The Journal of Machine Learning Research* 13.null (Feb. 2012), pp. 281–305. ISSN: 1532-4435. DOI: 10.5555/2188385.2188395.

[12] Eta S. Berner, ed. *Clinical Decision Support Systems - Theory and Practice*. 2nd. Health Informatics. New York, NY: Springer, 2007. DOI: 10.1007/978-0-387-38319-4. URL: http://www.springer.com/br/book/9781441922236.

[13] Kevin Beyer et al. "When Is "Nearest Neighbor" Meaningful?" In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1999, pp. 217–235. DOI: 10.1007/3-540-49257-7_15. URL: https://doi.org/10.1007/3-540-49257-7_15.

[14] Roger C. Bone et al. "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine". In: *Chest* 101.6 (1992), pp. 1644–1655.

[15] Carl R. Boyd, Mary Ann Tolson, and Wayne S. Copes. "Evaluating trauma care: the TRISS method". In: *Journal of Trauma and Acute Care Surgery* 27.4 (1987), pp. 370–378. DOI: 10.1097/00005373-198704000-00005. URL: http://dx.doi.org/10.1097/00005373-198704000-00005.

[16] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. DOI: 10.1007/bf00058655. URL: https://doi.org/10.1007/bf00058655.

[17] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[18] Stefan Candefjord, Linn Asker, and Eva-Corina Caragounis. “Mortality of trauma patients treated at trauma centers compared to non-trauma centers in Sweden: a retrospective study”. In: *European Journal of Trauma and Emergency Surgery* (July 2020). DOI: 10.1007/s00068-020-01446-6. URL: https://doi.org/10.1007/s00068-020-01446-6.

[19] Access Center for Financing, Agency for Healthcare Research Cost Trends, and Quality. “Total expenditures in millions by condition and age groups, United States, 2017”. In: *Medical Expenditure Panel Survey* (2017). [Generated interactively: Mon Nov 23 2020]. URL: https://www.meps.ahrq.gov/mepstrends/hc_cond_icd10/.

[20] Howard R. Champion, William J. Sacco, and Wayne S. Copes. “Injury Severity Scoring Again”. In: *Journal of Trauma and Acute Care Surgery* 38.1 (1995), pp. 94–95. DOI: 10.1097/00005373-199501000-00024. URL: http://dx.doi.org/10.1097/00005373-199501000-00024.

[21] Howard R. Champion et al. “A new characterization of injury severity.” In: *Journal of Trauma and Acute Care Surgery* 30.5 (1990), pp. 539–546.

[22] Howard R. Champion et al. “A revision of the Trauma Score”. In: *Journal of Trauma and Acute Care Surgery* 29.5 (1989), pp. 623–629. DOI: 10.1097/00005373-198905000-00017. URL: http://dx.doi.org/10.1097/00005373-198905000-00017.

[23] Howard R. Champion et al. “Trauma score”. In: *Crit. Care Med.* 9.9 (1981), pp. 672–676.

[24] M. N. Chawda et al. “Predicting outcome after multiple trauma: which scoring system?” In: *Injury* 35.4 (2004), pp. 347–358. ISSN: 0020-1383. DOI: 10.1016/s0020-1383(03)00140-2. URL: http://dx.doi.org/10.1016/S0020-1383(03)00140-2.

[25] Liangyou Chen et al. “Decision tool for the early diagnosis of trauma patient hypovolemia”. In: *Journal of Biomedical Informatics* 41.3 (2008), pp. 469–478. DOI: 10.1016/j.jbi.2007.12.002. URL: http://dx.doi.org/10.1016/j.jbi.2007.12.002.

[26] William S. Cleveland. “Robust Locally Weighted Regression and Smoothing Scatterplots”. In: *Journal of the American Statistical Association* 74.368 (Dec. 1979), pp. 829–836. DOI: 10.1080/01621459.1979.10481038. URL: https://doi.org/10.1080/01621459.1979.10481038.

[27] Committee on Medical Aspects of Automotive Safety. “Rating the Severity of Tissue Damage: I. The Abbreviated Scale”. In: *JAMA: The Journal of the American Medical Association* 215.2 (1971), pp. 277–280. DOI: 10.1001/jama.1971.03180150059012. URL: https://doi.org/10.1001/jama.1971.03180150059012.

[28] Committee on Trauma - American College of Surgeons. *Resources for Optimal Care of the Injured Patient 2014*. Chicago, 2014.

[29] National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington, DC: The National Academies Press, 2011. ISBN: 978-0-309-22222-8. URL: http://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research.

[30] Thomas M. Cover and Peter E. Hart. “Nearest neighbor pattern classification”. In: *Information Theory, IEEE Transactions on* 13.1 (1967), pp. 21–27.

[31] Cristiane de Alencar Domingues et al. “The role of the New Trauma and Injury Severity Score (NTRISS) for survival prediction”. In: *Rev Esc Enferm USP* 45.6 (2011), pp. 1353–1358.

[32] Sahibsingh A. Dudani. “The Distance-Weighted k-Nearest-Neighbor Rule”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.4 (Apr. 1976), pp. 325–327. DOI: 10.1109/tsmc.1976.5408784. URL: https://doi.org/10.1109/tsmc.1976.5408784.

[33] Sandrine Dudoit. *Lecture notes for PB HLTH C240D*. Oct. 2013.

[34] Brian J. Eastridge et al. “Death on the battlefield (2001–2011)”. In: *Journal of Trauma and Acute Care Surgery* 73 (2012), S431–S437. DOI: 10.1097/ta.0b013e3182755dcc. URL: https://doi.org/10.1097/ta.0b013e3182755dcc.

[35] Organisation for Economic Co-operation and Development (OECD). *Health spending (indicator)*. [Online; accessed 23-May-2017]. 2017. DOI: 10.1787/8643de7e-en. URL: https://data.oecd.org/healthres/health-spending.htm.

[36] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[37] Jerome H. Friedman. “Multivariate Adaptive Regression Splines”. In: *The Annals of Statistics* 19.1 (1991), pp. 1–67.

[38] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. “A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study”. In: *JAMA: The Journal of the American Medical Association* 270.24 (1993), p. 2957. DOI: 10.1001/jama.1993.03510240069035. URL: https://doi.org/10.1001/jama.1993.03510240069035.

[39] Jean-Roger Le Gall et al. “A simplified acute physiology score for ICU patients”. In: *Critical Care Medicine* 12.11 (1984), pp. 975–977. DOI: 10.1097/00003246-198411000-00012. URL: https://doi.org/10.1097/00003246-198411000-00012.

[40] Jean-Roger Le Gall et al. “The Logistic Organ Dysfunction System: A New Way to Assess Organ Dysfunction in the Intensive Care Unit”. In: *JAMA: The Journal of the American Medical Association* 276.10 (1996), pp. 802–810. DOI: 10.1001/jama.1996.03540100046027. URL: http://dx.doi.org/10.1001/jama.1996.03540100046027.

[41] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. “Variable selection using random forests”. In: *Pattern Recognition Letters* 31.14 (Oct. 2010), pp. 2225–2236. DOI: 10.1016/j.patrec.2010.03.014. URL: https://doi.org/10.1016/j.patrec.2010.03.014.

[42] Laurent G. Glance et al. “TMPM–ICD9”. In: *Annals of Surgery* 249.6 (2009), pp. 1032–1039. DOI: 10.1097/sla.0b013e3181a38f28. URL: https://doi.org/10.1097/sla.0b013e3181a38f28.

[43] James L. Guzzo et al. “Prediction of Outcomes in Trauma: Anatomic or Physiologic Parameters?” In: *Journal of the American College of Surgeons* 201.6 (2005), pp. 891–897. DOI: 10.1016/j.jamcollsurg.2005.07.013. URL: https://doi.org/10.1016/j.jamcollsurg.2005.07.013.

[44] Edward L. Hannan et al. “Validation of TRISS and ASCOT using a non-MTOS trauma registry”. In: *Journal of Trauma and Acute Care Surgery* 38.1 (1995), pp. 83–88.

[45] Frank Harrell. *Problems Caused by Categorizing Continuous Variables*. Vanderbilt Biostatistics Wiki. Mar. 18, 2017. URL: http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous (visited on 08/04/2017).

[46] Z. G. Hashmi et al. “The potential for trauma quality improvement: One hundred thousand lives in five years”. Presented at the 11th Annual Academic Surgical Congress, Jacksonville, Florida. Feb. 2016. URL: http://www.asc-abstracts.org/abs2016/15-12-the-potential-for-trauma-quality-improvement-one-hundred-thousand-lives-in-five-years/.

[47] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 2009.

[48] United States Department of Health, Centers for Disease Control Human Services, and National Center for Health Statistics Prevention. “Underlying Cause of Death 2017”. In: *Multiple Cause of Death File 2017* 20.2W (2018). [CDC WONDER Online Database; accessed 27-Nov-2020. Data are compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.] URL: http://wonder.cdc.gov/ucd-icd10.html.

[49] Melonie Heron. “Deaths: Leading causes for 2017”. In: *National Vital Statistics Reports* 68.6 (June 2019). URL: http://www.cdc.gov/nchs/products/nvsr.htm.

[50] Michael J. Hine et al. “Decision Making by Emergency Room Physicians and Residents: Implications for the Design of Clinical Decision Support Systems”. In: *International Journal of Healthcare Information Systems and Informatics* 4.2 (2009), pp. 17–35. DOI: 10.4018/jhisi.2009040102. URL: http://dx.doi.org/10.4018/jhisi.2009040102.

[51] Kwok M Ho et al. “A comparison of admission and worst 24-hour Acute Physiology and Chronic Health Evaluation II scores in predicting hospital mortality: a retrospective cohort study”. In: *Critical Care* 10.1 (Nov. 2005), R4. ISSN: 1364-8535. DOI: 10.1186/cc3913. URL: http://dx.doi.org/10.1186/cc3913.

[52] John B. Holcomb et al. “Transfusion of plasma, platelets, and red blood cells in a 1:1:1 vs a 1:1:2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial”. In: *JAMA* 313.5 (2015), pp. 471–482. DOI: 10.1001/jama.2015.12. URL: http://dx.doi.org/10.1001/jama.2015.12.

[53] Kai-Zhu Huang, Haiqin Yang, and Michael R. Lyu. *Machine learning: modeling data locally and globally*. Springer Science & Business Media, 2008.

[54] Rao R. Ivatury et al. "Patient Safety in Trauma: Maximal Impact Management Errors at a Level I Trauma Center". In: *The Journal of Trauma: Injury, Infection, and Critical Care* 64.2 (2008), pp. 265–272. DOI: 10.1097/ta.0b013e318163359d. URL: https://doi.org/10.1097/ta.0b013e318163359d.

[55] Jin Hee Jeong et al. "The new trauma score (NTS): a modification of the revised trauma score for better trauma mortality prediction". In: *BMC Surgery* 17.1 (2017), p. 77. DOI: 10.1186/s12893-017-0272-4. URL: https://doi.org/10.1186/s12893-017-0272-4.

[56] J. M. Jones, A. D. Redmond, and J. Templeton. "Uses and abuses of statistical models for evaluating trauma care". In: *Journal of Trauma and Acute Care Surgery* 38.1 (1995), pp. 89–93.

[57] MAC de Jongh, MHJ Verhofstad, and LPH Leenen. "Accuracy of different survival prediction models in a trauma population". In: *British Journal of Surgery* 97.12 (2010), pp. 1805–1813. ISSN: 1365-2168. DOI: 10.1002/bjs.7216. URL: https://doi.org/10.1002/bjs.7216.

[58] William A. Knaus et al. "APACHE – acute physiology and chronic health evaluation: a physiologically based classification system". In: *Critical Care Medicine* 9.8 (1981), pp. 591–597. DOI: 10.1097/00003246-198108000-00008. URL: https://doi.org/10.1097/00003246-198108000-00008.

[59] William A. Knaus et al. "APACHE II: A severity of disease classification system". In: *Critical Care Medicine* 13.10 (1985), pp. 818–829. DOI: 10.1097/00003246-198510000-00009. URL: https://doi.org/10.1097/00003246-198510000-00009.

[60] William A. Knaus et al. "The APACHE III Prognostic System". In: *Chest* 100.6 (1991), pp. 1619–1636. DOI: 10.1378/chest.100.6.1619. URL: https://doi.org/10.1378/chest.100.6.1619.

[61] Olive C. Kobusingye and Ronald R. Lett. "Hospital-based trauma registries in Uganda". In: *Journal of Trauma and Acute Care Surgery* 48.3 (2000), pp. 498–502.

[62] Yutaka Kondo et al. "Revised trauma scoring system to predict in-hospital mortality in the emergency department: Glasgow Coma Scale, Age, and Systolic Blood Pressure score". In: *Critical Care* 15.4 (2011), R191. ISSN: 1364-8535. DOI: 10.1186/cc10348. URL: http://dx.doi.org/10.1186/cc10348.

[63] Frank Kroezen et al. "Base Deficit-Based Predictive Modeling of Outcome in Trauma Patients Admitted to Intensive Care Units in Dutch Trauma Centers". In: *The Journal of Trauma: Injury, Infection, and Critical Care* 63.4 (2007), pp. 908–913. DOI: 10.1097/ta.0b013e318151ff22. URL: https://doi.org/10.1097/ta.0b013e318151ff22.

[64] Deborah A. Kuhls et al. "Predictors of mortality in adult trauma patients: the Physiologic Trauma Score is equivalent to the Trauma and Injury Severity Score". In: *Journal of the American College of Surgeons* 194.6 (2002), pp. 695–704. ISSN: 1072-7515. DOI: 10.1016/S1072-7515(02)01211-5. URL: http://dx.doi.org/10.1016/S1072-7515(02)01211-5.

[65] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007). DOI: 10.2202/1544-6115.1309. URL: http://dx.doi.org/10.2202/1544-6115.1309.

[66] Siu W. Lam et al. "Validation of a base deficit-based trauma prediction model and comparison with TRISS and ASCOT". In: *European Journal of Trauma and Emergency Surgery* 42.5 (2016), pp. 627–633. ISSN: 1863-9941. DOI: 10.1007/s00068-015-0592-y. URL: http://dx.doi.org/10.1007/s00068-015-0592-y.

[67] Erin LeDell, Maya Petersen, and Mark van der Laan. "Computationally Efficient Confidence Intervals for Cross-validated Area Under the ROC Curve Estimates". In: *Electronic Journal of Statistics* 9.1 (2015), pp. 1583–1607. DOI: 10.1214/15-EJS1035. URL: http://projecteuclid.org/euclid.ejs/1437742107.

[68] Rolf Lefering. "Development and validation of the revised injury severity classification score for severely injured patients". In: *European Journal of Trauma and Emergency Surgery* 35.5 (2009), pp. 437–447. ISSN: 1863-9941. DOI: 10.1007/s00068-009-9122-0. URL: http://dx.doi.org/10.1007/s00068-009-9122-0.

[69] Rolf Lefering et al. "Update of the trauma risk adjustment model of the TraumaRegister DGU: the Revised Injury Severity Classification, version II". In: *Critical Care* 18.5 (Sept. 2014), p. 476. DOI: 10.1186/s13054-014-0476-2. URL: https://doi.org/10.1186/s13054-014-0476-2.

[70] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

[71] Nehemiah T. Liu and Jose Salinas. "Machine Learning for Predicting Outcomes in Trauma". In: *SHOCK* 48.5 (Nov. 2017), pp. 504–510. DOI: 10.1097/SHK.0000000000000898. URL: https://doi.org/10.1097/SHK.0000000000000898.

[72] Ellen J MacKenzie, Sam Shapiro, and James N Eastham Jr. “Rating AIS severity using emergency department sheets vs. inpatient charts”. In: *Journal of Trauma and Acute Care Surgery* 25.10 (1985), pp. 984–988.

[73] Ellen J. MacKenzie et al. “A National Evaluation of the Effect of Trauma-Center Care on Mortality”. In: *New England Journal of Medicine* 354.4 (Jan. 2006), pp. 366–378. DOI: 10.1056/nejmsa052049. URL: https://doi.org/10.1056/nejmsa052049.

[74] John C Marshall et al. “Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome”. In: *Critical care medicine* 23.10 (1995), pp. 1638–1652.

[75] Stephen Milborrow. *earth: Multivariate Adaptive Regression Splines*. R package version 4.4.9. 2017. URL: https://CRAN.R-project.org/package=earth.

[76] Ross T. Miller et al. “The modified rapid emergency medicine score: A novel trauma triage tool to predict in-hospital mortality”. In: *Injury* (2017). DOI: 10.1016/j.injury.2017.04.048. URL: https://doi.org/10.1016/j.injury.2017.04.048.

[77] Charles N. Mock et al. “Trauma mortality patterns in three nations at different economic levels: implications for global trauma system development”. In: *Journal of Trauma and Acute Care Surgery* 44.5 (1998), pp. 804–814.

[78] Lorenzo Moja et al. “Effectiveness of Computerized Decision Support Systems Linked to Electronic Health Records: A Systematic Review and Meta-Analysis”. In: *American Journal of Public Health* 104.12 (2014), e12–e22. DOI: 10.2105/AJPH.2014.302164. URL: https://doi.org/10.2105/AJPH.2014.302164.

[79] Sandra Montmany et al. “Preventable deaths and potentially preventable deaths. What are our errors?” In: *Injury* 47.3 (2016), pp. 669–673. DOI: 10.1016/j.injury.2015.11.028. URL: https://doi.org/10.1016/j.injury.2015.11.028.

[80] Ernest E Moore et al. “Organ Injury Scaling III: Chest Wall, Abdominal Vascular, Ureter, Bladder, and Urethra”. In: *Journal of Trauma* 33.3 (1992), pp. 337–339.

[81] Ernest E Moore et al. “Organ Injury Scaling IV: Thoracic Vascular, Lung, Cardiac, and Diaphrag”. In: *Journal of Trauma* 36.3 (1994), pp. 299–300.

[82] Ernest E Moore et al. “Organ Injury Scaling VI: Extrahepatic Biliary, Esophagus, Stomach, Vulva, Vagina, Uterus (Nonpregnant), Uterus (Pregnant), Fallopian Tube, and Ovary”. In: *Journal of Trauma* 39.6 (1995), pp. 1069–1070.

[83] Ernest E Moore et al. "Organ Injury Scaling VII: Cervical Vascular, Peripheral Vascular, Adrenal, Penis, Testis, and Scrotum". In: *Journal of Trauma* 41.3 (1996), pp. 523–524.

[84] Ernest E Moore et al. "Organ Injury Scaling, II: Pancreas, duodenum, small bowel, colon, and rectum". In: *Journal of Trauma* 30.11 (1990), pp. 1427–1429.

[85] Ernest E Moore et al. "Organ Injury Scaling: Spleen and Liver (1994 Revision)". In: *Journal of Trauma* 38.3 (1995), pp. 323–324.

[86] Ernest E Moore et al. "Organ Injury Scaling: Spleen, Liver, and Kidney". In: *Journal of Trauma* 29.12 (1989), pp. 1664–1666.

[87] Sara Moore. *SLScreenExtra: A Collection of Additional Feature Selection Algorithms and Utilities for SuperLearner*. R package version 0.4.6. 2023. URL: https://github.com/saraemoore/SLScreenExtra.

[88] Sara Moore. *SuperSelector: Ensembled Feature Selection using Cross-Validated SuperLearner*. R package version 0.1.2. 2023. URL: https://github.com/saraemoore/SuperSelector.

[89] Sara Moore. *Yet Another Local Learner (YALL): A localized machine learning algorithm inspired by and built on Cross-Validated SuperLearner*. R package version 0.1.0. 2023. URL: https://github.com/saraemoore/yall.

[90] Rui P. Moreno et al. "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission". In: *Intensive Care Medicine* 31.10 (2005), pp. 1345–1355. ISSN: 1432-1238. DOI: 10.1007/s00134-005-2763-5. URL: http://dx.doi.org/10.1007/s00134-005-2763-5.

[91] Leonie de Munter et al. "Mortality prediction models in the general trauma population: A systematic review". In: *Injury* 48.2 (2017), pp. 221–229. DOI: 10.1016/j.injury.2016.12.009. URL: https://doi.org/10.1016/j.injury.2016.12.009.

[92] Lena M. Napolitano et al. "Systemic inflammatory response syndrome score at admission independently predicts mortality and length of stay in trauma patients". In: *The Journal of Trauma: Injury, Infection, and Critical Care* 49.4 (2000), pp. 647–652.

[93] Timothy C. Nunez et al. "Early Prediction of Massive Transfusion in Trauma: Simple as ABC (Assessment of Blood Consumption)?" In: *The Journal of Trauma: Injury, Infection, and Critical Care* 66.2 (Feb. 2009), pp. 346–352. DOI: 10.1097/ta.0b013e3181961c35. URL: https://doi.org/10.1097/ta.0b013e3181961c35.

[94] Turner Osler, Susan P Baker, and William Long. "A modification of the injury severity score that both improves accuracy and simplifies scoring". In: *Journal of Trauma and Acute Care Surgery* 43.6 (1997), pp. 922–926.

[95] Turner Osler et al. "A Trauma Mortality Prediction Model Based on the Anatomic Injury Scale". In: *Annals of Surgery* 247.6 (2008), pp. 1041–1048. DOI: 10.1097/sla.0b013e31816ffb3f. URL: https://doi.org/10.1097/sla.0b013e31816ffb3f.

[96] Turner Osler et al. "ICISS: an international classification of disease-9 based injury severity score". In: *Journal of Trauma and Acute Care Surgery* 41.3 (1996), pp. 380–388.

[97] Eric Polley et al. *SuperLearner: Super Learner Prediction*. R package version 2.0-26. 2019. URL: https://github.com/ecpolley/SuperLearner.

[98] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. DOI: 10.48550/ARXIV.1602.04938. URL: https://arxiv.org/abs/1602.04938.

[99] A. Rogers et al. "Increased mortality with undertriaged patients in a mature trauma center with an aggressive trauma team activation system". In: *European Journal of Trauma and Emergency Surgery* 39.6 (2013), pp. 599–603. DOI: 10.1007/s00068-013-0289-z. URL: http://dx.doi.org/10.1007/s00068-013-0289-z.

[100] Piotr Romanski, Lars Kotthoff, and Patrick Schratz. *FSelector: Selecting Attributes*. R package version 0.34. 2023. URL: https://github.com/larskotthoff/fselector.

[101] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Statistics in Medicine* 25.1 (2006), pp. 127–141. ISSN: 1097-0258. DOI: 10.1002/sim.2331. URL: http://dx.doi.org/10.1002/sim.2331.

[102] William J Sacco et al. "Comparison of alternative methods for assessing injury severity based on anatomic descriptors". In: *Journal of Trauma* 47.3 (1999), pp. 441–446. URL: https://www.ncbi.nlm.nih.gov/pubmed/10498295.

[103] Danielle Sartorius et al. "Mechanism, Glasgow Coma Scale, Age, and Arterial Pressure (MGAP): A new simple prehospital triage score to predict mortality in trauma patients". In: *Critical Care Medicine* 38.3 (2010), pp. 831–837. DOI: 10.1097/ccm.0b013e3181cc4a67. URL: https://doi.org/10.1097/ccm.0b013e3181cc4a67.

[104] Philip J. Schluter et al. "Trauma and Injury Severity Score (TRISS) Coefficients 2009 Revision". In: *The Journal of Trauma: Injury, Infection, and Critical Care* 68.4 (2010), pp. 761–770. DOI: 10.1097/ta.0b013e3181d3223b. URL: https://doi.org/10.1097/ta.0b013e3181d3223b.

[105] Christopher K. Senkowski and Mark G. McKenney. "Trauma scoring systems: a review". In: *Journal of the American College of Surgeons* 189.5 (1999), pp. 491–503. ISSN: 1072-7515. DOI: 10.1016/S1072-7515(99)00190-8. URL: http://dx.doi.org/10.1016/S1072-7515(99)00190-8.

[106] B. W. Silverman. *Density estimation for statistics and data analysis*. Vol. 26. Monographs on statistics and applied probability. London; New York: Chapman and Hall, 1986. ISBN: 0412246201.

[107] Oleg Sofrygin, Mark J. van der Laan, and Romain Neugebauer. *simcausal: Simulating Longitudinal Data with Causal Inference Applications*. R package version 0.5.5. 2020. URL: https://github.com/osofr/simcausal.

[108] Jaime Lynn Speiser et al. "A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling". In: *Expert Systems with Applications* 134 (Nov. 2019), pp. 93–101. DOI: 10.1016/j.eswa.2019.05.028. URL: https://doi.org/10.1016/j.eswa.2019.05.028.

[109] R. T. Spence et al. "Injury Severity Score coding: Data analyst v. emerging m-health technology". In: *South African Medical Journal* 106.10 (2016), p. 1037. DOI: 10.7196/samj.2016.v106i10.10597. URL: https://doi.org/10.7196/samj.2016.v106i10.10597.

[110] Matthew Sperrin et al. "Missing data should be handled differently for prediction than for description or causal explanation". In: *Journal of Clinical Epidemiology* 125 (Sept. 2020), pp. 183–187. DOI: 10.1016/j.jclinepi.2020.03.028. URL: https://doi.org/10.1016/j.jclinepi.2020.03.028.

[111] Graham Teasdale and Bryan Jennett. "Assessment of Coma and Impaired Consciousness: A Practical Scale". In: *The Lancet* 304.7872 (1974), pp. 81–84. DOI: 10.1016/s0140-6736(74)91639-0. URL: https://doi.org/10.1016/s0140-6736(74)91639-0.

[112] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society B* 58.1 (1996), pp. 267–288.

[113] Lyle Ungar. *CIS520 Machine Learning | Lectures / LocalLearning*. [Online; accessed 20-February-2023]. 2021. URL: https://alliance.seas.upenn.edu/~cis520/dynamic/2021/wiki/index.php?n=Lectures.LocalLearning.

[114] Gilmer Valdes et al. “The Conditional Super Learner”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (Dec. 2022), pp. 10236–10243. DOI: 10.1109/tpami.2021.3131976. URL: https://doi.org/10.1109/tpami.2021.3131976.

[115] Jean-Louis Vincent and Rui Moreno. “Clinical review: Scoring systems in the critically ill”. In: *Critical Care* 14.2 (2010), p. 207. ISSN: 1364-8535. DOI: 10.1186/cc8204. URL: http://dx.doi.org/10.1186/cc8204.

[116] Jean-Louis Vincent et al. “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure”. In: *Intensive care medicine* 22.7 (1996), pp. 707–710.

[117] Sandra M. Vioque et al. “Classifying errors in preventable and potentially preventable trauma deaths: a 9-year review using the Joint Commission’s standardized methodology”. In: *The American Journal of Surgery* 208.2 (2014), pp. 187–194. DOI: 10.1016/j.amjsurg.2014.02.006. URL: https://doi.org/10.1016/j.amjsurg.2014.02.006.

[118] T. Al West et al. “Harborview assessment for risk of mortality: an improved measure of injury severity on the basis of ICD-9-CM”. In: *Journal of Trauma and Acute Care Surgery* 49.3 (2000), pp. 530–541.

[119] Cydni N Williams, Susan L Bratton, and Eliotte L Hirshberg. “Computerized decision support in adult and pediatric critical care”. In: *World Journal of Critical Care Medicine* 2.4 (2013), pp. 21–28. DOI: 10.5492/wjccm.v2.i4.21. URL: http://dx.doi.org/10.5492/wjccm.v2.i4.21.

[120] David H. Wisner. “History and current status of trauma scoring systems”. In: *Archives of Surgery* 127.1 (1992), pp. 111–117. DOI: 10.1001/archsurg.1992.01420010133022. URL: http://dx.doi.org/10.1001/archsurg.1992.01420010133022.

[121] Huiyun Xiang et al. “Undertriage of major trauma patients in the US emergency departments”. In: *The American Journal of Emergency Medicine* 32.9 (2014), pp. 997–1004. DOI: 10.1016/j.ajem.2014.05.038. URL: http://dx.doi.org/10.1016/j.ajem.2014.05.038.

[122] David W. Yates. “ABC of major trauma: Scoring systems for trauma”. In: *British Medical Journal* 301.6760 (1990), pp. 1090–1094. DOI: 10.1136/bmj.301.6760.1090. URL: http://dx.doi.org/10.1136/bmj.301.6760.1090.

[123] Nedim Yücel et al. "Trauma Associated Severe Hemorrhage (TASH)-Score: Probability of Mass Transfusion as Surrogate for Life Threatening Hemorrhage after Multiple Trauma". In: *The Journal of Trauma: Injury, Infection, and Critical Care* 60.6 (June 2006), pp. 1228–1237. DOI: 10.1097/01.ta.0000220386.84012.bf. URL: https://doi.org/10.1097/01.ta.0000220386.84012.bf.

[124] Ying Zhang and Peter Szolovits. "Patient-specific learning in real time for adaptive monitoring in critical care". In: *Journal of Biomedical Informatics* 41.3 (2008), pp. 452–460. DOI: 10.1016/j.jbi.2008.03.011. URL: http://dx.doi.org/10.1016/j.jbi.2008.03.011.

[125] Jack E. Zimmerman et al. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients". In: *Critical Care Medicine* 34.5 (2006), pp. 1297–1310. DOI: 10.1097/01.ccm.0000215112.84523.f0. URL: https://doi.org/10.1097/01.ccm.0000215112.84523.f0.