Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2009-038 August 26, 2009

Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models

Caleb Hug

Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models

by

Caleb Wayne Hug

S.M., Massachusetts Institute of Technology (2006)
B.S., Whitworth College (2004)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

Author . . . .
Department of Electrical Engineering and Computer Science
May 19, 2009

Certified by . . . .
Peter Szolovits
Professor
Thesis Supervisor

Accepted by . . . .
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Detecting Hazardous Intensive Care Patient Episodes Using Real-time Mortality Models

by

Caleb Wayne Hug

Submitted to the Department of Electrical Engineering and Computer Science on May 19, 2009, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Abstract

The modern intensive care unit (ICU) has become a complex, expensive, data-intensive environment. Caregivers maintain an overall assessment of their patients based on important observations and trends. If an advanced monitoring system could also reliably provide a systemic interpretation of a patient's observations it could help caregivers interpret these data more rapidly and perhaps more accurately.

In this thesis I use retrospective analysis of mixed medical/surgical intensive care patients to develop predictive models. Logistic regression is applied to 7048 development patients with several hundred candidate variables. These candidate variables range from simple vitals to long term trends and baseline deviations. Final models are selected by backward elimination on top cross-validated variables and validated on 3018 additional patients.

The real-time acuity score (RAS) that I develop demonstrates strong discrimination ability for patient mortality, with an ROC area (AUC) of 0.880. The final model includes a number of variables known to be associated with mortality, but also computationally intensive variables absent in other severity scores. In addition to RAS, I also develop secondary outcome models that perform well at predicting pressor weaning (AUC=0.825), intraaortic balloon pump removal (AUC=0.816), the onset of septic shock (AUC=0.843), and acute kidney injury (AUC=0.742).

Real-time mortality prediction is a feasible way to provide continuous risk assessment for ICU patients. RAS offers similar discrimination ability when compared to models computed once per day, based on aggregate data over that day. Moreover, RAS mortality predictions are better at discrimination than a customized SAPS II score (Day 3 AUC=0.878 vs AUC=0.849, $p < 0.05$). The secondary outcome models also provide interesting insights into patient responses to care and patient risk profiles. While models trained for specifically recognizing secondary outcomes consistently outperform the RAS model at their specific tasks, RAS provides useful baseline risk estimates throughout these events and in some cases offers a notable level of predictive utility.

Thesis Supervisor: Peter Szolovits
Title: Professor

Acknowledgments

The research described in this thesis was funded in part by the National Library of Medicine (training grant LM 07092) and the National Institute of Biomedical Imaging and Bioengineering (grant R01 EB001659).

Many people helped make this work possible. I would first like to thank my lovely wife Kendra. She sacrificed many evenings to the time sink that graduate school can easily become. Her friendship, support, perspective (and food!) made the past five years an enjoyable journey — helping through the difficult moments and adding immeasurably to the joyous times. Kendra and I welcomed our first child, Hannah, during my time at MIT. The opportunity to witness the miracle of life and a child's first exploration of the world is an experience that I relish and something that has renewed my appreciation for many of the simplest things in life and inspired my work.

My research advisor, Peter Szolovits, was instrumental in guiding the research presented in this thesis. He allowed me the freedom and responsibility to pursue my own research directions. Many of the resulting digressions and tangents were invaluable contributions to my educational experience. But whenever I had questions, I could count on prescient insight and the rich experiential advice that stems from his illustrious career in medical informatics.

I would also like to acknowledge my other committee members, including Bill Long, Roger Mark, and Lucila Ohno-Machado. Each member provided constructive conversations, helpful advice, and excellent feedback regarding the research presented in this thesis. I am especially grateful for Bill's patience in listening to me as I bounced ideas off of him or sometimes ranted about data problems.

A number of other individuals at MIT also helped me with this work. Tom Lasko's friendship and advice was an immense encouragement to me as I first began to explore the field of medical informatics. Andrew Reisner provided helpful information for many of my clinical questions and was always a joy to talk with. Gari Clifford and Mauricio Villarroel provided much assistance with the data used in this study and provided many enjoyable conversations. I am blessed to consider all of these individuals my friends.

Furthermore, I would be remiss not to mention several others who played an essential role in my successful completion of this work. Kent Jones, Susan Mabry, and Howard Gage — computer science and mathematics faculty from Whitworth College — went out of their way to challenge and encourage me. While the past five years seemed to have passed so quickly, my parents and brothers (especially my oldest brother Joshua) were also wonderful sources of encouragement to me during the times with no apparent end in sight.

During my time at MIT I spent three years living as a graduate resident tutor in an undergraduate dorm. Living with the students of Conner 4 was a great experience and complimented my MIT education in many ways. While I'm excited to finally move off campus, I will miss the energy and community that characterize this group of talented young adults.

Studying in historic New England has nurtured a strong respect for my country. My story is a testament to the exceptionalism of the United States, a country built

on reverence for human liberty and filled with endless opportunities. As my graduate student tenure comes to an end, I marvel at the great privilege of graduate education in this country — an easily overlooked privilege made possible by the generous tax dollars of U.S. citizens. I will not forget the investment that my country has made.

Finally, I would like to thank my Heavenly Father. He has given me a rich life — blessed me with a wonderful family, a healthy body, and all of my worldly needs. Furthermore, He has challenged me with opportunities to grow personally, spiritually, and academically at a renowned institution. I could not ask for more opportunity. As king Solomon wrote in the book of Proverbs, "It is the glory of God to conceal a thing; but the honor of kings is to search out a matter." For a farm boy from the sticks, I indeed feel like I have been in the company of kings. And I believe uncovering the wonder of mathematics and the intricacies of the human body offer but a glimpse of the glory of God.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The modern intensive care unit (ICU) has become a complex, expensive, data-intensive environment. In this environment — where physician decisions often make the difference between life and death — tools that help caregivers interpret patterns in the data and quickly make the correct decision are essential. My objective in this research is to develop models that, given a set of observations, provide a systemic "understanding" of a patient's medical well-being and assist physicians in making more informed decisions. I use a data-driven approach to model the complex patient system by considering variables that range from therapeutic interventions to simple vitals to complex trends. Specifically, I develop several mortality models and compare them against a real-time mortality model. I then compare and contrast the ability of mortality models to predict acute patient events with models that were specialized to predict specific events. If real-time risk models can be successfully developed, physicians could have an immediate alert of jeopardized patient state — providing valuable time to intervene — or an indicator of a particular treatment regime's benefit to an individual patient.

1.1 Overview

While doctors routinely do an outstanding job of matching complex patterns observed in patient data to an applicable set of diagnoses and treatments, they are not perfect. Patients admitted to ICUs — a particularly vulnerable category of patients — require close monitoring due to an increased probability of life threatening events. The close attention of caregivers, necessary to provide high quality care, clearly exposes patients to the human errors known to be common in health care [40]. In fact, Rothschild et al. found that ICU patients suffer a large number iatrogenic injuries, especially failure to carry out intended treatment correctly [71]. A system that understands the patient's progression could potentially catch dangerous episodes and ultimately increase caregiver vigilance.

A national shortage of nurses and high turnover rates likely exacerbate human errors. These shortages are particularly evident in the ICU. One survey conducted in 2000 found that the nurse vacancy rate in critical care, at 14.6%, was higher than other locations [30]. In order to fill vacancies, temporary staff are commonly used in many hospitals. Furthermore, projections indicate that the current shortage of intensivists will shortly be a crisis [41]. In the ICU, the potential for information technology and medical informatics to supply decision support, enhance efficiency, and generally improve quality by utilizing relevant data is well understood (e.g., see [36]). In fact, companies such as VISICU (recently acquired by Philips Medical Systems) have emerged that seek to leverage the intensivist shortage by allowing a single intensivist to monitor up to 100 patients through a remote environment. A system that can effectively interpret a patient's data could help reduce the burden placed on caregivers and, as a result, help alleviate the intensivist shortage.

Despite the theoretical promise of comprehensive patient monitors, reality might present a more dire picture. One recent review by Ospina-Tascón et al. [64] questions the utility of recent monitoring progress by pointing to the systemic lack of randomized controlled trials and argues that, of the few conducted, most show negative results. Besides the clear ethical problems with conducting randomized controlled trials on obviously helpful monitors (e.g., electrocardiogram monitoring for patients with acute myocardial infarction), Ospina-Tasconón's review raises many questions regarding the utility of investing in monitoring development. An alternative interpretation that might be inferred from such criticisms of contemporary monitoring systems is that the systems are inadequate for the actual caregiver needs. Are advanced monitoring devices really helpful, or do they simply overwhelm the nurses and physicians with useless information that do not ultimately benefit patients?

Current monitors do indeed come at a cost. Concerned about sensitivity, monitors often sacrifice specificity. The trade-off between sensitivity and specificity can be seen by the documented prevalence of false alarms [58, 89, 90, 57]. An ancillary burden from devices with low specificity and high sensitivity is excessive background sound — Ryherd et al. found that the noise level in one neurological intensive care unit was significantly higher than recommended by the World Health Organization guidelines [74]. Caregivers surveyed as part of Ryherd's study overwhelmingly indicated that the noise adversely affected them and their patients. The prevalence of audible false alarms indicate that the wealth of observations taken in the ICU are poorly understood at the monitoring level. The problem of better interpreting observations in order to increase specificity has attracted considerable attention, and recent work, such as that by Zong et al., has demonstrated methods for dramatically reducing false alarm rates [95]. Modern monitors are able to corroborate related signals in order to limit spurious alarms. Current approaches, however, generally focus on better alarms on individual signals rather than to fuse information together to reflect the underlying patient condition and produce warnings such as suspected hypovolemia or septic shock.

Informatics can surely help. Looking at recent advances in informatics — such as the ability to quickly search the Internet and find highly relevant information — it is only a matter of time before relevant medical knowledge will be highly accessible from the bedside. The insurmountable electronic medical record obstacle has even shown signs of abating with several large industry players making substantial investments in the personal health record (PHR) arena and releasing promising solutions such as the Microsoft HealthVault, Google Health, or Dossia's Indivo system [60, 20, 10, 22]. Current U.S. government trends also indicate large investments in standards-based electronic health information systems. As standards emerge for electronic health records, detailed patient history will be available. If this additional patient information can be synthesized, better and more customized care could result. In general, the emerging innovations in the field of informatics point to an environment where a wealth of useful data will be available at the ICU bedside for use in systems that automatically assist caregivers.

One approach to understanding a patient is to focus on the patient's risk of death. A real-time risk model — or real-time *acuity* model — could track important changes in a patient's risk profile. More volatile patient states presumably have patterns that are associated with a greater risk of mortality. A real-time acuity score could also provide more frequent outcome prognoses than the current daily severity of illness scores. Clinically, the value of a real-time acuity score remains uncertain. How should a patient's care change if the score changes from a 50% chance of survival to a 60% chance of survival? Such changes are unlikely to be useful in determining the patient's care. However, if the model could detect (1) acute deterioration in the patient's state or (2) insidious changes in state over the course of a day, then it could potentially help interpret abundant ICU data more rapidly and perhaps more accurately.

In this thesis, I investigate a real-time general acuity model for intensive care patients. The acuity model that I explore is based on a patient's risk of near-term mortality. I first contrast my real-time acuity model with daily acuity models and existing severity of illness scores, and then I examine the performance of my general acuity model in the context of secondary outcomes. For comparison, a variety of models that predict secondary outcomes directly are developed and discussed. Unlike existing daily scores, which generally emphasize simplicity, my models utilize a variety of computationally intensive inputs as well as caregiver interventions. Furthermore, in contrast to a daily point score, a real-time acuity score can offer a detailed summary of a patient's risk profile over time.

1.2 Outline of Thesis

This thesis is organized into the following chapters:

- Chapter 2 provides an overview of existing severity of illness scores with a particular focus on the role that time plays in severity of illness scores.

- Chapter 3 describes the data and the preparation of the data that I use to create and validate the predictive models that I consider in this report.
- Chapter 4 discusses the general methodological framework that I follow to create and validate the predictive models.
- Chapter 5 develops and describes three types of mortality models: (1) daily mortality models, (2) a stationary daily mortality model, and (3) a real-time mortality model.
- Chapter 6 examines models trained and validated on specific secondary outcomes and compares their performance with the performance of the real-time mortality model developed in Chapter 5.
- Finally, Chapter 7 concludes my thesis with a summary of the contributions that it makes to the field of medical informatics and a discussion regarding future work.

Chapter 2

Background

2.1 Severity of Illness Scores

One area where researchers have utilized large amounts of ICU data is the development of severity of illness scores. Over the past 20 years, there has been a growing interest in severity of illness scores and several mature options have emerged, including the Acute Physiology and Chronic Health Evaluation (APACHE) [39], the Simplified Acute Physiology Score (SAPS) [45], the Mortality Prediction Model (MPM) [52], and several more recent generations of each of these scores [37, 38, 44, 51, 46]. The APACHE score was constructed using an expert clinical panel to select variables and denote levels of severity for each. The SAPS metric was designed similarly to APACHE, but its designers sought to match the APACHE performance using a simpler (and less time consuming to calculate) model. The APACHE and SAPS metrics both provide a point score at 24 hours after admission that indicates the illness severity for the patient. The MPM model took a different approach, using a more objective, forward stepwise selection methodology to select important variables. Unlike APACHE and SAPS, the MPM provides the patient's mortality probability directly and was constructed for multiple time points: at admission and 24 hours after admission. More recent MPM models have been constructed for 48 hours and 72 hours after admission [50]. Prominent severity of illness scores have been validated on large multi-center databases — or, in the case of SAPS and MPM, large international databases. Recent work by Ohno-Machado et al. in [63] provides a thorough review of severity of illness scores.

The original intent of severity scores was to compare groups of patients and to stratify patient populations between hospitals. Despite warnings from many of the original researchers and several studies (e.g., [76]), many caregivers have come to expect the availability of a severity score to assist them in treating individual patients. The fact is that despite their inadequacy for individual care, severity of illness scores are not going away [27]. Many researchers have validated the use of severity of illness scores in settings that deviate from their original design. Alternative settings have included populations such as coronary care patients or subarachnoid hemorrhage patients or days subsequent to the initial 24 hours after admission [80, 26, 79, 72]. The performance under these alternate settings has been generally moderate.

The underlying models behind existing severity metrics remain quite simplistic. Many of their original constraints are arguably unnecessary given the nearly ubiquitous availability of computing power today. The traditional advantage of SAPS, i.e., its simplicity, makes little difference in an environment where data are automatically collected and processed by a computer. In fact, using digital data, it is now feasible to include complicated derived features, such as long term trends or deviation from a patient's baseline, as possible inputs. Features that capture trends, patient-specific abnormalities, or important patterns in various observations should provide additional insight into the patient's underlying stability. On the other hand, caregivers appreciate simple models because of their comprehensibility and they are hesitant to use decision support systems that they do not understand. Another advantage of simplicity is the ability to calculate scores from widely available observations allowing scores to be easily implemented across different hospitals and diverse patient populations. In order to surmount the obstacles presented by a more abstruse system that requires advanced infrastructure for implementation, a real-time acuity score needs to offer clear benefits to the caregiver's daily tasks by providing sensitive but specific assessment calibrated for individual patients.

2.1.1 Organ Dysfunction Scores

While the intent of the general severity indexes has been to provide mortality risk assessment, complementary work has been done to develop organ dysfunction scores to assess patient morbidity. One such score, the sepsis-related organ failure assessment (SOFA) score, seeks to "describe a sequence of complications in the critically ill" [92, 91]. The SOFA score is limited to 6 organs by looking at respiration, coagulation, liver, cardiovascular, central nervous system, and renal measurements. For each organ, the score provides an assessment of derangement between 0 (normal) and 4 (highly deranged). One noteworthy feature of the SOFA score is that it uses the mean arterial pressure (MAP) along with vasopressor administration for the cardiovascular assessment. In contrast to the mortality risk provided by most severity of illness scores, the SOFA score aims to evaluate morbidity. Since its introduction, several studies have successfully applied the SOFA score to non-sepsis patients (e.g., trauma patients [1]) and the meaning of the SOFA acronym quickly morphed into Sequential Organ Failure Assessment.

Other organ dysfunction scores include the multiple organ dysfunction score (MODS), the logistic organ dysfunction score (LODS) and the multiple organ failure score [56, 43, 21]. Differentiating itself from the intervention-dependent SOFA score, the MODS score relies on what its authors refer to as the "pressure adjusted heart rate" (PAR), calculated by multiplying the heart rate by the ratio of central venous pressure to mean arterial pressure [56]. The LODS score was designed for use only during the first ICU day and combines the level of dysfunction of all organs into a single score. The association of organ dysfunction with mortality has prompted many papers to explore the use of organ dysfunction scores at predicting mortality with results that are, in general, only slightly worse than the general severity of illness scores [62, 18, 6, 48, 7, 96, 67, 86]. While the organ dysfunction scores are functionally similar to my objective in this research, the critical distinction is that I will approach the problem from the opposite direction; that is, I will look at a mortality model's ability to understand patient state whereas organ dysfunction scores were designed to reflect organ derangement and are often validated by their correlation with final patient outcome.

2.1.2 Machine versus Human

How do severity scores compare to humans? Relative performance between "objective" scores and humans is a difficult question that several studies have examined. When physicians have a low prediction of ICU survival ($< 10\%$), Rocker et al. found that the low prediction, often acted on by limiting life support, by itself predicts mortality better than the severity of illness metrics or organ dysfunction scores, thereby making the doctor's belief a self-fulfilling prophecy [68]. The advantage that physicians have at predicting mortality is supported by a variety of studies that show physicians generally outperform severity scores [83, 78, 76]. Comparisons between physicians and scoring systems all share the problem alluded to above: a physician's prognosis for an individual patient clearly influences the physician's actions. The coupling between a physician's prognosis and his or her actions is an unavoidable challenge inherent in the retrospective analysis of any intensive care episode. If the doctor is considered the gold standard, it is impossible to demonstrate improvement over his or her actions. Perhaps this observation, combined with prior experience, better calibration for individual patients, and consideration of factors not included in scoring systems is why physicians generally perform marginally better at predicting mortality. Some researchers, however, have argued from a resource utilization viewpoint that given what they consider to be reasonable performance from severity scores, automatic scores should be adopted as objective measures to prevent futile care in the costly ICU environment. It seems prudent that severity scores improve drastically — especially in terms of individual patient calibration — before such action is considered.

2.1.3 Modeling Survival

Most ICU outcome prediction models rely on logistic regression. For example, a variety of equations are available for SAPS and APACHE severity scores to convert the point score into a mortality probability. Logistic regression has the advantage of being straightforward and relatively easily to comprehend. Bayesian networks have also been used to better understand the structure of complex data. Such analysis has revealed interesting details about many complex systems. I have previously explored the application of survival models to an earlier release of the MIMIC II data. The advantage that survival techniques have, however, is limited by the absence of quality follow-up data (in general I only know which patients die in the hospital). Given the limited follow-up information, my analysis indicated that survival models perform nearly the same as logistic regression models at predicting outcome, but the fitting routines for survival models are less stable.

2.1.4 The Role of Time

A number of researchers have explored using daily severity of illness metrics. In 1993, Le Gall et al. suggested that despite likely being too time-consuming for most ICUs, daily scores would be the most "efficient way to evaluate the progression of risk of death" [44]. Rué et al. found that the mortality prediction on the current-day was the most informative — in fact, the mortality probability at admission and on previous days did not improve performance from the current day's score [72]. The importance of the current-day mortality prediction that Rué et al. observed corroborates Lemeshow et al.'s finding that the most important features change between the admission MPM model and the 24, 48 and 72 hour MPM models. The logistic regression equation also changes between 24-hour intervals to reflect an increasing probability of mortality [50]. From their observations, Lemeshow et al. make the general observation that a patient in the ICU with a "steady" clinical profile is actually getting worse.

Several others have examined the sequential assessment of daily severity scores. In 1989, Chang notably found that, using a set of criteria along with daily APACHE II scores, individual patient mortality could be predicted well with no false positives [9]. Lefering et al. argued against Chang's results and, while they found that Chang's metric could help identify high risk patients, their results caution against the use of such metrics for individual patients [47]. In Lefering et al.'s evaluation, in order to keep the false positive numbers low, the sensitivity of the estimates was severely limited. Lefering's results confirmed several previous findings such as those by Rogers et al. which caution against using daily severity scores for predicting individual outcome [69].

Ignoring the implications for individual patient prediction, others have confirmed the usefulness of daily severity scores. Wagner et al. showed strong results looking at daily risk predictions based on the APACHE III score and several additional variables such as the primary reason for ICU admission and treatment before ICU admission [94]. Wagner et al.'s study relied on over 17,440 patients from 40 U.S. hospitals. In another study by Timsit et al., daily SAPS II and the LOD score were combined to yield strong discrimination performance (ROC area of 0.826) and good calibration (Hosmer-Lemeshow C statistic of 7.14, p=0.5) [84].

The severity of illness studies pointed to above have several notable points of similarity. First, they heavily rely on existing models that have been widely adopted such as the SAPS and APACHE scores. While the wealth of studies validating existing severity scores is reassuring, the fundamental design of current severity scores is arguably obsolete. Second, besides concerns about using severity scores over periods that they were not intended for, the infrequency of existing severity scores (once per day) limits their utility for identifying acute changes in patient state.

2.2 Real-time Acuity

Following the progression from evaluation on only the first day to daily evaluation, the next step for severity scores might be pseudo real-time evaluation. Little work has been done directly to explore systemic real-time risk monitoring of ICU patients, apart from the quintessential bedside monitor that performs signal processing on an array of vital signs. Some reasons for the dearth of research in real-time risk assessment likely include the following obstacles (1) the difficulty in evaluating state tracking using heterogeneous inputs of varying temporal resolution, (2) the rich data necessary for such evaluation, (3) lack of quality data in a structured digital format. The emergence of rich, high-volume data repositories promises to rapidly mitigate the last two of these obstacles, and will hopefully provide leverage for progress on the first.

Recently, several researchers have augmented existing severity of illness metrics using readily available physiological measurements. Silva et al. defined a variety of "adverse events" based on blood pressure, oxygen saturation, heart rate, and urine output values deviating from a "normal range" for a fixed period of time. Using their real-time intermediate outcomes, Silva et al. showed enhanced mortality prediction performance [82]. Rivera-Fernández et al. defined similar physiologic alterations and also demonstrated strong performance [67]. In both cases, by using patterns of events prior to the current time the researchers were able to improve upon the performance of SAPS II. In a similar vein, Toma et al. have taken advantage of daily SOFA scores to find temporal organ failure patterns, termed "Episodes", that assist in predicting mortality [85, 86]. The studies by Toma et al., however, did not have access to the full daily records of the patients. Despite SAPS II calculations from the first day, they were unable to analyze patterns in many of the more predictive features relied upon by SAPS II.

Other researchers have explored models for predicting specific forms of deterioration such as work by Shavdia on predicting the onset of septic shock [81] or work by Eshelman et al. in providing predictive alerts for hemodynamic instability [17]. Several others have focused on predictions from high resolution trend data (e.g., 1 sample per minute) such as recent work by Cao et al. to predict hemodynamic instability from multi-parameter trends [8] or work by Ennett et al. to predict respiratory instability [16].

My goal in this thesis is to extend some of the work reviewed in this chapter. By utilizing a wealth of rich temporal data available from the MIMIC II database, I explore the development of real-time acuity models. I also explore the ability of models to predict specific clinically significant events. Through these endeavors, my work aims to contribute toward the development of advanced computer-assisted decision support in the ICU.

Chapter 3

Methods: Dataset Preparation

For the modeling and analysis presented in this thesis, I relied on data extracted from the MIMIC II database. While the MIMIC II database provides a rich collection of intensive care data, it can be difficult to understand these data and, like most real data sources that rely on human involvement, it contains a number of subtle quality issues. In preparing the data for use in this research, a variety of choices were necessary. For example, I corrected the arterial blood pressures with noninvasive measurements when the arterial line was obviously dampened. Such corrections helped make modeling with this data more reasonable. Other decisions, such as how I chose to integrate fluid inputs over time or how long I held values before I label them as missing, were also important. Understanding these decisions is important for any efforts that might try to reproduce the work discussed in this thesis.

This chapter provides a brief background of MIMIC II, a detailed summary of the MIMIC II data that I used and how I prepared it, and a number of important issues that I encountered while preparing the dataset. My hope is that this discussion will both enhance the reader's understanding of the data that my research is built on and assist future users of this data.

3.1 MIMIC II

The Multi-parameter Intelligent Monitoring for Intensive Care (MIMIC) [75] database was created to facilitate the development and evaluation of ICU decision-support systems. With data collection occurring over several years, the MIMIC II database now contains over 30,000 patients from a variety of care units at a Boston teaching hospital. New patients are constantly being added to this database; at the time that the data was extracted for this work a total of 26,647 patients were available. While one unique characteristic of this database is the high resolution waveforms for many of the patients, I do not currently use this information in my work. Instead, I rely exclusively on nurse-verified values[1] along with the intravenous medications, lab values, and ICD-9 codes.

MIMIC II also includes detailed free-text progress notes and discharge summaries for most patients. This text is frequently helpful when trying to better understand the context surrounding a particular patient's visit, the care regime the patient received, and irregularities in the numerical data.

Apart from the waveform data, MIMIC II is stored in a large relational database. In the following sections I will refer to tables (i.e., relations) and table attributes using a `fixed width` font.

3.2 Extracting the Variables

My first step was to translate the data from the relational database to a form directly suitable for modeling. The variables were collated in order to temporally synchronize them into a time-dependent matrix for each patient. Each column of this matrix represents a particular variable and each row (which I will refer to as an "instance") corresponds to a unique timestamp in a particular patient's stay. Many variables were charted hourly in the `ChartEvents` table. During sensitive episodes, however, this frequency often increases. For each unique time stamp (rounded to the nearest minute), an additional instance was created for the patient. Thus if a new observation (e.g., heart rate) was made at time t, then a unique instance is guaranteed to exist for time t in the matrix.

An important aspect in preparing the data was the method for handling variables with different temporal resolutions. For each variable I used a time-limited sample-and-hold approach. An upper time limit was specified for each variable to limit the maximum hold time. This maximum hold time was determined by independently examining the distributions of the observational frequencies for each variable. Figure 3-1 shows several examples of the observation-interval distributions that were used for this task. Hold limits were selected that covered all common measurement frequencies. For example, a chemistry variable such as BUN (most commonly measured once per day) was held for up to 28 hours. Similarly, variables that were more frequently updated, such as systolic blood pressure, were only held for 4 hours. When a variable observation was absent for a period greater than the hold window time, I labeled it as missing. Trusting that the caregivers made measurements more frequently when they were needed undoubtedly introduces additional noise into my dataset; but this negative is arguably negligible when weighted against the considerable reduction in data sparseness obtained when values were allowed to persist for a reasonable amount of time.

[1]The nurse-verified values are generally charted every hour, but this frequency varies greatly between variables and is patient-dependent.

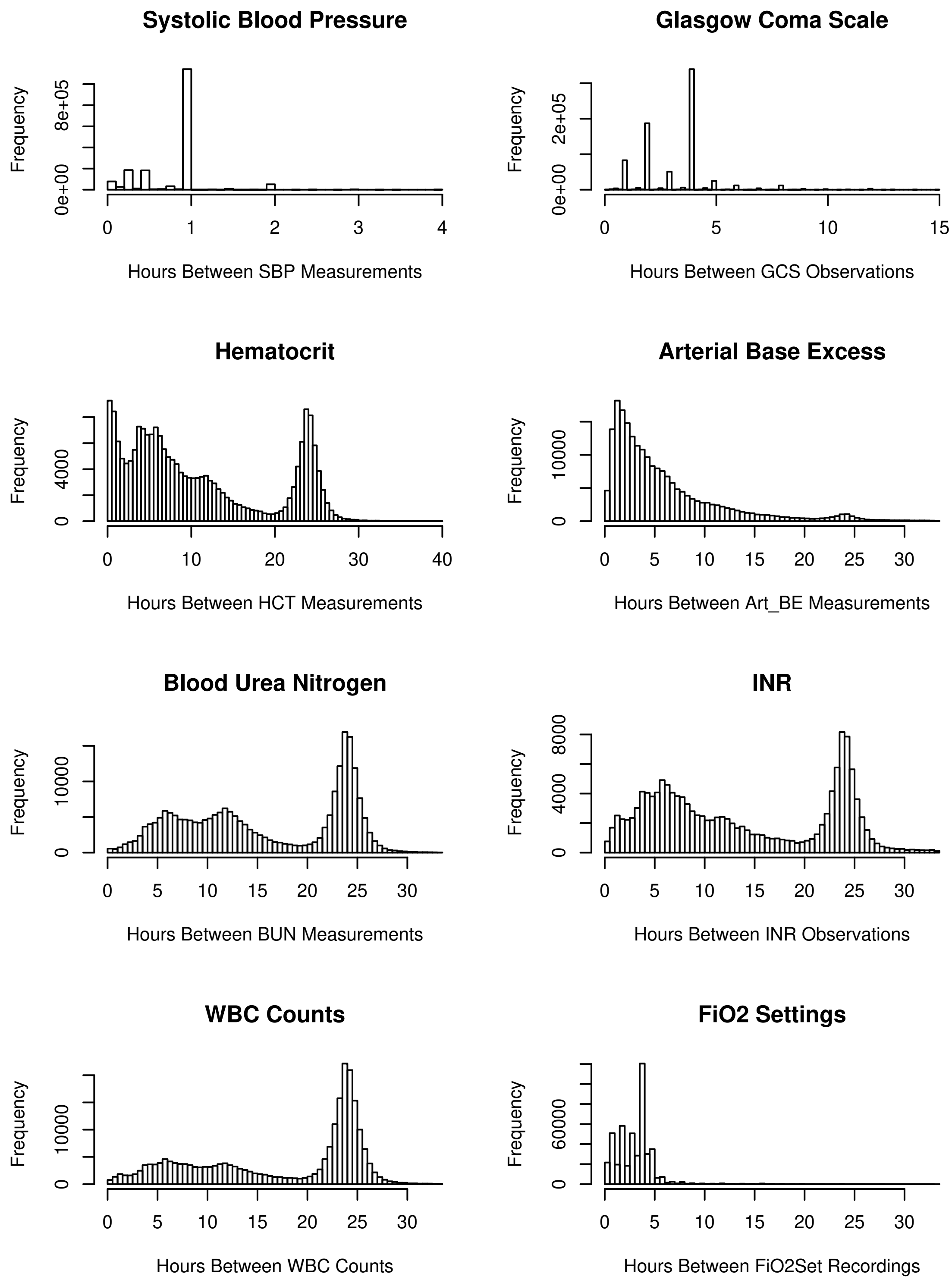


Figure 3-1: Observation frequency histograms

3.2.1 General ChartEvent Variables

The majority of the candidate variables for my models are located in the `ChartEvents` table. Many of the `ChartEvent` variables contain a numerical value, typically resulting from measurements taken from the patient. The numeric `ChartEvent` variables that I include in my dataset — such as the variable names, `ItemIDs`, hold limits and valid ranges — are provided in Table 3.1. The final column of this table is explained later in Section 3.3.

The valid ranges provided in Table 3.1 were found empirically by examining the individual distributions. Using the distributions, threshold points that discarded high and low outliers were selected. In some cases it was also necessary to consider the physiologic bounds of a particular variable. For example, obvious errors occasionally yielded a pH value of 0.076 instead of 7.6 or a temperature value of 37 "degrees Fahrenheit" instead of 98.6 degrees Fahrenheit.[2]

Table 3.1: Continuous and Ordinal `ChartEvent` Variables

Variable Name	ItemID	Hold Limit (hrs)	Units	Min Val	Max Val	Slope Wins (hrs)
Misc.						
Glasgow Coma Scale (GCS)	198	28	points	3	15	28
Weight	581	28	kg	20	300	28
AdmitWt	762	Const	kg	20	300	-
Cardiovascular						
SBP (NBPSys)	455	4	mmHg	30	250	4, 28
DBP (NBPDias)	455	4	mmHg	8	150	4, 28
MAP (NBPMean)	456	4	mmHg	20	250	4, 28
A-line SBP (SBP)	51	4	mmHg	30	300	4, 28
A-line DBP (DBP)	51	4	mmHg	8	150	4, 28
A-line MAP (MAP)	52	4	mmHg	20	170	4, 28
Heart Rate (HR)	211	4	BPM	20	300	4, 28
Resp Rate (RESP)	211	4	BPM	20	300	4, 28
SpO2	646	4	%	70	101	4, 28
CVP	113	4	mmHg	-5	50	4, 28
PAPMean	491	4	mmHg	0.1	120	4, 28
PAPsd	492	4	mmHg	0.1	120	4, 28
Cardiac Index (CrdIndx)	116	10	L/min/m^2	0.1	10	4, 28
SVR	626	10	dyn·s/cm^5	0.1	3200	4, 28
COtd	90	10	L/min	0.1	20	4, 28
COfick	89	10	L/min	0.1	20	4, 28
PCWP	504	10	mmHg	0.1	45	4, 28
PVR	512	10	dyn·s/cm^5	0.1	1000	4, 28
Chemistries						
Sodium (Na)	837, 1536	28	mEq/L	115	160	28
Potassium (K)	829, 1535	28	mEq/L	1	10	28
Chloride (Cl)	788, 1523	28	mEq/L	75	135	28
CO2	787	28	mEq/L	0.1	55	28
Glucose	811	28	mg/dL	0.1	500	28
BUN	781, 1162	28	mg/dL	0.1	180	28
Creatinine	791, 1525	28	mg/dL	0.1	40	28

[2]Some of these obvious errors appear to have been corrected in the most recent release of MIMIC II

Table 3.1 – continued from previous page

Variable Name	ItemID	Hold Limit (hrs)	Units	Min Val	Max Val	Slope Wins (hrs)
Magnesium (Mg)	821, 1532	28	mg/dL	0.01	5	28
AST	770	28	IU/L	10	1000	28
ALT	769	28	IU/L	10	1000	28
Calcium (Ca)	786, 1522	28	mg/dL	4	14	28
Ionized Ca (IonCa)	816	28	mmol/L	0	2.5	28
Total Bilirubin (TBili)	1538, 848	28	mg/dL	0.001	60	28
Direct Bilirubin (DBili)	803, 1527	28	mg/dL	0	50	28
Total Protein (TProtein)	849, 1539	28	g/dL	0.01	15	28
Albumin	772, 1521	28	g/dL	0.01	7	28
Lactate	818, 1531	28	mg/dL	0.2	40	28
Troponin	851	28	ng/mL	0.01	100	28
Hematology						
Hematocrit (HCT)	813	28	%	15	60	28
Hemoglobin (Hgb)	814	28	%	4	20	28
Platelets	828	28	10^9/L	0.1	1200	28
INR	815, 1530	28	-	0.01	12	28
Prothrombin time (PT)	824, 1286	28	s	0.01	36	28
PTT	825, 1533	28	s	10	151	28
WBC Count (WBC)	861, 1127, 1542	28	$10^3/\mu$L	0.01	70	28
RBC Count (RBC)	833	28	$10^6/\mu$L	1	7	28
Temp	678, 679	28	Deg F	80	110	28
Arterial Blood Gases						
Art Base Excess (Art BE)	776	28	mmol/L	-40	30	28
Art CO2	777	28	mEq/L	1	60	28
Art PaCO2	778	28	mmHg	5	100	28
Art PaO2	779	28	mmHg	0.1	500	28
Art pH	780, 1126	28	-	6.5	8.5	28
Ventilation						
FiO2Set	190	28	torr	0.1	1	28
PEEPSet	506	28	cmH20	0	50	28
Resp Rate Tot (RespTot)	615	28	BPM	0.1	50	28
Resp Rate Set (RespSet)	619	28	BPM	0.1	40	28
Resp Rate Spon (RespSpon)	614	28	BPM	0.001	40	28
Peak Insp Pres (PIP)	535	28	cmH2O	5	60	28
PlateauPres	543	28	cmH2O	5	60	28
Tidal Vol Obs (TidVolObs)	682	28	mL/B	100	1100	28
Tidal Vol Set (TidVolSet)	683	28	mL/B	50	1001	28
Tidal Vol Spon (TidVolSpon)	684	28	mL/B	0.1	1200	28
SaO2	834	28	%	80	101	28

3.2.2 Categorical Variables

A number of the MIMIC II `ChartEvents` variables are categorical in nature. These variables were handled separately as two types: ordinal and binary. If a variable contained a natural progression in its categories and this ordering was deemed potentially useful, the variable was labeled with integer values starting at 1 (least severe) and progressing to n (most severe). The remaining categorical variables, without a natural order, were coded using binary indicator variables. A binary indicator variable was marked 1 (True) if the corresponding categorical variable's value belongs to a specified subset of the possible values; otherwise, it retained its default value of 0 (False).

The categorical variables that I included, along with the coding schemes for the variables derived from them, are shown in Table 3.2. As in Table 3.1, Table 3.2 also includes the maximum number of hours that the variable will be held in the absence of an updated value (the "Hold Limit" column).

Table 3.2: Categorical `ChartEvent` variables. Type 'Ord" indicates ordinal variable and type "Bin" indicates binary variable. Value T indicates "True". If no matching category found, default value is 0 (Ord) or False (Bin).

Label (`ItemID`)	Variable	Type	Hold (hrs)	(Value) Category
Heart Rhythm (212)				
	hrmHB	Ord	3	(1) 1st Deg AV Block; (2) 2nd AVB Mobitz 2; (3) 2nd AVB/Mobitz I; (4) Wenckebach; (5) Comp Heart Block
	hrmPaced	Bin	3	(T) Paced; (T) A Paced; (T) AV Paced; (T) V Paced; (T) Zoll Paced
	hrmSA	Ord	3	(1) Parox Atr Tachy; (2) Sinus Arrhythmia; (3) Supravent Tachy; (4) Wand.Atrial Pace; (5) MultiFocalAtrTrach; (6) Atrial Fib; (7) Atrial Flutter
	hrmVA	Ord	3	(1) Junctional; (2) Idioventricular; (3) Vent. Tachy; (4) Ventricular Fib; (5) Asystole
Ectopy Type (161)				
	PVC	Bin	3	(T) PVC's; (T) V Quadrigeminy; (T) Vent. Trigeminy; (T) Vent. Bigeminy
	PAC	Bin	3	(T) PAC's; (T) A Quadrigeminy; (T) Atrial Trigeminy; (T) Atrial Bigeminy
	PNC	Bin	3	(T) PNC's; (T) N Quadrigeminy; (T) Nodal Trigeminy; (T) Nodal Bigeminy
Ectopy Frequency (159)				
	EctFreq	Bin	3	(T) Rare; (T) Occasional; (T) Frequent; (T) Runs Vtach
Code Status (128)				
	DNI	Bin	3	(T) Do Not Intubate
	NoCPR	Bin	3	(T) CPR Not Indicate
	DNR	Bin	3	(T) Do Not Resuscita
	ComfortMeas	Bin	3	(T) Comfort Measures
	OtherCode	Bin	3	(T) Other/Remarks
	FullCode	Bin	3	(T) Full Code
Risk for Falls (1484)				
	FallRisk	Bin	3	(T) Yes
Orientation (479)				
	orientation	Ord	5	(1) Oriented x 3; (2) Oriented x 2; (3) Oriented x 1; (4) Disoriented
	orientUnableAs	Bin	5	(T) Unable to Assess

Continued on next page

Table 3.2 – continued from previous page

Label (`ItemID`) Variable	Type	Hold (hrs)	(Value) Category
Riker SAS (1337)			
RikerSAS	Ord	5	(1) Unarousable; (2) Very Sedated; (3) Sedated; (4) Calm/Cooperative; (5) Agitated; (6) Very Agitated; (7) Danger Agitation
Ventilator Type (722)			
Vent	Bin	5	(T) 7200A; (T) Drager; (T) Other/Remarks; (T) Servo 900c
Ventilator Mode (720)			
VentMode	Ord	5	(1) Assist Control; (2) CMV; (3) CPAP; (4) CPAP+PS; (5) Pressure Control; (6) Pressure Support; (7) SIMV; (8) SIMV+PS; (9) TCPCV; (10) Other/Remarks
Pacemaker (516)			
pacemkr	Bin	28	(T) Epicardial Wires; (T) Permanent; (T) Transcutaneous; (T) Transvenous
Trach Size (690)			
trach	Bin	10	(T) #4; (T) #5; (T) #6; (T) #7; (T) #8; (T) #9; (T) #10; (T) Other/remarks
Skin Color (643)			
paleSkin	Bin	10	(T) Pale; (T) Ashen; (T) Dusky; (T) Cyanotic
flushSkin	Bin	10	(T) Flushed; (T) Mottled
jaundiceSkin	Bin	10	(T) Jaundiced
Skin Integrity (644)			
impairedSkin	Bin	10	(T) Absent; (T) Impaired; (T) Other/Remarks
IABP Setting (225)			
iabp	Bin	28	(T) 1:1; (T) 1:2; (T) 1:3; (T) 1:4
iabpVal	Ord	0	(1) 1:4; (2) 1:3; (3) 1:2; (4) 1:1
Service Type (1125)			
svOther	Bin	28	(T) Other
svCSICU	Bin	28	(T) CSICU
svNSICU	Bin	28	(T) NSICU
svMICU	Bin	28	(T) MICU
svMSICU	Bin	28	(T) MSICU
svCCU	Bin	28	(T) CCU
svCSRU	Bin	28	(T) CSRU

3.2.3 Medications

Intravenous medications administered during a given patient's stay are recorded in the MIMIC II `MedEvents` table. Unlike the observational variables found in the `ChartEvents` table, the `MedEvents` table records active intervention by the caregivers. These medications are often quite important in interpreting the observational variables. Table 3.3 lists the medications that are commonly given during a patient's

ICU stay. Each of these medications were included in my dataset.

Table 3.3: `MedEvent` variables (Intravenous Medications)

Medication	`ItemID`	Units	Medication	`ItemID`	Units
Aggrastat	110	mcg/kg/min	Lepirudin	177	mg/kg/hr
Amicar	111	gm/hr	Levophed	47	mcg/min
Aminophylline	3	mg/hr, mg/kg/hr	Levophed-k	120	mcg/kg/min
Amiodarone	112	mg/min	Lidocaine	48	mg/min
Amrinone	40	mcg/kg/min	Midazolam	124	mg/hr
Argatroban	173	mcg/kg/min	Milrinone	125	mcg/kg/min
Ativan	141	mg/hr	Morphine Sulfate	126	mg/hr, mg/kg/hr
Atracurium	113	mg/kg/hr	Narcan	148	mcg/kg/min
Bivalirudin	174	mg/kg/hr	Natrecor	172	mcg/kg/min
Cisatracurium	114	mcg/kg/min, mg/kg/hr	Neosynephrine	127	mcg/min
Dilaudid	163	mg/hr	Neosynephrine-k	128	mcg/kg/min
Diltiazem	115	mg/hr	Nicardipine	178	mcg/kg/min
Dobutamine	42	mcg/kg/min	Nitroglycerine	49	mcg/min
Dopamine	43	mcg/kg/min	Nitroglycerine-k	121	mcg/kg/min
Doxacurium	116	mg/kg/hr	Nitroprusside	50	mcg/kg/min
Epinephrine	44	mcg/min	Pancuronium	129	mg/kg/hr
Epinephrine-k	119	mcg/kg/min	Pentobarbitol	130	mg/kg/hr
Esmolol	117	mcg/kg/min	Precedex	167	mcg/kg/hr
Fentanyl (Conc)	149	mcg/hr	Procainamide	52	mg/min
Fentanyl	118	mcg/hr	Propofol	131	mcg/kg/min
Heparin	25	U/hr	Reopro	134	mcg/kg/min, mcg/min
Insulin	45	U/hr	Sandostatin	133	mcg/hr
Integrelin	142	mcg/kg/min	TPA	135	mg/min
Ketamine	151	mcg/kg/hr, mcg/kg/min	Vasopressin	51	U/hr, U/min
Labetolol	122	mg/min	Vecuronium	138	mg/kg/hr
Lasix	123	mg/hr			

One implementational difficulty in adding these variables to our dataset was understanding their duration. In general, medications are administered at a certain dose per unit time, and this dose is repeated every hour (even if the dose does not change). Often, when a medication is discontinued, a zero dose is recorded for the last value. In other instances the `Stopped` column is marked with “Stopped” or “D/C’d”. In more difficult cases, there is no indication that the medication ended and I considered the last recorded value to be the end. However, there are exceptions to these rules. One case arises when the medication `ItemID` changes but the same medication is being administered with different units. For example, Neosynephrine (`ItemID` 127) given in mcg/min might be changed to Neosynephrine-k (`ItemID` 128) given in mcg/kg/min.

Each medication was added to the dataset in three ways. Some medications are only administered in per-kilogram units while other medications are measured in absolute dose. A number of medications, however, are entered using either method. Consequently, I used the weight of a patient to add the absolute dose variable for each medication and the dose per-kilogram variable for each medication. Finally, I also mapped each medication to a more generic category and added a binary indicator variable to indicate if that type of medication was present (e.g., if Neosynephrine is being given, then the Sympathomimetic_agent variable is flagged).

3.2.4 Input/Output Variables

Patient Input/Output (IO) observations are recorded in the MIMIC II `IOEvents` table. A related table, `TotalBalEvents`, provides patient IO balances and 24-hour summaries. Using observations from these tables, two types of variables were created.

First, IO variables were added using the the `IOEvents` table. These variables are listed in Table 3.4. For the generic "Output" variables, all outputs recorded from the patient over the given window were summed (these are `ItemID`s with a `Category` value of null in the `D_IOItems` table). Similarly, for the "Input" variables, all recorded inputs over the given window were summed. To preserve the ability to later sum the total IO for a given time range (e.g., hourly urine output), each variable was included with and without a hold window. The variables that do not have a hold window have suffixes of "B".

Table 3.4: `IOEvents` variables

Variable Name	Hold (hrs)	Units	Window Length (min)
AllInput	4	ml	all previous
Input_60	4	ml	60
InputB_60	0	ml	60
InputB	0	ml	-
AllOutput	4	ml	all previous
Output_60	4	ml	60
OutputB_60	0	ml	60
OutputB	0	ml	-
UrineOut	4	ml	-
UrineOutB	0	ml	-
InputRBCs	4	ml	-
InputRBCsB	4	ml	-
InputOtherBlood	4	ml	-
InputOtherBloodB	0	ml	-

The general "Input" and "Output" variables contain all the recorded IV inputs and the all of the recorded outputs, respectively. It is often helpful to separate the input and output into subcategories. To accomplish this, three variables that only looked at special types of IO were created. These variables were named UrineOut, InputRBCs, and InputOtherBlood, and are included in Table 3.4. To understand the measurements that contribute to the specific IO variables, Table 3.5 lists the constituent `ItemID`s for these variables.

Table 3.5: `ItemID`s used for summary IO variables

Variable Name	Item Labels	Units	`ItemID`
UrineOut			
	Urine Out Foley	ml	55
	Urine Out Void	ml	69
	Urine Out Suprapubic	ml	715
	OR Out OR Urine	ml	61
	Urine Out Rt Nephrostomy	ml	57
	Urine Out Lt Nephrostomy	ml	57
	Urine Out Incontinent	ml	85
	Urine Out LleoConduit	ml	473
	Urine Out Other	ml	405
	Urine Out Straight Cath	ml	428
	Urine Out Ureteral Stent #1	ml	428
InputRBCs			
	packed RBC's	ml	144
	OR Packed RBC's	ml	172
	Packed RBC's 375.0ml	ml	398
InputOtherBlood			
	Platelets	ml	179
	OR Platelets	ml	224
	Platelets 440.0ml	ml	3955
	Fresh Frozen Plasma	ml	163
	Cryoprecipitate	ml	319
	Whole blood	ml	221
	Other Blood Products	ml	221

A second type of IO variable that I included in my dataset represents longer-term IO summaries. These variables were extracted from the `TotalBalEvents` table. The `TotalBalEvents` table contains IO balances and 24-hour summations for various IO items in the `IOEvents` table. Daily summations were calculated at 23:00 for each day of a patient's stay. The variables, along with their hold windows are listed in Table 3.6.

Table 3.6: `TotalBalEvents` variables

Variable Name	Units	`ItemID`	Hold Window (hrs)
TotIn24	ml	1	28
TotOut24	ml	2	28
TotIV24	ml	18	28
UrOut24	ml	26	28
Bal24	ml	27	28
LOSBal	ml	28	28

3.2.5 Demographic Variables

In addition to the variables discussed above, there were a number of additional variables that were helpful to include in my dataset. Many of these variables loosely fell under the demographic category, and included indicators such as chronic illnesses from ICD-9 codes or the physical location of the patient (e.g., Medical ICU). These demographic variables are described in Table 3.7.

Table 3.7: Demographic Variables

Variable Name	Description and Source
Sex	`D_Patients Sex`
Age	First ICU admission year - year of birth (`CensusEvents InTime` - `D_Patients DOB`)
hospTime	Minutes in hospital prior to ICU admission; found by using the difference between the first ICU admission (`CensusEvents InTime`) and the last hospital admission (`Admissions Adm_Dt`) that is before the first ICU admission
AIDS	Present if `ICD9 Code` matches regular expression "ˆ042"
HemMalig	Hematologic Malignancy; present if `ICD9 Code` matches regular expression "ˆ20[0-8]"
MetCarcinoma	Metastatic Carcinoma; present if `ICD9 Code` matches regular expressions "ˆ1[4-5][0-9]", "ˆ1[6-7][0-5]", "ˆ179", or "ˆ1[8-9][0-9]"
SICU	Physically located in the T-SICU (`CUID` = 74, or 53)
MSICU	Physically located in the MSICU (`CUID` = 72)
MICU	Physically located in the MICU (`CUID` = 70, 69, or 126)
CCU	Physically located in the CCU (`CUID` = 1 or 3)
CSRU	Physically located in the CSRU (`CUID` = 54, 124, or 125)

3.3 Derived Variables

In addition to the variables that were directly extracted from the data, a number of additional variables were calculated. There were numerous motivations for including these variables. First, meta-information, such as the presence or absence of measurements is often quite informative. Second, it is often helpful to interpret a variable's value relative to another variable's value. Another motivating factor is the loss of information when a single variable (e.g., Heart Rate) is observed in isolation without examining prior history; often the temporal behavior of a variable shares equal importance with the variable's value.

3.3.1 Meta Variables and Calculated Variables

First, to capture the presence or absence of particular measurements, I created a number of indicator variables. These variables were labeled 1 (True) if the corresponding variable was available and labeled 0 (False) if the variable was missing. The names of these indicator variables end with a capital "M". The variable "CVPM", for example, indicates if CVP measurements are available.

Many potentially interesting variables can be calculated from the variables discussed thus far. Some simple examples include the BUN-to-Creatinine ratio or the pulse pressure. More complex calculations include the cumulative time that the patient has spent on vasopressors, the number of vasopressors that patient is on, or the hourly urine output rate. These variables, and many others, are described in Table 3.8.

3.3.2 Variables from Literature

A number of papers have suggested features that may help predict patient mortality. For example, Rivera-Fernández et al. suggest several types of events and demonstrate that the number of times that each of these events occur can help enhance mortality prediction models [67]. Silva et al. suggested a variety of similar events that they used with artificial neural networks to predict mortality [82]. For my dataset, I added a number of calculated variables that were inspired by Rivera's and Silva's work. These new variables are listed in Table 3.9.

3.3.3 Slopes, Ranges, and Baseline Deviations

Finally, I calculated variables that attempt to capture the temporal behavior of various variables. For example, I added variables that indicate the relative change over time for the continuous `ChartEvent` variables found in Table 3.1. I did this by using the raw sampled values (with no holding) to calculate the per-minute slope of the best-fit line for various fixed-length windows prior to the particular instance. For

Table 3.8: Derived Variables

Variable Name(s)	Description
VasopressorsM	The patient is on at least one vasopressor (Vasopressin, Neosynephrine, Levophed, Dopamine, or Epinephrine)
PressorsM	The patient is on at least one pressor (Vasopressin, Neosynephrine, Levophed, Dopamine, or Epinephrine, Dobutamine, Milrinone, or Amrinone)
SedativesM	The patient is on at least one sedative (Propofol, Pentobarbitol, Ativan, Midazolam, Ketamine, Dilaudid, Fentanyl, Morphine Sulfate)
CVPM	A CVP measurement is available
COtdM	A COtd measurement is available
PCWPM	A PCWP measurement is available
CrdIndxM	A Cardiac Index value is available
PAPmeanM	A PAPmean measurement is available
HCTM	A HCT measurement is available
LactateM	A lactate measurement is available
MechVent	The patient is mechanically ventilated
SBPm, DBPm, and MAPm	Merged blood pressure values. When the invasive blood pressure is not available, the noninvasive pressure is used. To correct for arterial-line dampening, a noninvasive value is used instead of an invasive value if the invasive systolic value is more than 15% less than the noninvasive systolic value or the invasive diastolic value is more than 15% greater than the noninvasive diastolic value.
VasopressorSum.std	Each vasopressor dose is standardized (by dividing the dose by the mean dose) and then summed together
PressorSum.std	Each pressor dose is standardized (by dividing the dose by the mean dose) and then summed together
Pulse Pressure (PulsePres)	SBPm - DBPm
Est. Cardiac Output (ECO)	0.5 * (HR * (SBPm - DBPm))/MAPm
ECOSlope	The slope of the best-fit line over the preceding six-hours
Shock Index (ShockIdx)	HR/SBPm
UrineByHr	Hourly urine output
BUN:Creatinine (BUNtoCr)	BUN/Creatinine
PaO2:FiO2 (PaO2toFiO2)	Ordinal value indicating PaO2:FiO2 ratio: (0) patient is not ventilated or ventilated and ratio is greater than 300; (1) patient is ventilated and ratio is between 300 and 200; (2) patient is ventilated and ratio is between 200 and 100; and (3) patient is ventilated and ratio is less than 100
DopSm, DopMd, DopLg	Small (less than 2 mcg/kg/min), Medium (between 2 and 10 mcg/kg/min), and Large (greater than 10 mcg/kg/min) doses of dopamine
VentLen	The number of contiguous minutes prior to the current time that patient has been on a mechanical ventilator
VentLenC	The cumulative number of minutes prior to the current time that the patient has been on a mechanical ventilator
PressorTime	The number contiguous minutes prior to the current time that the patient has received vasopressor medications
CumPressorTime	The cumulative number of minutes prior to the current time that the patient has received vasopressor medications
SBPm.pr	The ratio of the average SBP while on vasopressor medications to the average SBP while not on vasopressor medications (up to current time)
MAPm.pr	The ratio of the average MAP while on vasopressor medications to the average MAP while not on vasopressor medications (up to current time)
PressD01	Vasopressor medications were first initiated during the first 24 hours in the ICU
PressD12	Vasopressor medications were first initiated during the second day in the ICU
PressD24	Vasopressor medications were first initiated during the third or forth day in the ICU
PressD4	Vasopressor medications were first initiated after the forth day in the ICU
BPcor	The correlation between the MAP and the pressorSum.std up to the current time
PressorCnt	The total number of vasopressors that the patient is on up to the current time

Table 3.9: Variables from Literature

Variable Name	Description	Normal Range	win (m)
SBPm.oor30c	Minutes continuously out of range	90-180 mmHg	30
SBPm.oor30t	Total minutes out of range	90-180 mmHg	30
SBPm.oor120c	Minutes continuously out of range	90-180 mmHg	120
SBPm.oor120t	Total minutes out of range	90-180 mmHg	120
SpO2.oor30c	Minutes continuously out of range	$\geq$ 90%	30
SpO2.oor30t	Total minutes out of range	$\geq$ 90%	30
SpO2.oor120c	Minutes continuously out of range	$\geq$ 90%	120
SpO2.oor120t	Total minutes out of range	$\geq$ 90%	120
HR.oor30c	Minutes continuously out of range	60-120 bpm	30
HR.oor30t	Total minutes out of range	60-120 bpm	30
HR.oor120c	Minutes continuously out of range	60-120 bpm	
HR.oor120t	Total minutes out of range	60-120 bpm	120
UrineByHr.oor60c	Minutes continuously out of range	$\geq$ 30	60
UrineByHr.oor120c	Minutes continuously out of range	$\geq$ 30	120
SBPThreshCnt	Number of SBP threshold events	90-180 mmHg	-
SBPThreshCntN	Number of hourly SBP threshold events[a]	90-180 mmHg	-
SBPThreshCntF	Fraction of instances with SBP threshold event	90-180 mmHg	-
SpO2LowCnt	Number of low SpO_2 values	$\geq$ 90%	-
SpO2LowCntN	Number of low hourly SpO_2 values[a]	$\geq$ 90%	-
SpO2LowCntF	Fraction of instances with low SpO_2	$\geq$ 90%	-
HRThreshCnt	Number of HR threshold events	60-120 bpm	-
HRThreshCntN	Number of hourly HR threshold events[a]	60-120 bpm	-
HRThreshCntF	Fraction of instances with HR events	60-120 bpm	-
UrLowCnt	Number of low hourly urine events	$\geq$ 0.5 ml/kg/hr	-

[a] When multiple observations are available in less than one hour, the worst observation is used

each of these window lengths (e.g., 28 hr), a new variable was added. The slope windows that were used for particular variables are indicated in the last column in Table 3.1.

Two other ways that I explored for capturing the history of a variable included calculating the range of a patient's previous values and calculating the deviation of a value from the patient's evolving baseline. The range variables indicate the difference between maximum and minimum values seen previously in the patient's stay. The deviation from baseline variable is a little more elaborate. For each instance over a patient's stay, this variable represents the current value minus the mean value of previous instances in the patient's stay. Range and baseline deviation variables were added for the following subset of previously defined variables:

Table 3.10: Variables with Range and Baseline Deviation Calculations

SBPm	DBPm	MAPm
HR	Weight	HCT
Hgb	INR	Art_pH
BUN	LOSBal	Lactate
GCS	PT	Input_60
Creatinine		

3.4 Preliminary Dataset

Using the variables described above, I put together my preliminary dataset. In doing this, I excluded patients who did not contain commonly observed variables. This process eliminated a number of patients whose data were incomplete or poorly represented (e.g., patients who died or were discharged only after a few hours in the unit). I required the following criteria to be met:

- At least one BUN observation (19275 patients)
- At least one GCS observation (18735 patients)
- At least one Hematocrit observation (18850 patients)
- At least one HR observation (26029 patients)
- At least one IV medication recorded in `MedEvents` (14833 patients)
- Receive adult care (are not neonates) (19878 patients)

By requiring all these criteria to be met, the total number of patients was reduced from 26647 to 13923.

3.4.1 Descriptive Statistics

In order to better understand the preliminary dataset, I calculated a number of descriptive statistics. For brevity, I limit my graphical analysis to a subset of the 438 variables described in this chapter (for a complete list of the variables in the final dataset, see Appendix B). Figure 3-2 provides histograms describing demographic information in the preliminary dataset. In addition, it is potentially helpful in understanding the patient population to visualize the distributions for a few variables in this preliminary dataset. Figures 3-3 and 3-4 provide histograms with descriptive statistics for the variables that are required for SAPS II.

3.4.2 Multiple Hospital Visits

While most of the MIMIC II ICU patients only have one hospital admission, many patients have multiple hospital visits. For the purposes of this research, I limited my analysis to a given patient's first recorded ICU admission. This generally corresponds to the first hospital admission information available.[3] Due to the snapshot nature of the data, there is no guarantee that the first recorded ICU visit is actually the patient's first ICU visit (e.g., they may have visited a year prior to the start of the data collection). This approach does, however, omit ICU readmissions where prior ICU information is known. Table 3.11 provides the number of hospital and ICU admissions I have for the patients in the `CensusEvents` table, the preliminary dataset discussed here, and the final dataset presented in the following section. Figure 3-5 shows that most patients have only one recorded admission and that about one quarter of the patients are responsible for all readmissions[4] (and the number of these readmissions has a long tail with a max of 33[5]). An alternative strategy might be to use the first ICU stay from the last known hospital admission for a patient, but this limits the follow-up information for a number of patients. There remains a strong case, however, for using the first ICU admission of a hospital stay as a number of studies have shown that ICU readmissions are correlated with increased mortality and hospital length of stay [70, 59].

3.4.3 Mortality

Another necessary decision for my dataset was how to define mortality. Many of the severity of illness metrics (e.g., SAPS II) provide a prediction of hospital mortality, but as noted above, the MIMIC II data often includes multiple hospital visits for

---

[3] Generally only hospital admissions that include at least one ICU admission are captured by MIMIC II.

[4] Since the `CensusEvents` table has a separate entries for every time a patient enters/leaves the ICU, it was necessary to define an ICU stay as an ICU period that ignores gaps of up to a maximum of 24 hours (e.g., periods where the patient is in the operating room).

[5] This max occurs for `SUBJECT_ID` 13033 over a time of five years

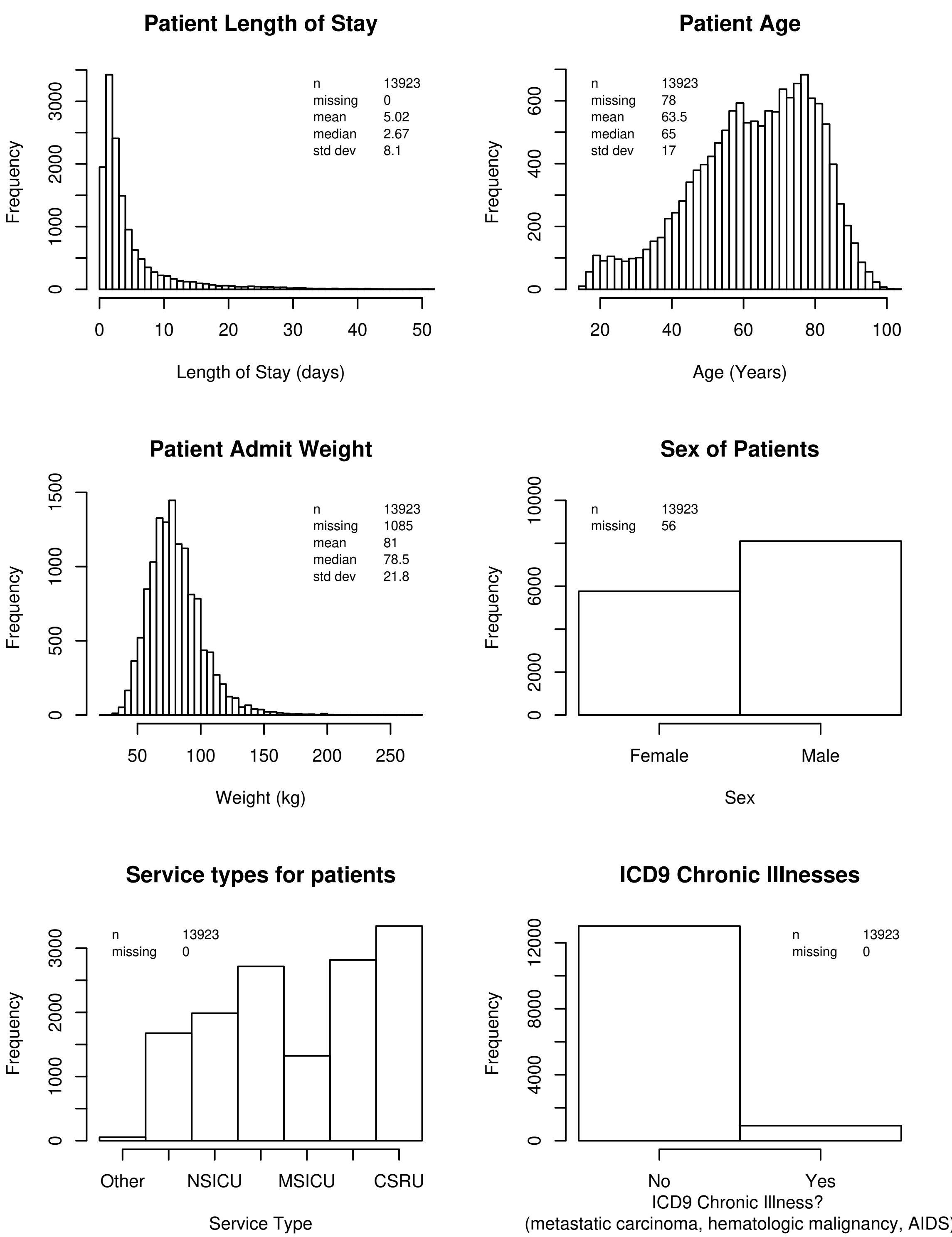


Figure 3-2: Histograms for demographic information

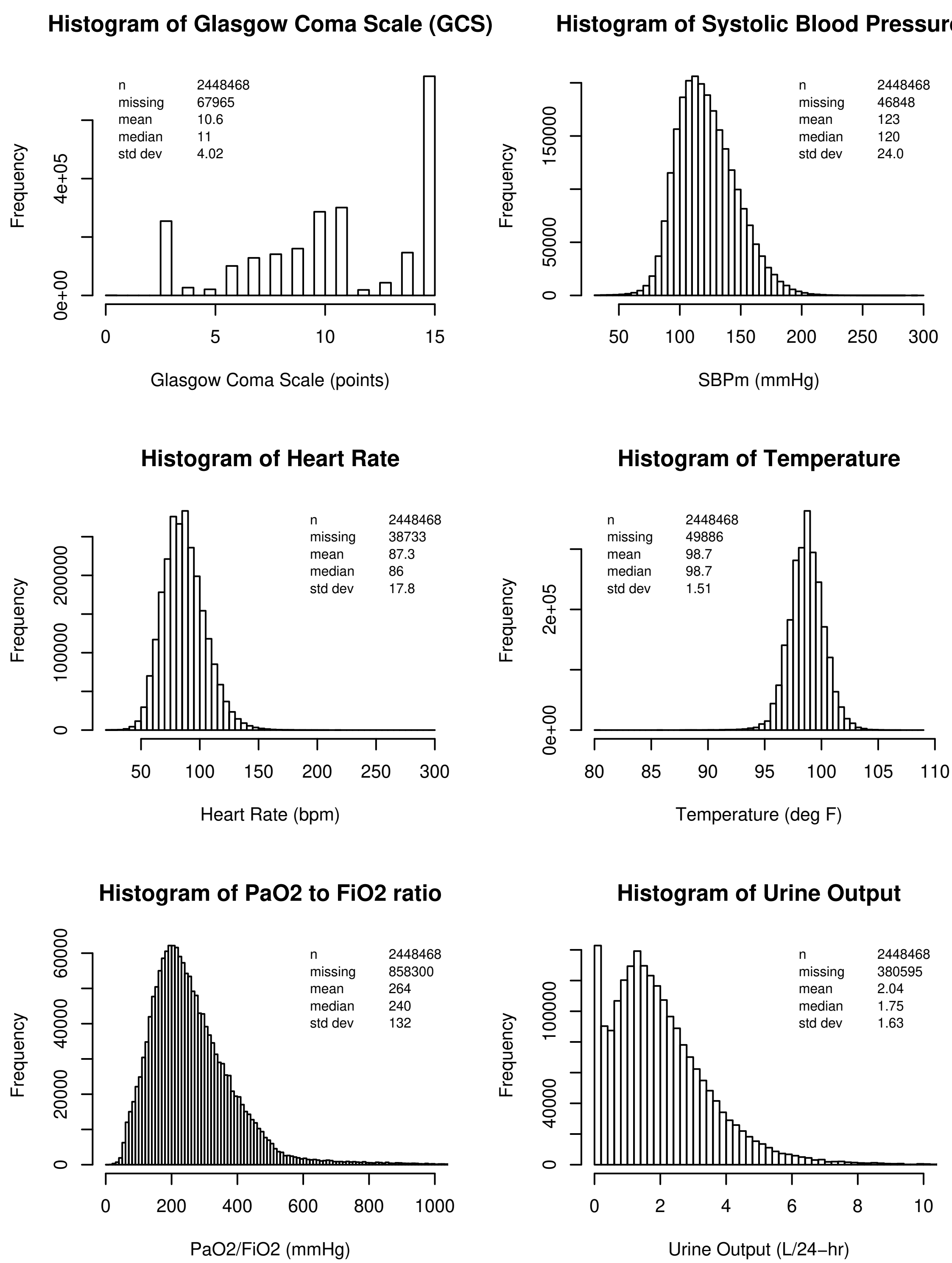


Figure 3-3: SAPS II Variables

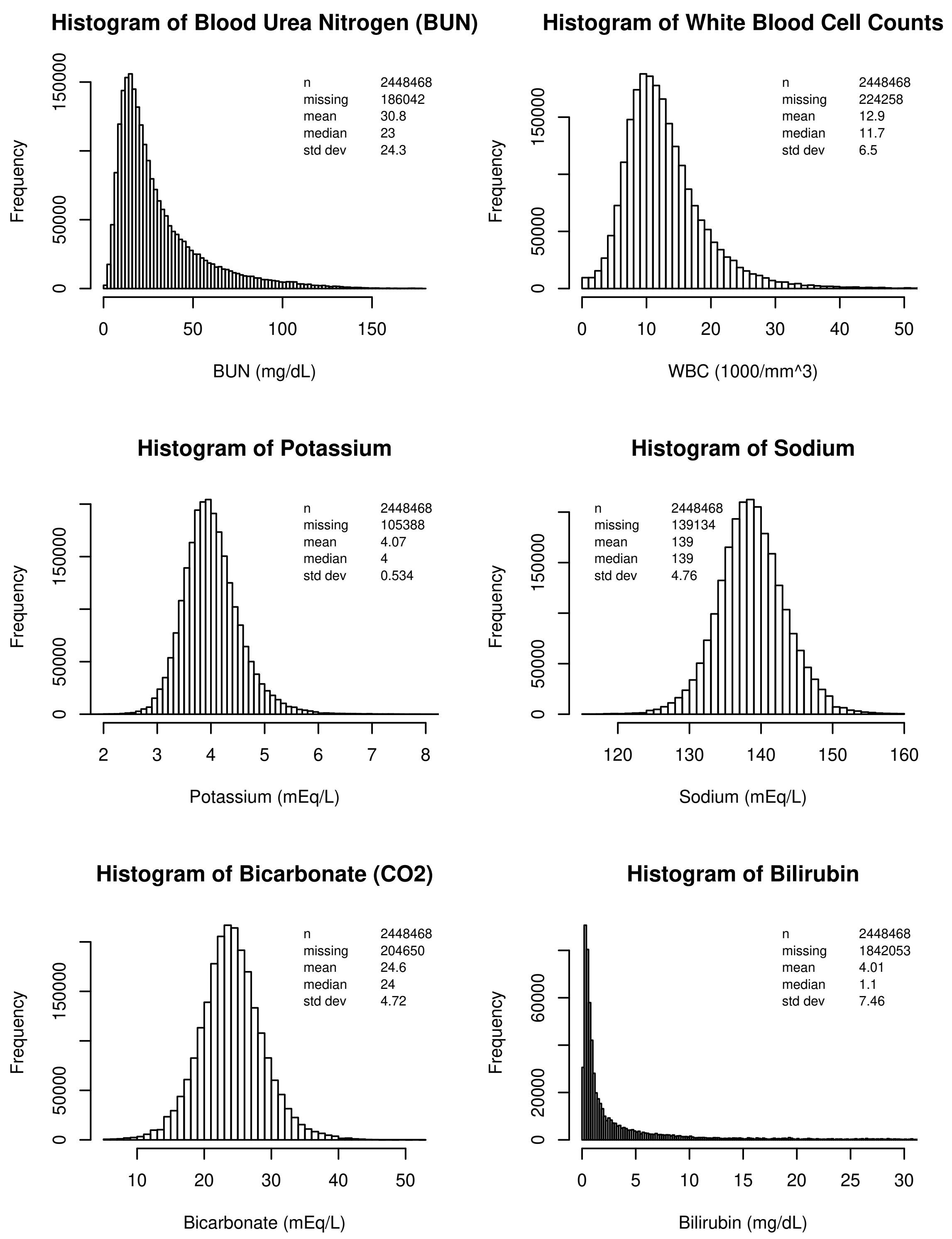
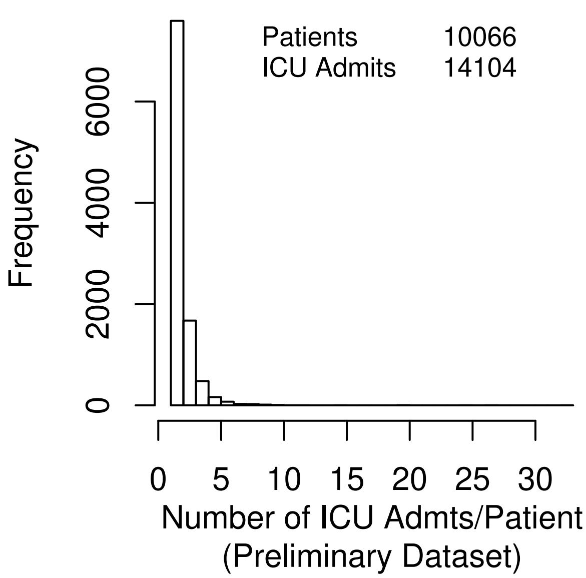


Figure 3-4: SAPS II Variables (cont)

Table 3.11: Hospital and ICU Admissions

Dataset	Patients	Hosp Admts (avg/pt)	ICU Admts (avg/pt)
`CensusEvents`	25,642	29,602 (1.15)	33,492 (1.31)
Preliminary	13,923	17,499 (1.26)	19,594 (1.41)
Final	10,066	12,693 (1.26)	14,104 (1.41)



Dataset	Patients	Readmitted (%)
`CensusEvents`	25,642	4,477 (17.5%)
Preliminary	13,923	3,462 (24.9%)
Final	10,066	2,474 (24.6%)

Figure 3-5: ICU Readmissions

a given patient. Some patients, for example, are recorded as having died over one year after their first ICU discharge. To limit such cases, one might define death as "died within the ICU or within 30 days of discharge". This limit marks a patient as alive if he or she is still in the hospital 30 days after ICU discharge. If the patient is not in the hospital at this point, the hospital discharge status of the patient is used to indicate mortality: if the patient was discharged alive (censored) they are marked as survived; otherwise they are marked as expired. Figure 3-6 illustrates the change in mortality rate as patients stay in the ICU for longer periods of time. This figure includes both hospital mortality (i.e., the patient died at any point during any recorded visit) and the within-30-days-of-ICU-discharge mortality. As the figure shows, the two mortality rates track each other closely for the first several days. However, it is clear that patients who stay in the ICU longer are more apt to remain in the hospital for a significant period of time before dying. For the remainder of this work, references to "mortality" indicate death in the ICU or within the following 30 days.

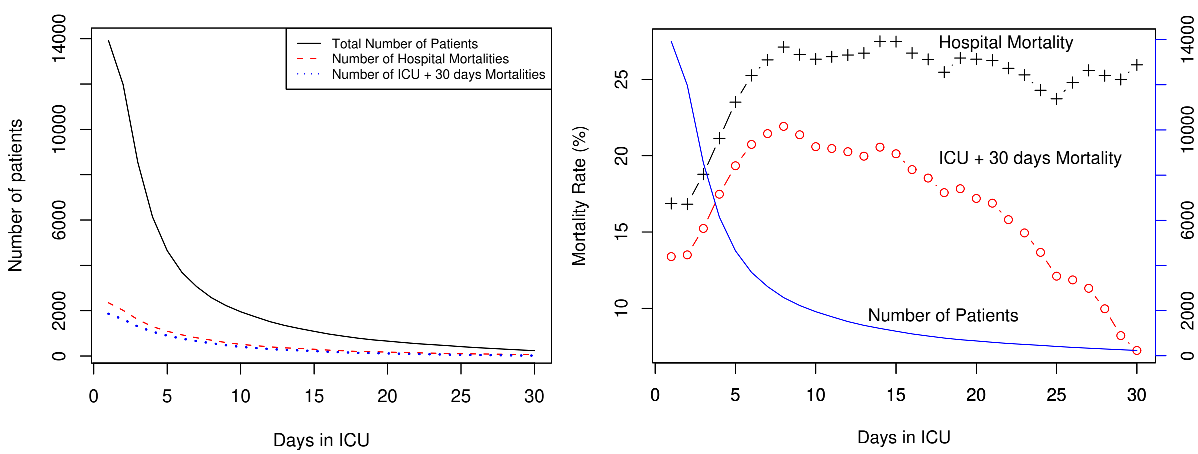


Figure 3-6: Patient counts versus the number of days spent in the ICU (left) and mortality rate versus the number of days spent in the ICU (right). For each patient, only the first ICU stay of the first recorded hospital visit is considered. "ICU + 30 day mortality" excludes deaths that occur after long post-ICU discharge hospitalizations. If a patient leaves the hospital alive within this 30-day period, they are assumed to have survived.

Table 3.12: Outcome Variables

Variable Name	Description and Source
Censored	The last `Expire_Flg` in the `Admissions` table indicates that patient left the hospital alive
Died	The patient dies in the ICU or dies within 30 days of discharge to the hospital floor. Censoring within this period is assumed to be equivalent to survival

3.5 Final Dataset

3.5.1 Patient Selection

For the preparation of my preliminary dataset, every effort was made to include all reasonably complete patients. For the final dataset — to be used for training and validating my models — I added a handful of important limitations to the scope of the data. The range of ailments that warrants intensive care is quite large; the limitations that I imposed were helpful in focusing my modeling efforts. In general, patients were excluded entirely if they were thought to have a different set of risks and concerns than the majority of the cases. The criteria I used for dropping entire patients (many of which were redundant) are listed in Table 3.13. The rules in Table 3.13 generally flag patients with severe trauma or neurological problems. While these patients require the close monitoring of an ICU, the root insult to their body is quite different from a medical patient or a heart patient and they often warrant different interpretations of physiological responses than other patients.

Table 3.13: Final Dataset: Entire Patient Exclusions

Drop Rule	Number of Patients
Neurosurgery patients (NSICU Service)	1987
Trauma patients (CSICU Service)	1676
Chronic Renal Failure (An ICD-9 code of 585)	225
Discharge summary contains “brain death”	56
Discharge summary contains “comatose”	41
Discharge summary contains “brain dead”	38
Discharge summary contains “brain steam dead”	2

In addition to dropping entire patients, other cases arose where it was helpful to drop only portions of a patient’s stay. For example, periods of a patient’s stay where the patient received limited care (e.g., comfort measures only) should clearly be treated differently than cases where the caregivers were trying everything possible to help the patient survive. The rules that I used for discarding such periods of a patient’s stay are included in Table 3.14.

Table 3.14: Final Dataset: Partial Patient Exclusions

Drop Rule	Number of Rows
In the ICU for longer than seven days	728739
Received limited treatment, including	198942
CMO (“comfort measures only”)	
DNR (“do not resuscitate”)	
DNI (“do not intubate”)	
“no CPR” or “other code”	
Received hemodialysis or hemofiltration	139561

The motivation for these rules generally follows the reasoning for excluding entire patients. For example, as Figure 3-6 indicates by plotting the mortality rate versus the number of days spent in the ICU, most patients leave the ICU within seven days of admission. For patients that do not leave in this 7-day window, the 30-day mortality rate starts to noticeably decrease as caregivers are able to successfully prolong the patient’s life while the patient remains in a compromised state often dependent on various interventions.

3.5.2 Dataset Summary

The final dataset — after applying all of the exclusions mentioned above — is summarized in Table 3.15. In addition, Appendix B lists the 438 individual variables with brief summary statistics. Figure 3-7 provides an updated version of Figure 3-6 for the final dataset.

Table 3.15: Preprocessed Data

Number of Patients	10,066
Number of Rows	1,044,982
Number of Features	438

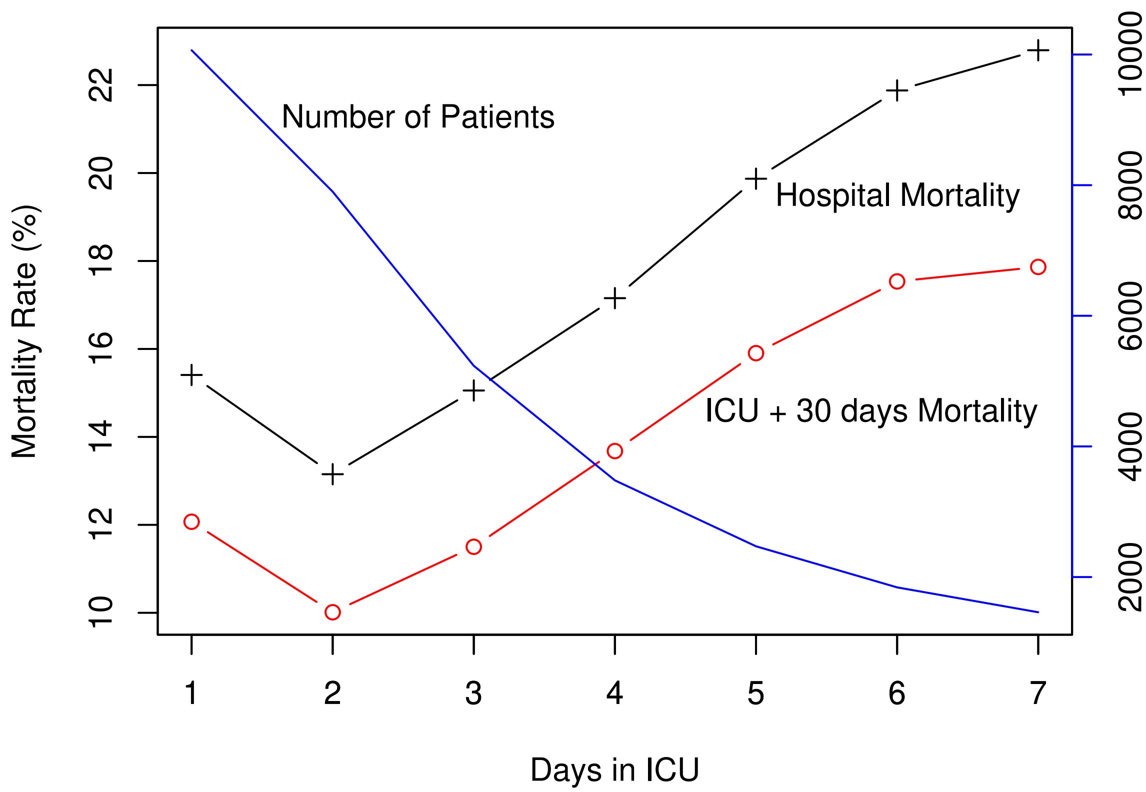


Figure 3-7: Mortality rate versus the number of days spent in the ICU: Final dataset

Chapter 4

Methods: Modeling

This chapter provides the framework that I used for creating predictive patient models. All of the models I discuss in this thesis rely on the methodology in this chapter. The discussion is kept general as it is necessary to defer details that depend on the specific types of models discussed in the following chapters. While the final dataset mentioned in the preceding chapter is the basis for all of the models that I consider, some of the models that I discuss later will place further limitations on this data and define a variety of outcomes to predict. For example, to predict the weaning of vasopressive medications, a model might be trained only on patients who are receiving vasopressive medications. Consequently, the methodology that follows is kept at a general level.

The patient models I developed are based on logistic regression. Logistic regression models the log odds ("logit") of a binary variable Y (e.g., mortality) using a linear combination of covariates (explanatory variables), $\boldsymbol{X}$:

$$\log\left(\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right)=\boldsymbol{X}\boldsymbol{\beta}.$$

This model can easily be rearranged to provide the probability of the outcome, $P(Y=1|\boldsymbol{X})$, as follows

$$P(Y=1|\boldsymbol{X})=\frac{1}{1+\exp\left(-\boldsymbol{X}\boldsymbol{\beta}\right)}.$$

The variable weights, $\boldsymbol{\beta}$, are typically fit using maximum likelihood.

There were three motivating factors for using logistic regression: (1) logistic regression results in transparent models that are easily understood and familiar to many within the medical profession; (2) logistic regression is a powerful modeling technique that can perform quite robustly at difficult prediction tasks; (3) the training—via iterative maximum likelihood estimation of the regression coefficients—is generally tractable. Given my large high-dimensional dataset, logistic regression allowed me to create interpretable, strongly performing models that were trained on all of the available training data.

In [31] I explored Cox Proportional Hazards survival models on an earlier release of the MIMIC II data. Survival models offered little benefit over logistic regression and the maximum likelihood estimation process for the Cox models often suffered from convergence problems. The primary limitation I encountered was quality followup data.[1] Even with followup data, some have argued that survival methods are inappropriate for the ICU because some patients—who ultimately die in the ICU—experience prolonged survival in the ICU that does not benefit them [77].

4.1 Model Construction

For my modeling, I started with a number of important assumptions. First, most of the models that I created assume that the covariates are stationary. This is not entirely accurate, as the ICU population is expected to change as some patients are discharged and others remain in the unit over time. Some individual variables, such as BUN, do demonstrate small trends over time in the unit. Work by Kayaalp et al., however, found that stationary ICU models often perform better than non-stationary models [35]. One way to avoid the stationarity assumption is to build daily models based on daily aggregates of the covariates. This is the approach that SAPS II takes by looking at representative (typically *worst*) values from the first 24 hours. A second important assumption, which is fundamental in logistic regression, is that the observations are independent of each other and linearly related to the logit of the dependent variable (i.e., outcome Y). The methodology that I employed largely follows the multivariate logistic regression methodology suggested by Ruttimann in [73].

4.1.1 Development and Validation Splits

To facilitate independent training and validation, I randomly partitioned the dataset into 70% development patients and a 30% validation patients. My models were trained exclusively using the development set. The validation set was used only to validate my final models on previously unseen patients.

4.1.2 Model Selection

The concept of the "best model" has garnered considerable attention in statistical modeling. The trade-off between model complexity and goodness of fit is typically a significant concern. Furthermore, with 438 different variables, a significant challenge

[1] The latest release of MIMIC II contains slightly better followup data by supplying a unique subject identifier to track the same patient across multiple hospital visits. For patients who leave the hospital alive on their last recorded visit, however, no followup is known. Work is underway to establish followup status by using the social security death records.

exists in finding the best subset of variables to include in a predictive model. An exhaustive search over the space of all possible models is combinatorially infeasible, so a number of commonly used (but not necessarily optimal) strategies are employed to simplify the search problem. I first filtered the candidate variables, then examined univariate models and finally performed backwards elimination to arrive at a final model. Each of these steps is further explained below.

Variable Filtering

Given the inclusive nature of my dataset preparation, a number of the dataset variables were included that had limited availability or were irrelevant. I eliminated a variety of such covariates by applying simple filters. The filters removed three categories of covariates from consideration. The first two filters addressed the problem of missing data: (1) I excluded covariates that were available for less than 80% of the development patients; (2) I removed covariates that had an average per-patient availability of less than than 60% of the patient instances (e.g., tests that were only executed on the first day for a typical patient and unavailable for subsequent days). The third filter removed irrelevant variables: (3) variables that remained effectively constant across the development patients were dropped.

Univariate Analysis

Given a binary outcome of interest, Y, the set of potential covariates was further reduced by selecting the most significant individual covariates. After ranking the covariates based on significance, I used a fixed significance threshold (e.g., p=0.05) to keep the top covariates for inclusion in my initial multivariate model. My ranking was based on the Wald Z score of each covariate obtained from a univariate logistic regression model trained to predict Y.[2] The Wald statistic, Z, is defined as the coefficient estimate for the univariate model, $\hat{\beta}$, divided by the estimated standard error of $\hat{\beta}$,

$$Z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}.$$

In addition to their original form, dataset covariates were evaluated for a variety of functional forms by applying transformations. The best form, in terms of the Wald statistic, was used for each variable. If the different transformations were nearly identical to the original form, then the original form was preferred. The transformations considered for each covariate included the following:

- Inverse (i)
- Absolute value (abs)

[2] p-values are easily obtained by comparing the squared Wald Z statistic against the χ^2 distribution with one degree of freedom

- Value squared (sq)
- Square root of value (sqrt)
- Logarithm of absolute value (la)
- Absolute deviation from mean (derangement) (am)
- Logarithm of absolute deviation from mean (lam)

While most values in the dataset were greater than or equal to 0, the absolute values in the above list were used for the few variables, such as Arterial Base Excess, that do drop below 0. To prevent logarithms of zero, values that were transformed with the logarithm were first shifted by adding a value of 0.0001.

The choice of the specific p-value threshold used for univariate screening warrants additional discussion. Many researchers have suggested using a rather liberal p-value such as 0.25 while others have been more conservative with lower p-values such as 0.05 [29]. The more stringent p-value thresholds avoid covariates of questionable importance, while a more liberal threshold admits covariates that may become important when considered along with other covariates. In general, the amount of data used for this research yields small p-values and most of the variables are significant at the 0.05 level.[3]

Collinearity Analysis

Using the top covariates (in their best form), I next screened the covariates to identify collinear or highly correlated covariates. This was done by first keeping only the best variable (based on univariate ranking) from variables that were clearly correlated—such as number of critical systolic blood pressure events over slightly different window lengths. After this first pass was completed, Spearman's rank correlation, ρ, was used to create a large correlation matrix. Spearman's rank correlation coefficient is a nonparametric measure of correlation that will detect monotonic relationships. Starting with the most significant univariate variables, correlation coefficients with other variables were examined. If a variable with less importance had a ρ value greater than 0.8, it was discarded.

With the variables that remained after filtering, univariate ranking, and collinearity analysis, an initial multivariate model was fit to the data. First, however, the model fitting process typically required manual removal of variables that caused singularity problems. While the collinearity analysis removed strong pairwise correlations, in the context of several hundred covariates other more subtle correlations arose that prevented the $\boldsymbol{\beta}$ estimation process from converging. With these considerations,

[3]For example, using the development split of the final dataset described in the previous chapter and logistic regression on mortality, a p-value threshold of 0.05 only eliminates around 15 of the 438 possible covariates

an initial multivariate model—typically containing several hundred variables—was trained.

Backward Elimination

With an initial model, backward elimination was next performed to simplify the model and remove variables (covariates) with marginal contribution. Backward elimination simplifies a large model by greedily removing the weakest variables. I used Akaike's Information Criterion (AIC) to eliminate the weakest features. The AIC metric penalizes the log likelihood of a candidate model by subtracting the number of parameters that were estimated for the model. An alternative to AIC is the Bayesian Information Criterion (BIC). BIC places more emphasis on model parsimony by multiplying the AIC complexity penalty by $\frac{1}{2}\log(n)$, where n is the sample size used to train the model [24]. Backward elimination proceeded by iteratively eliminating the least significant variable until removing the least significant variable caused the AIC value of the model to surpass the typical AIC threshold of 0. When the AIC threshold of 0 was reached, no more variables were removed and the model fitted with the selected set of variables was retained.

Sensitivity Analysis

By progressively increasing the AIC threshold from 0, I evaluated the sensitivity of the model to the number of covariates that it included. A plot of model performance versus the number of covariates provided a reasonable estimate of asymptotic upper bound on performance and the fewest number of covariates necessary to offer strong performance. In the course of the sensitivity analysis, if a simpler model was found that performed comparably to the more complex model, the complex model was discarded in favor of the simpler model.

4.1.3 Final Model

The model construction process above was repeated 5 times on randomly selected, unique 80%-training and 20%-validation partitions of the development data (5-fold cross-validation). This provided a check against over-fitting the development data. From the 5 models created, the union of the top features from each model was used to form a new model on all of the development data. Backward elimination on this new model was performed one last time.

As a final step, manual refinement of the model with human expertise was often helpful to simplify the model. For example, in some cases the backward elimination procedure would result in multiple variables that measure similar phenomenon. This might happen when several I/O variables (e.g., `allinput`, `Bal24` and `LOSBal`) were highly significant in the model, even when dropping one of the three may have a negligible effect on the model's performance. In other cases, however, seemingly similar variables were used together by the model to get at a different measurement. For example, by including the 24-hour fluid balance and the fluid balance for the entire stay, the model might better understand recent changes in the fluid balance. The effect in model performance was carefully considered before manually removing variables.

I also examined the availability of the covariates that were automatically selected for the models. Preference was given to covariates with high availability. As an example, variables derived from the MIMIC II `TotalBalEvents` table were often unavailable for the first ICU day while similar variables that were manually integrated from the MIMIC II `IOEvents` table were available for the same day. When model performance was similar, I used the more frequently available inputs. In other cases, marginally important variables that were frequently missing were manually examined and removed if the change in model performance was negligible. Frequently missing variables that resulted in significant performance improvement can be easily identified in plots that show the AUC performance versus the number of covariates: large jumps in these smooth curves generally represent an increase in missing observations. Large jumps were rare but when present influenced the choice of covariates for the final model.

Before validating against the held-out validation data, the model resulting from the above model selection process was examined to assess its fit on the development data. The model assessment was done using ROC curves and the Hosmer-Lemeshow test. In addition, bootstrapping (150 samples with replacement) was used to validate a number of goodness of fit statistics such as the logistic calibration curve slope and intercept (predicted probability versus actual probability). Each of these performance metrics is discussed further in the following section. Models that performed well were considered final and further validated on the held-out test data.

4.2 Model Validation

After the final model was found using the approach outlined above, I used the held-out data in order to validate the model's performance on unseen data. As in the model construction, I relied on two primary metrics to validate model performance. First, I looked at the Receiver Operating Characteristics (ROC) curve. As a second metric, I examined the calibration of the fitted model using the Hosmer-Lemeshow goodness of fit test and calibration plots.

4.2.1 Discrimination

To evaluate the model's discriminatory ability, I looked at ROC curves. ROC curves graphically illustrate the performance of a classifier by plotting the sensitivity versus specificity for different thresholds on the classifier's output. By looking at the area under the ROC curve (AUC), one can summarize the general performance of the classifier.[4] The curve allows one to evaluate the performance of the model when, for example, a given specificity is required. The ROC curve can be thought of as evaluating the adequacy of *risk ranking* for the classifier. The discrimination ability of two classifiers on the same subjects can be compared using a test suggested by DeLong [42]. The test examines the difference between the ROC areas and the variance of the difference. Since the two curves are partly correlated, the test must also correct for the correlation between the curves.

In many cases, it can be helpful to examine the positive predictive value (PPV) and the negative predictive value (NPV) for a classifier. The PPV is defined as the number of true positives at a given threshold for the model divided by the total number of positive predictions (true positives and false positives) from the model. Similarly, the NPV indicates the number of true negatives predicted by the model divided by the total number of negatives predicted by the model. In contrast to the AUC from an ROC curve, the PPV and NPV depend on the classification threshold chosen and the prevalence of the outcome of interest.

4.2.2 Calibration

In addition to discrimination performance, it is important for models to demonstrate strong calibration. Calibration can be viewed as evaluating the adequacy of the individual *risk estimates*. Several calibration tests exist, such as the common Hosmer-Lemeshow test [49, 29].

The Hosmer-Lemeshow test compares the observed frequencies and the estimated expected frequencies for a set of risk groups. The number of groups, g, is typically 10 and are often referred to as "deciles of risk", indicating that the highest 10% of the predictions will be in one group, the next 10% will make up the next group, and so on. The calibration statistic based on these groups will be referred to as H. In addition to grouping the outputs based on risk, it is also common to group them based on fixed cut-points over the output range (i.e., probability deciles). For example, the with $g = 10$, the probability deciles would be [0.1, 0.2], (0.2, 0.3], and so on. This version of the Hosmer-Lemeshow statistic will be referred to as C. In either case, the overall fit of the model is evaluated by comparing the test statistic to the Pearson χ^2 distribution with g degrees of freedom for the validation data or a Pearson χ^2 distribution with $g - 2$ degrees of freedom for the development data. Statistically significant p-values for the Hosmer-Lemeshow test indicate poor model calibration.

Many fault the grouping choice for the Hosmer-Lemeshow as arbitrary. An alternative method to evaluate calibration, proposed by Cox [11, 61], examines the relationship between the predicted probability and the actual probability. Instead of looking within local subgroups of the predictions, this method makes a more global assessment of model fit. The actual *calibrated* probability, P_c, is found by fitting the

[4]The AUC is equivalent to the Mann-Whitney statistic.

logit of the probability estimates, $L = \text{logit}(\hat{P})$, against the actual outcome, Y:

$$P_c = P(Y = 1|\boldsymbol{X}\hat{\beta}) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 L))}.$$

In a correctly calibrated model, $\gamma_0 = 0$ and $\gamma_1 = 1$. Since the development data will "fit" the data globally, it is necessary to use bootstrapping or cross-validation to obtain bias-corrected estimates of γ_0 and γ_1. Using $\hat{P}$ and $\hat{P}_c$ (bias-corrected), one index of unreliability is E_{max}, defined as

$$E_{max} = \max |\hat{P} - \hat{P}_c|.$$

The corrected estimates of γ_0 and γ_1 provide useful insight into the level of overfitting present in the model: (1) the bias-corrected γ_0 (intercept) is positive when the predicted probabilities are, on average, too high, and negative when the predicted probabilities are, on average, too low; (2) the bias-corrected γ_1 (slope) is greater than 1 when the predicted probabilities are, on average, too close to the mean, and less than 1 when the predicted probabilities are, on average, too extreme.

When validating against the development data, I used bootstrapping with 150 samples (with replacement). A number of statistics were calculated for each sample [24]. Of these statistics, I focused on three in particular: (1) the *Intercept*, γ_0, for the fitted logistic calibration curve, (2) the *Slope*, γ_1, for the fitted logistic calibration curve, and (3) the maximum absolute difference in predicted and calibrated probabilities $E_{\max}$. The intercept, slope, and error allow one to quantify the correction needed to calibrate the model predictions. Additional statistics that measure overfitting are also calculated on the bootstrap samples and are described in Appendix A.

When validating against the held-out data, the validation data was used to make one estimate of the above statistics (instead of the 150 estimates created by bootstrapping). To plot P_c versus $\hat{P}$ on validation data, I used Harrell's *val.prob* function included in the R Design package [66, 25]. The *val.prob* function shows the ideal calibration line (slope of 1 and intercept of 0), along with the fitted logistic calibration curve and a smooth nonparametric calibration curve fit using *lowess* smoothing. The relative frequency distribution of the probabilities (divided into 101 bins from 0 to 1) is shown along the x-axis. In addition, the mean probabilities for the deciles of risk used for the Hosmer-Lemeshow H statistic were added to the plots.

4.3 Other Severity of Illness Scores

For ICU models that predict patient mortality, it is useful to compare performance against other ICU severity of illness scores. While a number of scores exist, I chose to compare against SAPS II [44]. SAPS II was developed on a large set of ICU patients from Europe and North America and it is quite common, especially in European countries. It is also simpler than the more complex Apache score. Despite its simplicity, some of the variables were not directly available from the MIMIC II data. The SAPS I score does not suffer from the same data difficulties, but the added refinement of SAPS II (e.g., finer tuned granularity for the individual contributions), along with the wealth of literature looking at its application to various patient populations, make it considerably more attractive than SAPS I. This section summarizes the approach I took to calculate SAPS II.

4.3.1 SAPS II Calculation

For my comparisons, SAPS II scores were calculated for the same set of patients that I used to construct and validate a given model. In calculating SAPS II, I followed the description provided by Le Gall et al. in [44]. The specific variables that are used in SAPS II are listed in Table 4.1. Two fields were not directly available in the MIMIC II data: "chronic diseases" (Metastatic cancer, Hematologic malignancy, AIDS) and "type of admission" (Scheduled surgical, Medical, Unscheduled surgical).

I was able to consistently identify chronic illnesses by searching through the patient ICD-9 codes. The "type of admission" variable, however, was difficult to automatically determine. The possible values for this variable were defined by Le Gall et al. as follows:

- Scheduled surgery: patients who were scheduled for surgery at least 24 hours in advance

- Unscheduled surgery: patients who were scheduled for surgery within 24 hours of the operation.

- Medical admissions: patients who had no surgery within 1 week of admission to the intensive care.

Often, even with manual review of the available progress/discharge notes, extracting type of admission information was difficult.

For SAPS II calculations, most variables are considered normal (i.e., do not increase the score) if they are unavailable. Two variables, however, were required by Le Gall et al. for a patient to be included in their analysis. These included the "type of admission" variable (discussed above) and, for patients who are ventilated or have continuous positive airway pressure (CPAP), the PaO_2/FiO_2 ratio. For MIMIC II patients, it was typically not a problem to calculate the PaO_2/FiO_2 ratio for ventilated or CPAP patients. If the PaO_2/FiO_2 ratio was unavailable, the patient was excluded. The type of admission variable, however, presented a greater challenge. I attempted to resolve this challenge with similar variables that were found important for stratifying risk. First, without the information to distinguish between types of admission, the score for type of admission was left at zero in my score calculations. To compensate for this omission, two additional variables indicating ICU service type (svCSRU and svMICU) were included in the logistic regression model that predicts mortality from SAPS II. These two indicator variables were found by my mortality models to be highly important at stratifying ICU patient risk and were also highly significant in conjunction with SAPS II for predicting mortality. The resulting "pseudo-SAPS II" offers less than an ideal comparison with published SAPS II results; with the inclusion of the ICU service type variables and omission of the type of admission to the SAPS II calculation, a small difference in SAPS II mortality prediction performance can be expected. The CSRU service, however, should have a strong correlation with elective (scheduled) surgery patients and, similarly, the MICU service should have a strong correlation with medical admissions.

More recent severity of illness metrics, such as APACHE III, have found that context variables, such as prior location or major disease category, are helpful in predicting patient outcomes [38]. While having a relatively small maximum contribution of 8 points (unscheduled surgery) Le Gall et al. specifically note that the "type of admission" information was important enough to be required for patients in their model development and validation. Consequently, the SAPS II calculation included in this work should be subject to additional scrutiny when compared with other metrics that utilize more complete patient context information. But for the purposes of this work, the pseudo-SAPS II mortality predictions should provide a useful relative comparison; the additional patient context that is unavailable for SAPS II is also unavailable to my models.

Given the above considerations, the final SAPS II score was found by summing the measures of derangement (points) for each individual SAPS II variable. The list of SAPS II variables, including the maximum point contribution for each, is shown in Table 4.1. Using the development patients with SAPS II scores, svCSRU indicators, and svMICU indicators, a logistic regression model was trained to predict mortality. An additional model, without the ICU service type variables, was also examined. To compare with the equation published in [44], I also included the logarithm of the SAPS II score, ln(SAPS II + 1), as a covariate in both logistic regression equations. The logit equation published by Le Gall et al. in [44] is:

$$\text{logit} = -7.7631 + 0.0737(\text{SAPS II score}) + 0.9971[\ln(\text{SAPS II score} + 1)].$$

Table 4.1: SAPS II Variables

Variable	Max Points
Age	18
Heart rate	11
Systolic BP	13
Body temperature	3
PaO2:FiO2 (if ventilated or continuous positive airway pressure)	11
Urinary output	11
Serum urea nitrogen level	10
WBC count	12
Serum potassium	3
Serum sodium level	5
Serum bicarbonate level	6
Bilirubin level	9
Glasgow Coma Score[a]	26
Chronic diseases	17
Type of admission	8

[a]If the patient is sedated, the estimated GCS prior to sedation

Chapter 5

Mortality Models

Many severity of illness scores have been built to provide standard ICU patient risk assessments. These models have generally focused on the first 24 hours of a patient's ICU stay. If applied later in a patient's stay, they are still performed on a daily basis. While useful for stratifying patient risk between hospitals, daily predictions do not allow the score to closely track a patient's ICU progression. Acute events, such as the onset of septic shock or decompensation, may occur between daily scores.

Furthermore, most existing models emphasize simplicity in their inputs. In many ways this is an obsolete requirement that stems from a time when the scores were calculated by hand. Simplicity was also considered important for portability between different hospitals. Today, in an era of digitally collected data, a computer should be able to help analyze complex data patterns and assist caregivers in continually assessing patient risk.

This chapter focuses on building models that predict patient mortality. I first review the data that were used for training and validation, and then I present daily mortality models and "real-time" mortality models. For comparison purposes, I also present the customized SAPS II model described in the Chapter 4. I end the chapter with a number of comparisons between the models presented and a discussion of my findings.

5.1 The Data

Two datasets were used for the mortality models that I develop in this chapter. One dataset used all observations from the final dataset described in Chapter 3. These data were used to develop and validate the "real-time" models. In the second dataset, the final dataset was aggregated by day. The aggregated data were used to develop and validate daily models. Demographic information for patients included in the final dataset is provided in Table 5.1.

The breakdown for the final dataset, after splitting it into development and validation sets, is described in Table 5.2. The number of variables for the development

Table 5.1: Final dataset description. Prior hospital time was calculated using the difference between the ICU admission time and the corresponding hospital admission time. ICU LOS: ICU length of stay.

Age (yrs)	Male %	Female %	Prior Hospital Time (days)	ICU LOS (days)
65 ± 16	58.6	40.9	1.8 ± 3.6	2.8 ± 2.1

CSRU %	CCU %	MICU %	MSICU %	CSICU %	NSICU %
31.7	23.2	22.3	10.9	0	0

Table 5.2: Real-time data

	Patients	Mortality	Rows	Variables
Final Dataset	10066	12.1% (1215)	1044982	438
Development Set	7048	12.1% (853)	736218	200
Validation Set	3018	12.0% (362)	308764	200

and validation sets represent those selected using the univariate analysis and filtering described in Chapter 4.

For the daily aggregate data, Table 5.3 describes the breakdown between the development and validation sets. In aggregating the data, the higher-frequency values were summarized using a number of summary functions. These functions included *min*, *max*, and *mean*. During univariate ranking of the resulting summary variables, only the best of the three summary variables were kept. In addition, the standard deviation of the variable over each day was independently included. As in the table for the real-time data, the number of variables for the development and validation sets reflects univariate analysis and filtering. The initial variable screening process was repeated independently (yielding slightly different results) for the aggregate models described in the following section. Table 5.3 indicates the number of variables used for the Stationary Daily Acuity Score (`SDAS`), which is the first model described in this chapter.

5.1.1 Outcomes

As noted in the data preparation, the definition of mortality that my models used is ICU death or death within 30 days following ICU discharge. If a patient was discharged from the hospital alive within 30 days of his or her ICU stay (i.e., censored), survival was assumed. There are, of course, a number of cases where this survival assumption could be problematic (e.g., discharge to hospice care) but such cases were

Table 5.3: Aggregated daily data. *Number of variables used for the `SDAS` model; the number of variables after univariate analysis and filtering for the `DAS`n models varied.

	Patients	Mortality	Rows	Variables
Final Dataset	10066	12.1% (1215)	32480	1752
Development Set	7048	12.1% (853)	22888	349*
Validation Set	3018	12.0% (362)	9592	349*

difficult to avoid given the current constraints of MIMIC II.

5.2 Daily Acuity Scores

I first examined mortality models based on daily patient data. Two types of daily models were explored. The first daily model was developed and validated on all of the daily data (up to 7 days). This model will be referred to as the Stationary Daily Acuity Score (`SDAS`), as it assumes that the data's joint probability distribution does not change between days. A second class of daily models, referred to as the Daily Acuity Score (`DAS`n), was developed for individual patient days $n \in \{1, 2, 3, 4, 5\}$. The training and validation for `SDAS` and `DAS`n are considered in this section. For each model type, I provide a brief overview of the learning process, followed by a description of the final model and performance on the development data. After I describe `SDAS` and `DAS`n, I examine their performance on the held-out validation data.

5.2.1 `SDAS` Model

The `SDAS` model considered observations from all ICU days. After the filtering, univariate ranking, and collinearity analysis described in chapter 4, the daily aggregate development data was reduced from 1752 to 349 variables. Using 5-fold cross-validation on the development data, five initial models were developed and examined. Figure 5-1 shows the performance of these five models on each validation fold (20% of development data) as the number of covariates increases. These figures were generated by backward elimination using a progressively increasing AIC threshold (starting at zero). In general, the validation performance closely tracks the training performance and little overfitting was observed.

The models developed on Fold 2 and Fold 3 appear particularly strong. The 30-covariate model from Fold 2 is shown in detail in Model 5.1. In the description of this model, the daily summary function (e.g., *max*) lies between the variable name and, if present, the type of transformation applied to the variable. For example, the variable `GCS_max_sq` should be interpreted as the maximum daily GCS value squared. To interpret the model, a positive coefficient should be understood as increasing the

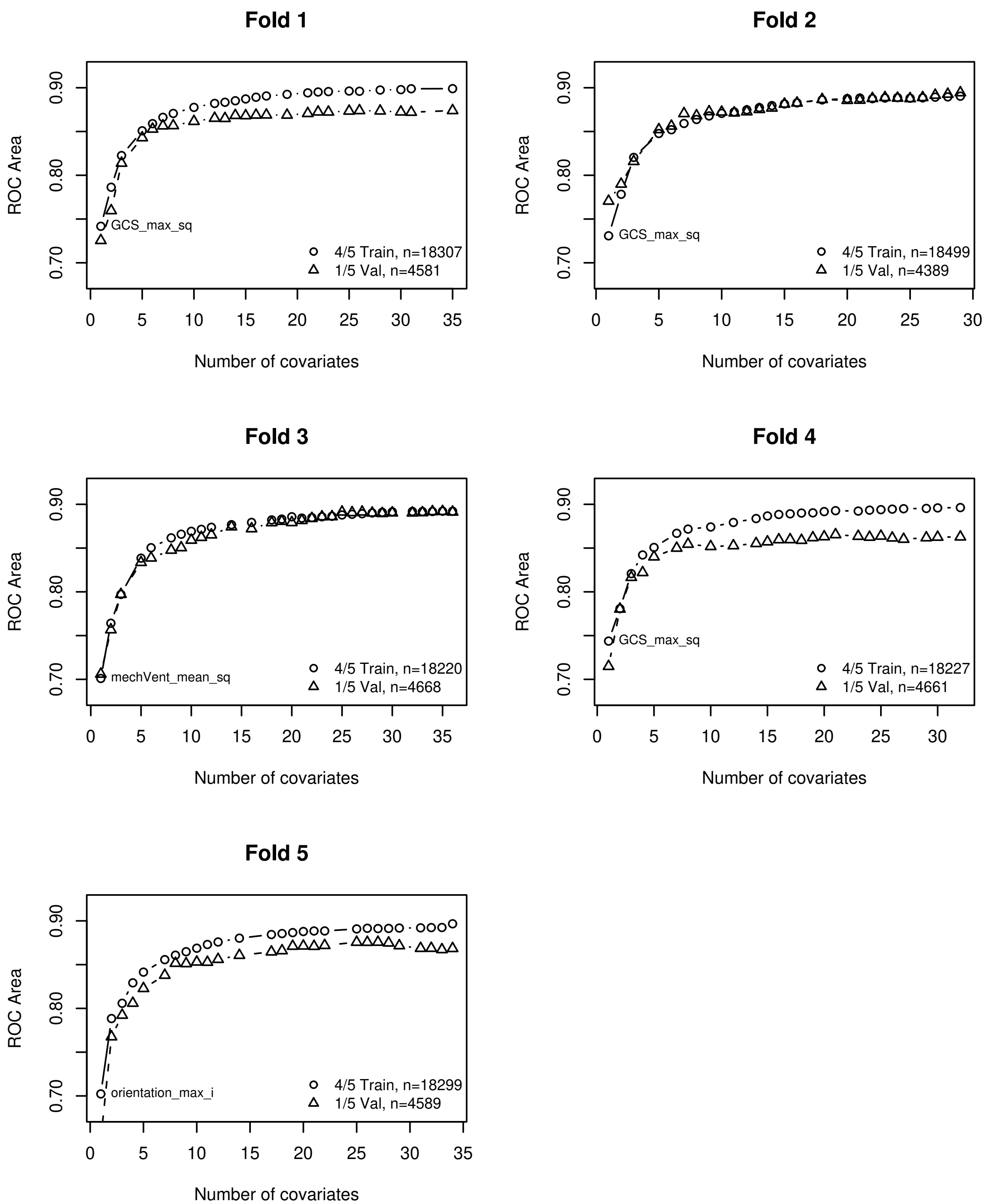


Figure 5-1: SDAS model selection. Sensitivity to number of covariates on each cross-validation fold. The covariate(s) from the simplest model are marked on the training curves.

probability of mortality (positive correlation) while a negative coefficient means less risk of mortality (negative correlation).

Model 5.1 includes a number of interesting covariates. By examining the Wald Z scores, however, it appears that the model can likely be improved. For example, the contributions from the largest length of stay fluid balance (`LOSBal_max`, $Z = 3.94$) and the largest 24 hour fluid balance (`Bal24_max`, $Z = -3.68$) may largely counteract each other and the model may benefit from removing at least one of these variables and possibly replacing it with an input variable (a mean hourly output for the day, `OutputB_60_mean_sqrt`, is already included). Similarly, the meaningfulness of the `pressD01_sd_sq` variable (that is, the squared standard deviation of the points marked 1 following the first pressor infusion and marked 0 before any pressors) is questionable as it decreases risk for patients who have a pressor started in the middle of their first day, but increases risk for patients who receive pressors early or late on their first day.

Examining all five folds from Figure 5-1, it is clear that there was negligible impact on the validation performance by reducing the number of covariates to about 25. Considering this, I took the top 25 covariates from the models for cross-validation folds 1, 2, 3, and 5 (excluding fold 4) and created a model using these covariates. By performing backward elimination one last time on this model a final model was selected. As done with each cross-validation fold in Figure 5-1, a plot of performance versus the number of covariates in a given model was created. This plot is shown in Figure 5-2.

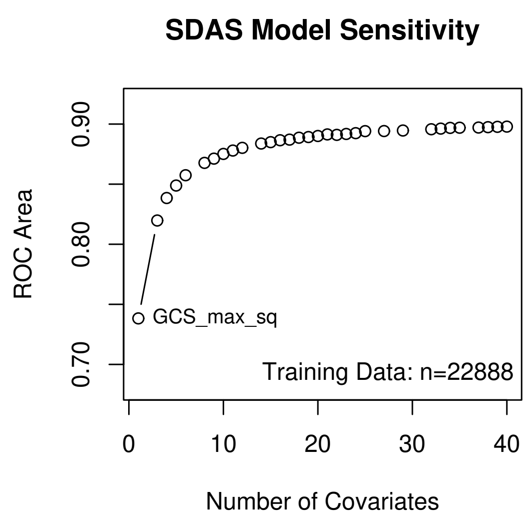


Figure 5-2: `SDAS` model selection (all development data)

The models in Figure 5-2 indicate that most of the performance was captured with about 35 inputs. This model was chosen for additional refinement. Upon examination, it was found to contain a number of pressor-related

Model 5.1 SDAS Model for Fold 2 with 30 Covariates

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
20172	1e-09	5415.11	30	0	0.893	0.785
Gamma	Tau-a	R2	Brier			
0.787	0.176	0.439	0.076			

	Coef	S.E.	Wald Z	P
INR_mean_i	-1.795e+00	1.423e-01	-12.61	0.0000
GCS_max_sq	-7.485e-03	6.000e-04	-12.47	0.0000
OutputB_60_mean_sqrt	-6.561e-02	6.885e-03	-9.53	0.0000
pacemkr_max	-1.084e+00	1.183e-01	-9.16	0.0000
svCSRU_max	-9.516e-01	1.208e-01	-7.88	0.0000
GCSrdv_mean	-1.138e-01	1.528e-02	-7.45	0.0000
pressD01_mean_am	-2.774e+00	3.893e-01	-7.13	0.0000
Platelets_Slope_1680_min	-5.493e+00	8.615e-01	-6.38	0.0000
pressD01_sd_sq	-5.085e+00	8.678e-01	-5.86	0.0000
sedatives_mean_sq	-4.375e-01	8.455e-02	-5.17	0.0000
Bal24_max	-4.493e-05	1.222e-05	-3.68	0.0002
CV_HRrng_max	-3.267e-03	1.083e-03	-3.02	0.0026
Intercept	4.292e-01	4.085e-01	1.05	0.2934
Milrinone_perKg_min_sq	3.523e+00	1.113e+00	3.17	0.0015
LOSBal_max	2.247e-05	5.703e-06	3.94	0.0001
hrmVA_max	3.410e-01	6.767e-02	5.04	0.0000
MBPm.pr_min_am	1.904e+00	3.711e-01	5.13	0.0000
Mg_min_sq	1.067e-01	1.798e-02	5.93	0.0000
beta.Blocking_agent_mean_lam	2.418e-01	3.955e-02	6.11	0.0000
Na_mean_am	5.214e-02	8.415e-03	6.20	0.0000
mechVent_mean_sq	7.183e-01	1.047e-01	6.86	0.0000
RESP_mean_sq	9.226e-04	1.293e-04	7.13	0.0000
Platelets_mean_i	2.512e+01	3.512e+00	7.15	0.0000
Lasix_max_lam	2.550e-01	3.457e-02	7.38	0.0000
CO2_mean_i	2.038e+01	2.741e+00	7.43	0.0000
jaundiceSkin_mean_la	1.523e-01	2.014e-02	7.56	0.0000
hospTime_min_sqrt	6.860e-03	7.939e-04	8.64	0.0000
pressorSum.std_mean_sqrt	7.758e-01	7.225e-02	10.74	0.0000
SpO2.oor30.t_mean_sqrt	4.929e-01	4.095e-02	12.04	0.0000
BUNtoCr_min_sqrt	2.867e-01	2.323e-02	12.34	0.0000
Age_min_sq	2.258e-04	1.450e-05	15.57	0.0000

variables: (1) `Sympathomimetic_agent_min`, (2) `pressorSum.std_mean_sqrt`, (3) `Dopamine_perKg_mean_sqrt`, (4) `pressD01_mean_am`, and (5) `pressD01_sd_sq`. Since the influence of the `Dopamine_perKg_mean_sqrt` and `pressorSum.std_mean_sqrt` were the smallest, they were removed from the model with little consequence. The `pressD01_sd_sq` input was also removed with no appreciable change in model performance. Similarly, the `sedatives_mean_sq` variable was removed in favor of only the `RikerSAS_mean` variable, and the automatically selected `TotOut24_min_sqrt` variable (from the `TotalBalEvents` table) was replaced with the similar, and more frequently available, `Alloutput_max_la` variable (manually integrated from the `IOEvents` table).[1] These changes ultimately had little impact on the fit to the training data but were felt to help simplify the model.

Additionally, comparison of the covariates in the model with inputs to other severity scores revealed several noticeable omissions — namely WBC Count, AIDS, Metastatic Carcinoma, and Hematologic Malignancy. I manually adding `WBC`, `AIDS`, `MetCarcinoma`, and `HemMalig` to the model. Of these, each had a small, albeit significant ($p < 0.005$), contribution to the model. Hematologic Malignancy had the most influence (Wald Z score of 4.98). The other three additions were significant but had Wald Z score less than the other contributions (Wald $Z < 3$). Given the clear role that these variables have in increasing the risk of mortality, they were included in the final model with with minimal concern for overfitting. The final model is described in detail in Model 5.2.

`SDAS` Validation (Development Data)

Next, I examined the goodness of fit for Model 5.2 on the development data. Figure 5-3 shows the ROC curve for the model on the development data. With an area of 0.898, the model does quite well at discriminating between patients who survive and patients who expire. It should be noted that Figure 5-3 contains multiple predictions for most patients. If the predictions are limited to day 1 (only one prediction per patient), the AUC decreases slightly to 0.890.

The calibration of `SDAS` on development data was also considered. Tables 5.4 and 5.5 provide details on the Hosmer-Lemeshow statistic for the deciles of risk (H) and the fixed probability deciles (C). While the model performs reasonably well in general, several of the deciles have large deviations between the number of deaths that were predicted and the number of deaths that were observed. With the large number of observations, these differences are statistically significant as shown by the χ^2 values below each table. If the validation is limited to the first ICU day, the calibration improves somewhat with p-values of 0.017 and 0.116 for the deciles of risk (H) and deciles of probability (C), respectively. In contrast, the calibration is better for the subsequent days with p-values ≥ 0.1 for days 2 through 5 (using H).

[1] The smoothness of the curve in Figure 5-2 (i.e., no large drops) indicates that the models were generally insensitive to covariates with a large number of missing values

Model 5.2 Final SDAS model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
20130	3e-10	5619.28	35	0	0.898	0.797
Gamma	Tau-a	R2	Brier			
0.798	0.177	0.456	0.074			

	Coef	S.E.	Wald Z	P
GCS_max_sq	-0.0064668	5.032e-04	-12.85	0.0000
INR_mean_i	-1.8734049	1.458e-01	-12.85	0.0000
pacemkr_max	-0.9337190	1.179e-01	-7.92	0.0000
svCSRU_max	-0.9137522	1.250e-01	-7.31	0.0000
RikerSAS_mean	-0.3430971	5.151e-02	-6.66	0.0000
Platelets_Slope_1680_min	-5.8856843	8.839e-01	-6.66	0.0000
urineByHr_mean_sqrt	-0.0584113	9.453e-03	-6.18	0.0000
GCSrdv_mean	-0.0902717	1.552e-02	-5.82	0.0000
GCSrng_min_am	-0.0812232	1.459e-02	-5.57	0.0000
pressD01_mean_am	-1.6132643	3.005e-01	-5.37	0.0000
CV_HRrng_max	-0.0061979	1.216e-03	-5.10	0.0000
Insulin_sd_sq	-2.1686950	4.372e-01	-4.96	0.0000
alloutput_max_la	-0.0890330	2.265e-02	-3.93	0.0001
MetCarcinoma_min	0.4468763	1.567e-01	2.85	0.0043
WBC_mean_am	0.0147036	5.149e-03	2.86	0.0043
AIDS_min	0.5954305	1.991e-01	2.99	0.0028
Intercept	1.5314512	4.529e-01	3.38	0.0007
MBPm.pr_min_am	1.4601630	3.518e-01	4.15	0.0000
HemMalig_min	0.6032027	1.212e-01	4.98	0.0000
RESP_mean_sq	0.0006615	1.324e-04	5.00	0.0000
hrmVA_max	0.3520834	6.823e-02	5.16	0.0000
PaO2toFiO2_mean	0.2672376	4.336e-02	6.16	0.0000
Na_mean_am	0.0549066	8.506e-03	6.45	0.0000
Mg_min_sq	0.1173220	1.815e-02	6.46	0.0000
ShockIdx_max	0.5742182	8.853e-02	6.49	0.0000
Platelets_mean_i	24.0719462	3.560e+00	6.76	0.0000
hospTime_min_sqrt	0.0057514	8.158e-04	7.05	0.0000
day_min_sq	0.0170075	2.372e-03	7.17	0.0000
jaundiceSkin_mean_la	0.1469141	2.045e-02	7.18	0.0000
CO2_mean_i	19.3845272	2.682e+00	7.23	0.0000
Lasix_max_lam	0.2523702	3.444e-02	7.33	0.0000
beta.Blocking_agent_mean_lam	0.2918077	3.923e-02	7.44	0.0000
Sympathomimetic_agent_min	0.8576883	9.254e-02	9.27	0.0000
SpO2.oor30.t_mean_sqrt	0.4059329	4.128e-02	9.83	0.0000
BUNtoCr_min_sqrt	0.2829088	2.348e-02	12.05	0.0000
Age_min_sq	0.0002601	1.495e-05	17.40	0.0000

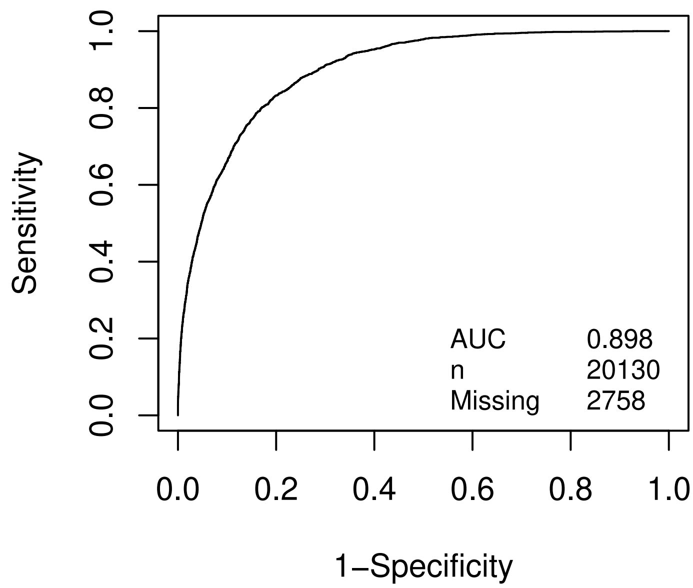


Figure 5-3: `SDAS` ROC curve (development data). AUC = the area under the curve; n = the total number of available predictions used for curve; Missing = number of missing predictions.

Table 5.4: `SDAS` Hosmer-Lemeshow H risk deciles (all days)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[0.000203,0.00335)	0.002	2	4.2	2011	2008.8	2013
2	[0.003353,0.00682)	0.005	3	9.9	2010	2003.1	2013
3	[0.006825,0.01281)	0.010	11	19.2	2002	1993.8	2013
4	[0.012812,0.02277)	0.017	24	34.8	1989	1978.2	2013
5	[0.022771,0.03971)	0.031	53	61.5	1960	1951.5	2013
6	[0.039706,0.06691)	0.052	104	104.8	1909	1908.2	2013
7	[0.066911,0.11297)	0.088	198	176.7	1815	1836.3	2013
8	[0.112972,0.20128)	0.152	324	305.3	1689	1707.7	2013
9	[0.201280,0.40232)	0.285	610	574.7	1403	1438.3	2013
10	[0.402321,0.99876]	0.634	1239	1276.9	774	736.1	2013

$\chi^2 = 24.47,\ d.f. = 8;\ p = 0.002$

Table 5.5: SDAS Hosmer-Lemeshow C probability deciles (all days)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.026	334	359.5	13270	13244.5	13604
2	(0.1,0.2]	0.142	376	353.1	2105	2127.9	2481
3	(0.2,0.3]	0.246	339	307.2	912	943.8	1251
4	(0.3,0.4]	0.346	274	266.1	494	501.9	768
5	(0.4,0.5]	0.447	275	272.9	335	337.1	610
6	(0.5,0.6]	0.550	215	221.5	188	181.5	403
7	(0.6,0.7]	0.651	204	216.7	129	116.3	333
8	(0.7,0.8]	0.749	196	193.9	63	65.1	259
9	(0.8,0.9]	0.849	179	191.0	46	34.0	225
10	(0.9,1]	0.950	176	186.2	20	9.8	196

$\chi^2 = 27.08,\ d.f. = 8;\ p = 0.001$

Table 5.6: SDAS bootstrapped goodness of fit statistics (development data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.797	0.799	0.795	0.004	0.793	150
R^2	0.456	0.460	0.454	0.006	0.450	150
Intercept	0.000	0.000	-0.017	0.017	-0.017	150
Slope	1.000	1.000	0.986	0.014	0.986	150
E_{max}	0.000	0.000	0.006	0.006	0.006	150
D	0.279	0.282	0.277	0.004	0.275	150
U	0.000	0.000	0.000	0.000	0.000	150
Q	0.279	0.282	0.277	0.004	0.275	150
B	0.074	0.073	0.074	-0.001	0.075	150

Finally, to further validate the model, bootstrapping with 150 samples (with replacement) was performed to assess the model's fit on the development data. Table 5.6 summarizes the result of this validation procedure. The columns of this table indicate the performance on the entire development data ("Original Index"), the performance on the bootstrapped training sample ("Training Sample"), the performance on the bootstrapped test sample ("Test Sample"), the difference between the two bootstrapped performances ("Optimism"), and the corrected statistic ("Corrected Index"). The level of optimism for individual statistics helps quantify the amount overfitting in the model. The statistics listed in Table 5.6 are briefly described in Appendix A.

5.2.2 DASn Model

My second type of model, based on daily aggregate data, specifies a separate model for each ICU day. These models allow us to explore the daily stationarity assumption of the SDAS model. As noted earlier, these models are referred to as DASn, where n indicates the day for which a particular model was trained. I followed a similar procedure on the DASn models as I did on the SDAS model above. However, due to the decreasing amount of data for each subsequent ICU day, I made a slight change in the feature selection methods. First, I performed the feature selection for the first three days as described using 5-fold cross-validation. However, with a similar number of candidate covariates and a much smaller set of observations (limited to one day), the greedy backward selection process often inadvertently deleted important covariates too early. To mitigate this problem, I performed a second stage of cross-validation using a limited set of candidate variables. This set included candidate variables selected from the 5-fold cross-validation on days 1, 2, and 3, and variables included in the final SDAS model (Model 5.2). This restricted the number of possible variables to a total of 83 for each DASn model (60 variables from days 1 through 3, 35 variables from SDAS, and an overlap of 12). As a result of the limitation on candidate variables, DAS4 and DAS5 had a slight disadvantage which was felt to be unavoidable given their limited data.

Appendix C shows the initial cross-validation performance of the models trained on each of the five folds for each of the first three days. Appendix C also contains sensitivity plots (performance as a function of number of covariates) for the second stage cross-validation that was limited to 83 variables.

As noted, since most patients leave the ICU within a few days, the number of patients available on later days was limited. Consequently, the Wald Z scores for covariates decreased as the ICU day increased. This led to more parsimonious models for later days. Consequently, only the first five days were explored because the decreasing number of patients after five days made modeling difficult and comparisons less meaningful.

Plots showing the sensitivity of model performance to the number of covariates are given below in Figure 5-4 for each of the five daily models (days 1 through 5). Models 5.3, 5.4,5.5,5.6, and 5.7 describe the final models. As in the SDAS model, WBC Count, AIDS, Metastatic Carcinoma, and Hematologic Malignancy were manually added to the model (if not already present). These variables were not always helpful and were only included if their individual p-value was less less than 0.1. As a reminder, the slope variables use units of change per *minute*; consequently, some of the coefficients for slope variables (e.g., Na_Slope_1680_mean in Model 5.5) are large.

To compare the sets of covariates included for different days I aligned them in terms of ranked Wald Z scores for each day. This was done for the daily DASn models shown in Models 5.3, 5.4, 5.5, 5.6, and 5.7. The alignment is shown in Figure 5-5.

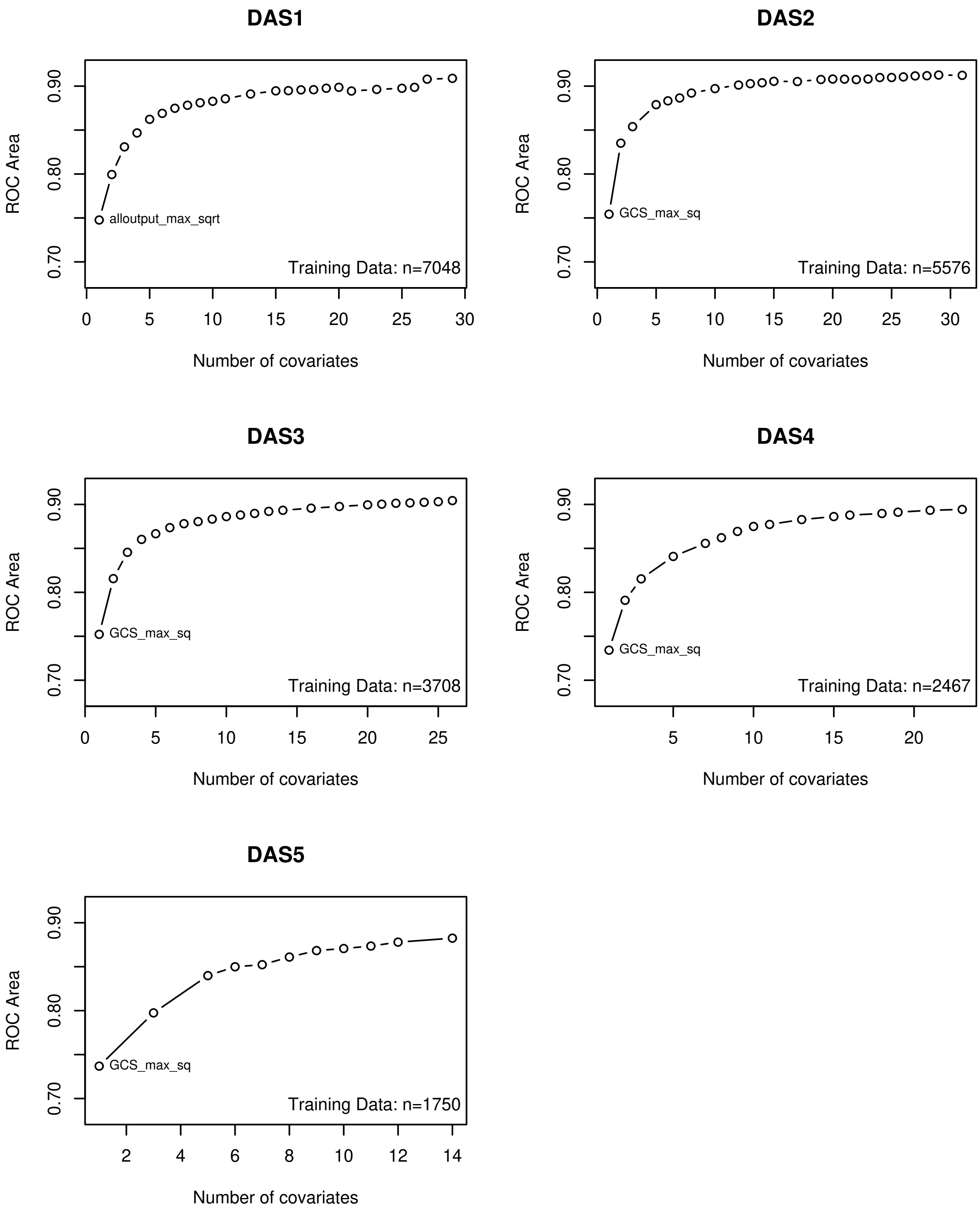


Figure 5-4: `DAS`n model selection (all development data)

Model 5.3 Final DAS1 model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
6364	2e-06	1609.3	22	0	0.9	0.8
Gamma	Tau-a	R2	Brier			
0.802	0.157	0.447	0.066			

	Coef	S.E.	Wald Z	P
GCS_max_sq	-0.0072777	8.114e-04	-8.97	0.0000
alloutput_max_sqrt	-0.0262779	3.053e-03	-8.61	0.0000
GCSrdv_mean	-0.3697803	4.745e-02	-7.79	0.0000
INR_mean_i	-1.5899254	2.558e-01	-6.22	0.0000
pacemkr_max	-0.9916790	2.072e-01	-4.79	0.0000
Insulin_mean_la	-0.0685988	1.792e-02	-3.83	0.0001
LactateM_sd_i	-0.0000371	1.107e-05	-3.35	0.0008
Intercept	0.1527354	6.926e-01	0.22	0.8255
MetCarcinoma_min	0.6146280	2.600e-01	2.36	0.0181
dopLg_mean_la	0.0841886	2.606e-02	3.23	0.0012
HemMalig_min	0.7508461	2.261e-01	3.32	0.0009
Lasix_max_lam	0.2567816	7.492e-02	3.43	0.0006
CO2_mean_i	15.0695137	4.323e+00	3.49	0.0005
Platelets_mean_i	26.1330763	6.959e+00	3.76	0.0002
SpO2_mean_am	0.1398357	3.270e-02	4.28	0.0000
PaO2toFiO2_mean	0.3556279	8.301e-02	4.28	0.0000
hrmVA_max	0.4476574	1.029e-01	4.35	0.0000
hospTime_min_sqrt	0.0065811	1.474e-03	4.46	0.0000
ShockIdx_mean_sq	0.6911597	1.489e-01	4.64	0.0000
Na_max_am	0.0722140	1.467e-02	4.92	0.0000
jaundiceSkin_mean_la	0.2107372	4.217e-02	5.00	0.0000
BUNtoCr_min_sqrt	0.2962917	4.515e-02	6.56	0.0000
Age_min_sq	0.0001780	2.684e-05	6.63	0.0000

Model 5.4 Final DAS2 model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
5179	0.002	1305.37	24	0	0.91	0.821
Gamma	Tau-a	R2	Brier			
0.823	0.149	0.463	0.06			

	Coef	S.E.	Wald Z	P
GCS_max_sq	-0.0116307	8.848e-04	-13.14	0.0000
INR_max_i	-2.2266609	2.892e-01	-7.70	0.0000
pacemkr_max	-1.1589393	2.168e-01	-5.35	0.0000
GCSrdv_mean	-0.1544221	3.357e-02	-4.60	0.0000
alloutput_min_sqrt	-0.0181251	3.952e-03	-4.59	0.0000
UrineOutB_max_sqrt	-0.0330005	1.069e-02	-3.09	0.0020
CV_HRrng_max	-0.0083858	2.722e-03	-3.08	0.0021
Insulin_mean_la	-0.0496108	1.927e-02	-2.57	0.0100
AIDS_min	0.9434952	4.231e-01	2.23	0.0257
MetCarcinoma_min	0.6917244	3.027e-01	2.28	0.0223
Intercept	1.7166820	6.954e-01	2.47	0.0136
RESP_mean_sq	0.0007393	2.853e-04	2.59	0.0096
HemMalig_min	0.8108085	2.517e-01	3.22	0.0013
hrmVA_max	0.5340514	1.611e-01	3.32	0.0009
Platelets_max_am	0.0021881	6.535e-04	3.35	0.0008
ShockIdx_mean_sq	0.6411199	1.902e-01	3.37	0.0008
hospTime_min_sqrt	0.0062001	1.745e-03	3.55	0.0004
Sympathomimetic_agent_min	0.6345344	1.762e-01	3.60	0.0003
Na_max_am	0.0682271	1.816e-02	3.76	0.0002
SpO2.oor30.t_mean_sqrt	0.3533329	9.198e-02	3.84	0.0001
jaundiceSkin_mean_la	0.1731920	4.284e-02	4.04	0.0001
beta.Blocking_agent_mean_lam	0.3351855	8.097e-02	4.14	0.0000
Amiodarone_min_am	3.0663969	6.177e-01	4.96	0.0000
Age_min_sq	0.0002075	3.311e-05	6.27	0.0000
BUNtoCr_mean_sqrt	0.3341310	4.989e-02	6.70	0.0000

Model 5.5 Final DAS3 model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
3526	0.003	964.81	26	0	0.904	0.809
Gamma	Tau-a	R2	Brier			
0.811	0.169	0.463	0.069			

	Coef	S.E.	Wald Z	P
GCS_max_sq	-0.0118516	9.892e-04	-11.98	0.0000
alloutput_max_sqrt	-0.0211350	3.189e-03	-6.63	0.0000
INR_max_i	-2.1664111	3.317e-01	-6.53	0.0000
GCSrng_mean_sq	-0.0067824	1.661e-03	-4.08	0.0000
svCSRU_max	-0.8094515	2.598e-01	-3.12	0.0018
Platelets_Slope_1680_min	-5.3871425	1.968e+00	-2.74	0.0062
pressD01_mean_am	-5.2563880	1.918e+00	-2.74	0.0061
MetCarcinoma_min	0.8145271	3.686e-01	2.21	0.0271
Platelets_max_am	0.0017373	7.684e-04	2.26	0.0238
temp_mean_am	0.2003711	8.752e-02	2.29	0.0220
jaundiceSkin_mean_la	0.1106708	4.724e-02	2.34	0.0192
Intercept	3.0339128	1.256e+00	2.41	0.0157
Na_Slope_1680_mean	92.5097926	3.760e+01	2.46	0.0139
Mg_min_sq	0.1260452	5.001e-02	2.52	0.0117
InputB_mean_sqrt	0.0551695	2.000e-02	2.76	0.0058
SBPm.oor30.t_max_sq	0.0003288	1.171e-04	2.81	0.0050
beta.Blocking_agent_mean_lam	0.2764307	9.746e-02	2.84	0.0046
MBPm.pr_min_am	2.9511087	1.030e+00	2.87	0.0042
Na_max_am	0.0620932	2.148e-02	2.89	0.0038
HCTrdv_max	0.0635881	2.155e-02	2.95	0.0032
Lasix_perKg_max_sqrt	3.1703119	9.820e-01	3.23	0.0012
Sympathomimetic_agent_min	0.6961636	2.127e-01	3.27	0.0011
RESP_mean_sq	0.0010697	3.236e-04	3.31	0.0009
hospTime_min_sqrt	0.0081331	1.978e-03	4.11	0.0000
BUNtoCr_min_sqrt	0.2633152	5.732e-02	4.59	0.0000
Age_min_sq	0.0001728	3.633e-05	4.76	0.0000
SpO2.oor30.t_mean_sqrt	0.5154426	1.015e-01	5.08	0.0000

Model 5.6 Final DAS4 model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
2351	0.002	711.56	20	0	0.892	0.784

Gamma	Tau-a	R2	Brier
0.786	0.192	0.467	0.078

	Coef	S.E.	Wald Z	P
GCS_max_sq	-7.201e-03	0.0012819	-5.62	0.0000
Platelets_Slope_1680_min	-1.394e+01	2.8825629	-4.84	0.0000
svCSRU_max	-1.470e+00	0.3113389	-4.72	0.0000
RikerSAS_mean	-5.760e-01	0.1416631	-4.07	0.0000
GCSrng_mean_sq	-7.241e-03	0.0017872	-4.05	0.0001
INR_mean_i	-1.813e+00	0.4500497	-4.03	0.0001
pressD01_mean_am	-8.687e+00	2.1800306	-3.98	0.0001
Insulin_sd_sq	-3.347e+00	1.5514995	-2.16	0.0310
urineByHr_mean_sqrt	-4.050e-02	0.0246355	-1.64	0.1002
Intercept	2.880e+00	1.3636750	2.11	0.0347
temp_mean_am	2.229e-01	0.0955437	2.33	0.0197
jaundiceSkin_mean_la	1.477e-01	0.0550705	2.68	0.0073
ShockIdx_max	7.627e-01	0.2664474	2.86	0.0042
MBPm.pr_min_am	3.238e+00	1.1310217	2.86	0.0042
hospTime_min_sqrt	6.546e-03	0.0022385	2.92	0.0034
Platelets_mean_i	2.962e+01	9.6453810	3.07	0.0021
CO2_mean_i	3.455e+01	8.8649392	3.90	0.0001
Sympathomimetic_agent_min	1.251e+00	0.2350043	5.32	0.0000
SpO2.oor30.t_mean_sqrt	6.704e-01	0.1255200	5.34	0.0000
BUNtoCr_mean_sqrt	3.361e-01	0.0603984	5.56	0.0000
Age_min_sq	2.402e-04	0.0000412	5.83	0.0000

Model 5.7 Final DAS5 model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
1690	4e-05	524.36	15	0	0.883	0.766

Gamma	Tau-a	R2	Brier
0.767	0.212	0.45	0.091

	Coef	S.E.	Wald Z	P
GCS_max_sq	-7.754e-03	1.361e-03	-5.70	0.0000
svCSRU_max	-1.633e+00	3.159e-01	-5.17	0.0000
RikerSAS_mean	-7.203e-01	1.497e-01	-4.81	0.0000
Platelets_Slope_1680_min	-1.313e+01	3.398e+00	-3.87	0.0001
INR_mean_i	-1.865e+00	4.900e-01	-3.81	0.0001
GCSrng_mean_sq	-6.519e-03	1.808e-03	-3.61	0.0003
Intercept	-2.188e+00	9.628e-01	-2.27	0.0231
HemMalig_min	6.189e-01	3.600e-01	1.72	0.0856
Mg_min_sq	1.578e-01	6.055e-02	2.61	0.0092
MBPm.pr_min_am	3.467e+00	1.158e+00	2.99	0.0028
ShockIdx_max	8.850e-01	2.754e-01	3.21	0.0013
InputOtherBloodB_mean_lam	3.944e-01	1.134e-01	3.48	0.0005
BUNtoCr_mean_sqrt	2.580e-01	6.446e-02	4.00	0.0001
CO2_mean_i	4.318e+01	9.452e+00	4.57	0.0000
SpO2.oor30.t_mean_sqrt	8.058e-01	1.421e-01	5.67	0.0000
Age_min_sq	2.873e-04	4.602e-05	6.24	0.0000

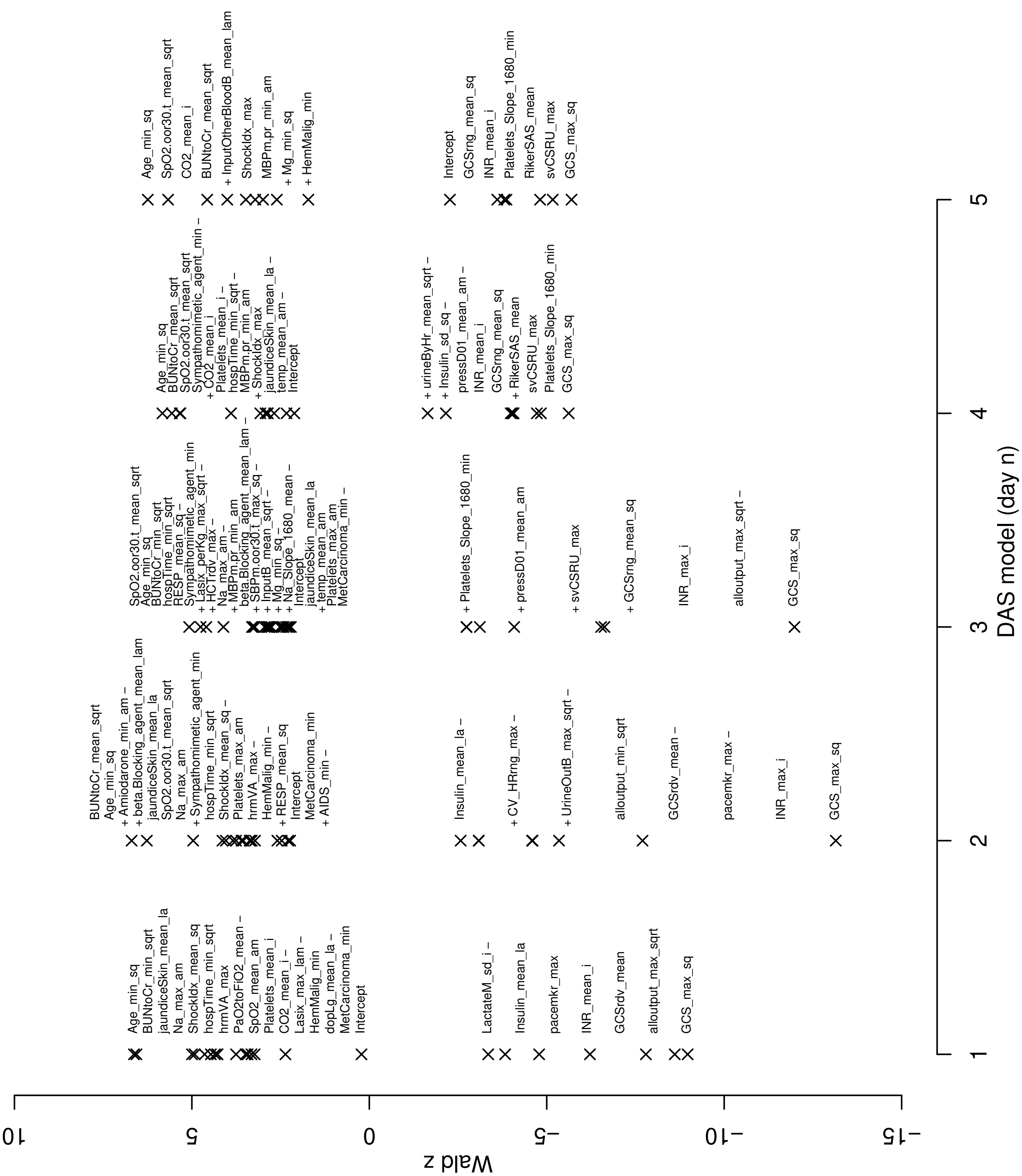


Figure 5-5: Ranked comparison of DASn inputs over days $n \in \{1, 2, 3, 4, 5\}$. Input names are ranked (but equally spaced) for the positive Wald Z scores and the negative Wald Z scores for each model. On a given day, variables absent from the previous day are prefixed with "+", and variables absent from the following day suffixed with "-".

DASn Validation (Development Data)

As with the SDAS model, we first validated the DASn models on the development data to examine their goodness of fit. Figure 5-6 provides the ROC curves for the DASn models. Table 5.7 shows the number of valid observations, "Obs", the number of missing observations (where at least one variable was missing), the number of variables in the model, the AUC performance, and the Hosmer-Lemeshow H statistics for each model. It also lists the *d.f.* used for the Hosmer-Lemeshow χ^2 comparison and the resulting p-values. The Hosmer-Lemeshow goodness of Fit tests using the deciles of risk can be found for each of the DASn models in Appendix D.

Table 5.7: DASn model characteristics (development data)

Day	Obs	Missing	Vars	AUC	H	p (*d.f.*)
1	6364	684	22	0.900	13.03	0.043 (6)
2	5179	397	24	0.910	9.07	0.170 (6)
3	3526	182	26	0.904	8.84	0.183 (6)
4	2351	116	20	0.892	4.36	0.499 (5)
5	1690	60	15	0.883	3.59	0.732 (6)

To assess the fit, I also examined goodness of fit statistics from bootstrapping with 150 samples (with replacement) on the development data. The results are shown in Tables 5.8, 5.9, 5.10, 5.11, and 5.12.

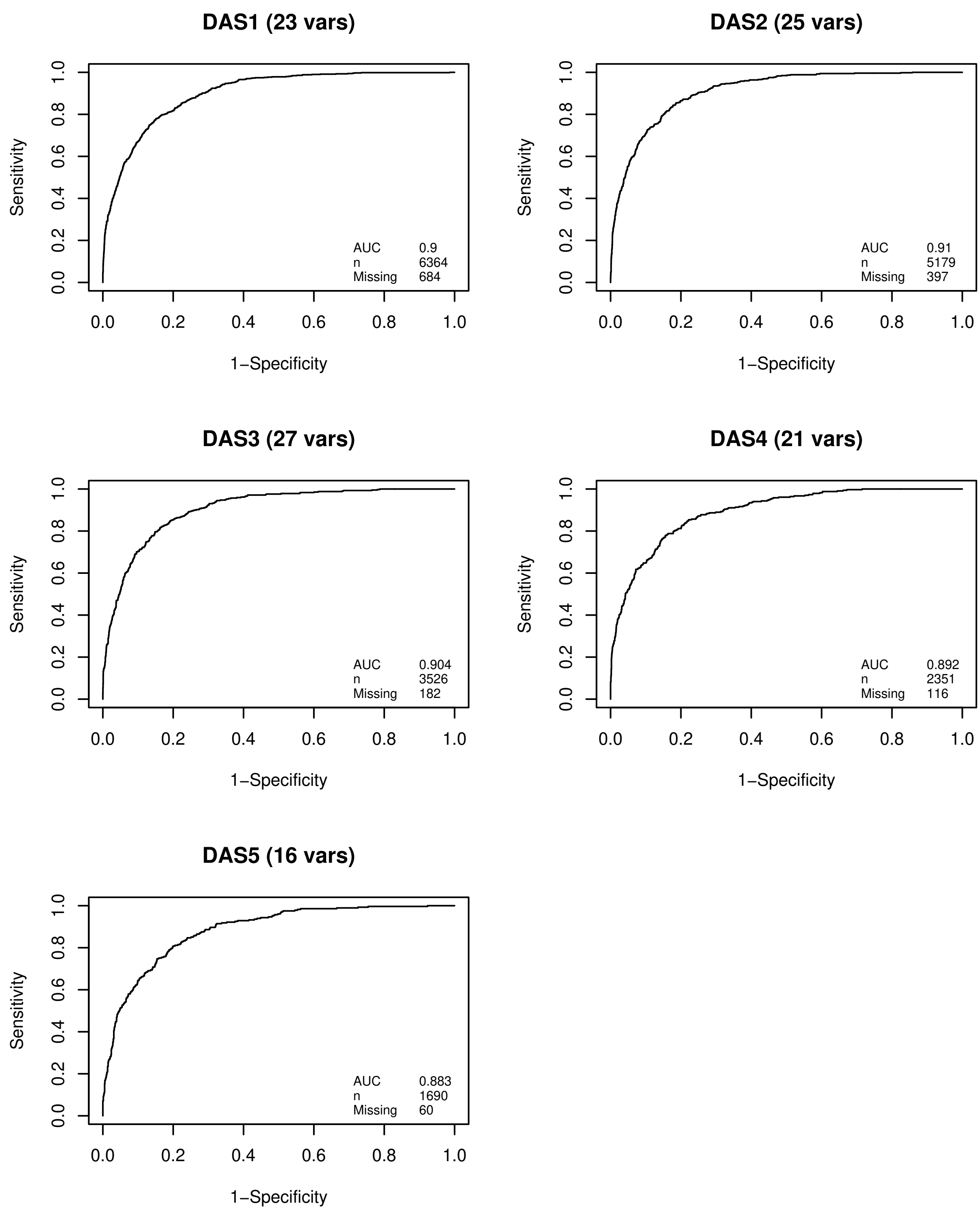


Figure 5-6: `DAS`n ROC curves (development data)

Table 5.8: DAS1 bootstrapped goodness of fit statistics (development data). *not all samples converged

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples*
D_{xy}	0.800	0.802	0.797	0.006	0.794	108
R^2	0.447	0.451	0.442	0.009	0.437	108
Intercept	0.000	0.000	-0.029	0.029	-0.029	108
Slope	1.000	1.000	0.980	0.020	0.980	108
E_{max}	0.000	0.000	0.010	0.010	0.010	108
D	0.253	0.256	0.249	0.007	0.246	108
U	0.000	0.000	0.000	0.000	0.000	108
Q	0.253	0.256	0.249	0.007	0.246	108
B	0.066	0.066	0.067	-0.001	0.067	108

Table 5.9: DAS2 bootstrapped goodness of fit statistics (development data). *not all samples converged

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples*
D_{xy}	0.821	0.825	0.816	0.010	0.811	126
R^2	0.463	0.472	0.456	0.017	0.447	126
Intercept	0.000	0.000	-0.061	0.061	-0.061	126
Slope	1.000	1.000	0.961	0.039	0.961	126
E_{max}	0.000	0.000	0.020	0.020	0.020	126
D	0.252	0.258	0.247	0.011	0.241	126
U	0.000	0.000	0.000	-0.001	0.000	126
Q	0.252	0.258	0.247	0.012	0.241	126
B	0.060	0.059	0.061	-0.002	0.062	126

Table 5.10: DAS3 bootstrapped goodness of fit statistics (development data). *not all samples converged

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples*
D_{xy}	0.809	0.815	0.801	0.014	0.795	149
R^2	0.463	0.474	0.452	0.022	0.441	149
Intercept	0.000	0.000	-0.056	0.056	-0.056	149
Slope	1.000	1.000	0.947	0.053	0.947	149
E_{max}	0.000	0.000	0.022	0.022	0.022	149
D	0.273	0.280	0.266	0.014	0.259	149
U	-0.001	-0.001	0.000	-0.001	0.000	149
Q	0.274	0.281	0.265	0.015	0.259	149
B	0.069	0.067	0.070	-0.003	0.072	149

Table 5.11: DAS4 bootstrapped goodness of fit statistics (development data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.784	0.793	0.777	0.016	0.769	150
R^2	0.467	0.480	0.456	0.024	0.443	150
Intercept	0.000	0.000	-0.068	0.068	-0.068	150
Slope	1.000	1.000	0.943	0.057	0.943	150
E_{max}	0.000	0.000	0.025	0.025	0.025	150
D	0.302	0.313	0.294	0.019	0.284	150
U	-0.001	-0.001	0.001	-0.002	0.001	150
Q	0.303	0.313	0.293	0.020	0.283	150
B	0.078	0.077	0.080	-0.003	0.082	150

Table 5.12: DAS5 bootstrapped goodness of fit statistics (development data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.766	0.770	0.759	0.011	0.755	150
R^2	0.450	0.458	0.440	0.018	0.431	150
Intercept	0.000	0.000	-0.046	0.046	-0.046	150
Slope	1.000	1.000	0.955	0.045	0.955	150
E_{max}	0.000	0.000	0.018	0.018	0.018	150
D	0.310	0.318	0.302	0.016	0.294	150
U	-0.001	-0.001	0.001	-0.002	0.001	150
Q	0.311	0.319	0.301	0.018	0.293	150
B	0.091	0.090	0.093	-0.002	0.094	150

5.2.3 Held-out Validation

The last step for the daily models was to validate them on separate validation data. To evaluate discrimination performance, the ROC curves for the `SDAS` model (on all days, and on individual days 1 through 5) are provided in Figure 5-7. The ROC curves for the individual `DAS`n models (days 1 through 5) are provided in Figure 5-8.

The H calibration statistics for these models on the held-out validation data are listed in Table 5.13. For each model and day, all predictions available were used for the calibration calculation (i.e., the groups of patients were not matched between models). Performance values for matched patient groups are shown later in this chapter when I compare models directly with each other.

Table 5.13: Hosmer-Lemeshow calibration summaries for `SDAS` and `DAS`n (validation data)

Model	Days	H	p (*d.f.*)
`SDAS`	all	51.10	6.70e-08 (9)
`SDAS`	1	14.39	0.045 (7)
`SDAS`	2	15.31	0.018 (6)
`SDAS`	3	5.01	0.542 (6)
`SDAS`	4	10.37	0.110 (6)
`SDAS`	5	7.55	0.183 (5)
`DAS`n	1	24.84	0.001 (7)
`DAS`n	2	17.52	0.008 (6)
`DAS`n	3	11.03	0.087 (6)
`DAS`n	4	11.91	0.064 (6)
`DAS`n	5	6.65	0.248 (5)

Table 5.14 shows the individual deciles for the `SDAS` mortality predictions on ICU days 1 through 7. Table 5.15 shows the individual deciles for `DAS1` on the first ICU day.

The Hosmer-Lemeshow test is often sensitive to the choice of binning used for the predictions. An alternative to the Hosmer-Lemeshow test is to plot the actual probability versus the predicted probability. Figures 5-9, 5-10, 5-11, 5-12 and 5-13 show calibration plots for `SDAS` and `DAS`n for days 1 through 5. For comparison with the development data bootstrapped statistics in Tables 5.6, 5.8, 5.9, 5.10, 5.11, and 5.12, the uncorrected probabilities were used. The corrected slope and intercept in these tables could be applied to shrink the probabilities and improve the calibration performance — although examination of Figures 5-9, 5-10, 5-11, 5-12 and 5-13, indicates that the correction would still be optimistic.

In the figures, the risk deciles from the Hosmer-Lemeshow tests are represented by triangles (referred to as “grouped observations” in the legends). In addition to the logistic fit between the actual and the predicted probabilities (dashed line), the

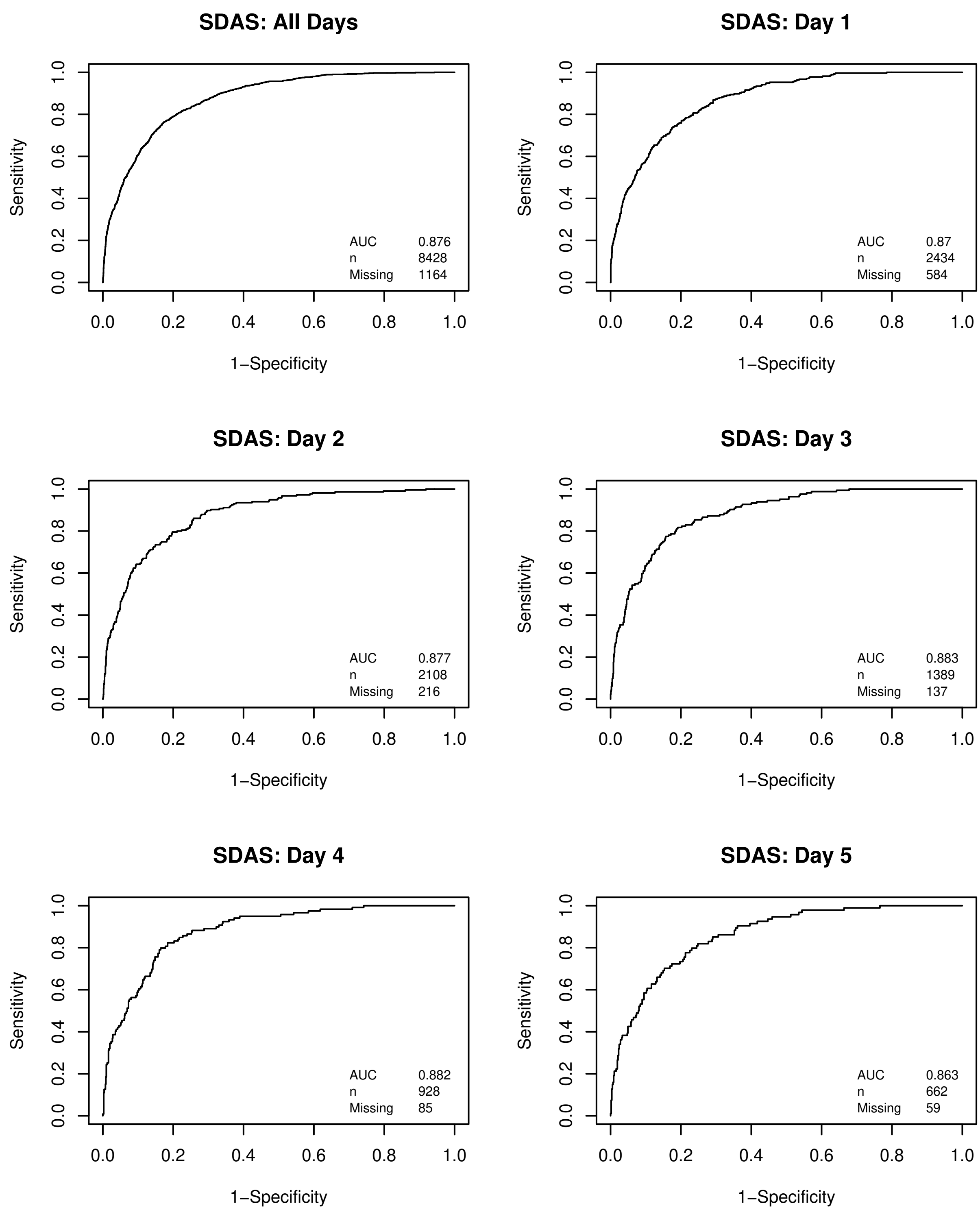


Figure 5-7: `SDAS` ROC curves (validation data)

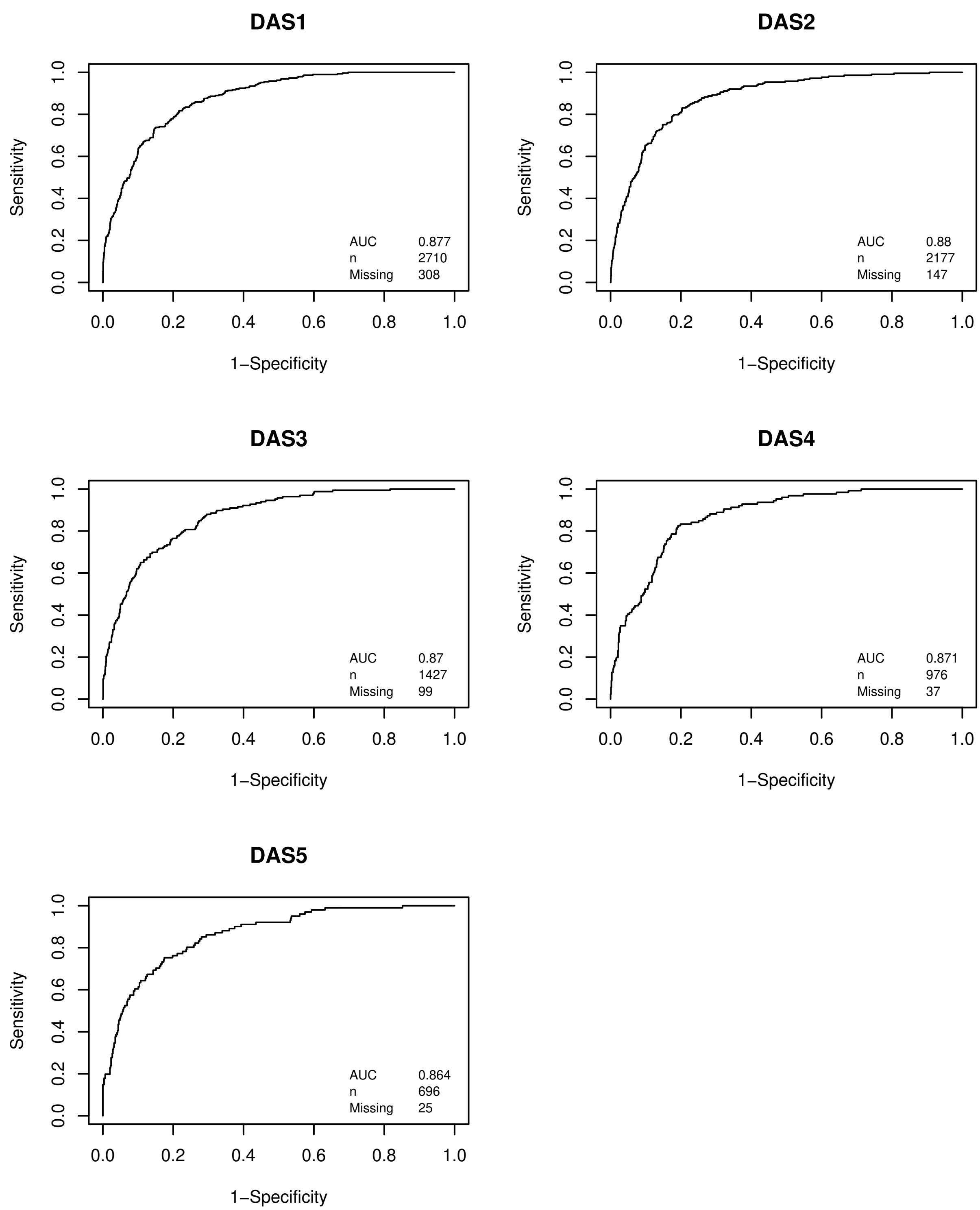


Figure 5-8: `DAS`n ROC curves (validation data)

Table 5.14: `SDAS` Hosmer-Lemeshow H deciles of risk (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-2	[0.000290,0.00623)	0.003	3	5.6	1683	1680.4	1686
3	[0.006233,0.01158)	0.009	7	7.2	836	835.8	843
4	[0.011578,0.02054)	0.016	19	13.3	824	829.7	843
5	[0.020540,0.03527)	0.028	28	23.2	814	818.8	842
6	[0.035274,0.05916)	0.046	44	38.8	799	804.2	843
7	[0.059156,0.09543)	0.076	80	63.8	763	779.2	843
8	[0.095427,0.17333)	0.130	136	109.7	707	733.3	843
9	[0.173330,0.37287)	0.256	233	216.2	610	626.8	843
10	[0.372869,0.99862]	0.635	456	535.1	386	306.9	842

$\chi^2 = 51.10,\ d.f. = 9;\ p = 0.000$

Table 5.15: DAS1 Hosmer-Lemeshow H deciles of risk (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-4	[9.42e-05,0.01731)	0.006	7	6.6	1077	1077.4	1084
5	[1.73e-02,0.02830)	0.023	8	6.2	263	264.8	271
6	[2.83e-02,0.04404)	0.036	15	9.6	256	261.4	271
7	[4.40e-02,0.07342)	0.057	18	15.5	253	255.5	271
8	[7.34e-02,0.12691)	0.097	42	26.3	229	244.7	271
9	[1.27e-01,0.28972)	0.191	66	51.8	205	219.2	271
10	[2.90e-01,0.99931]	0.553	134	149.8	137	121.2	271

$\chi^2 = 22.96,\ d.f. = 7;\ p = 0.002$

calibration plots also provide a nonparametric fit that is shown as a dotted line. The relative frequencies for each predicted probability are indicated by histogram along the bottom. Further details about the statistics that are listed can be found in Appendix A.

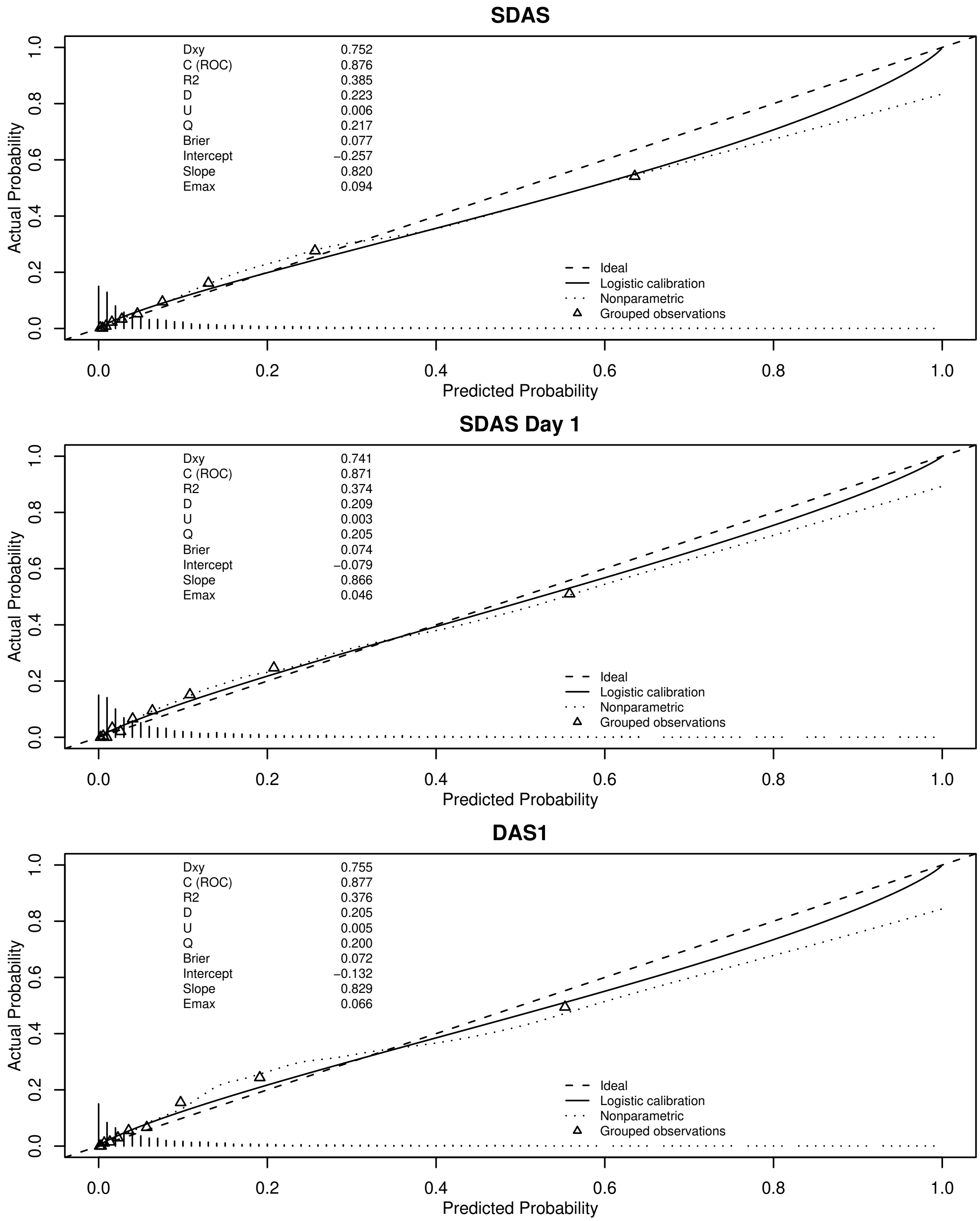


Figure 5-9: Calibration plots for `SDAS`, `SDAS` day 1, and `DAS1` (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

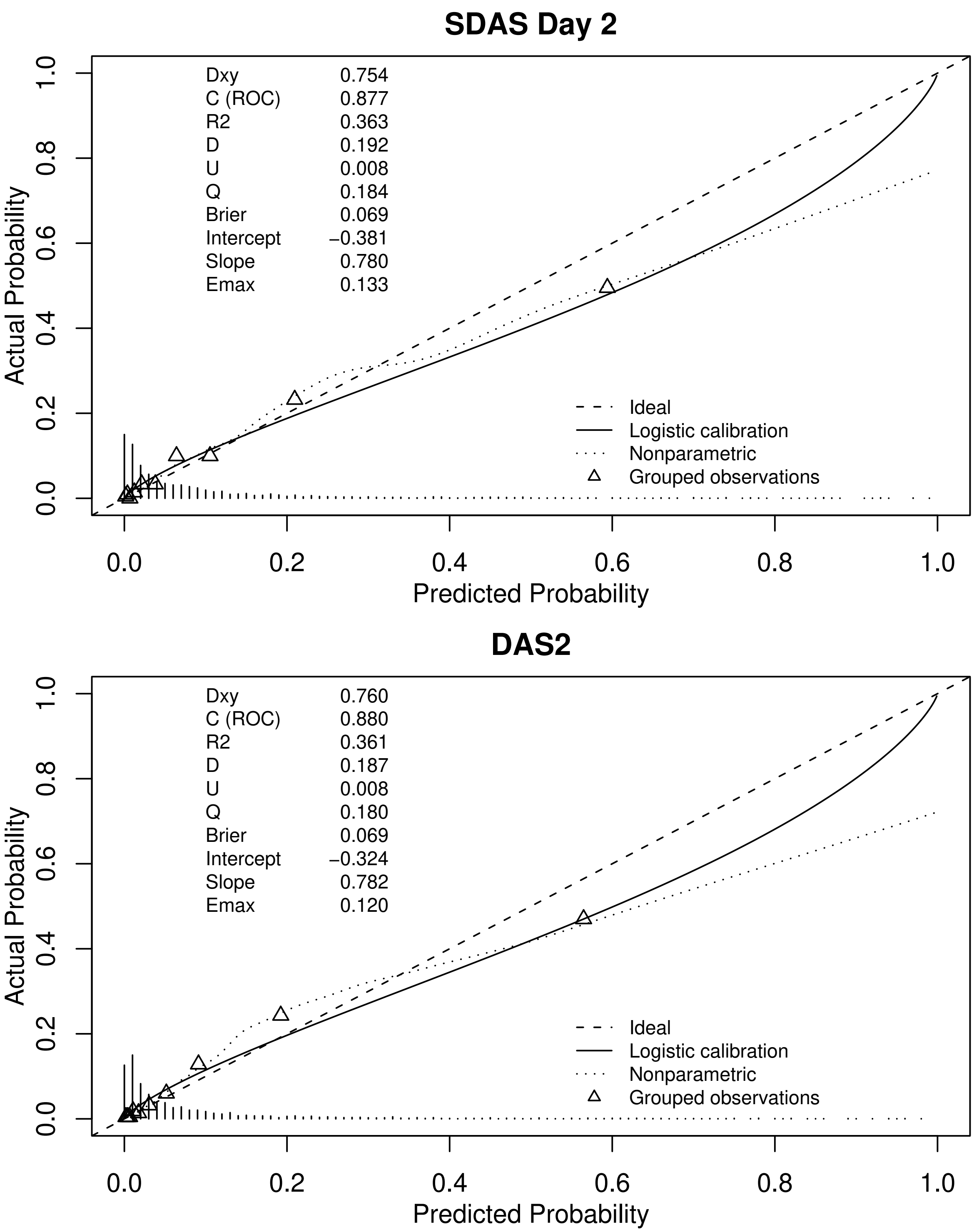
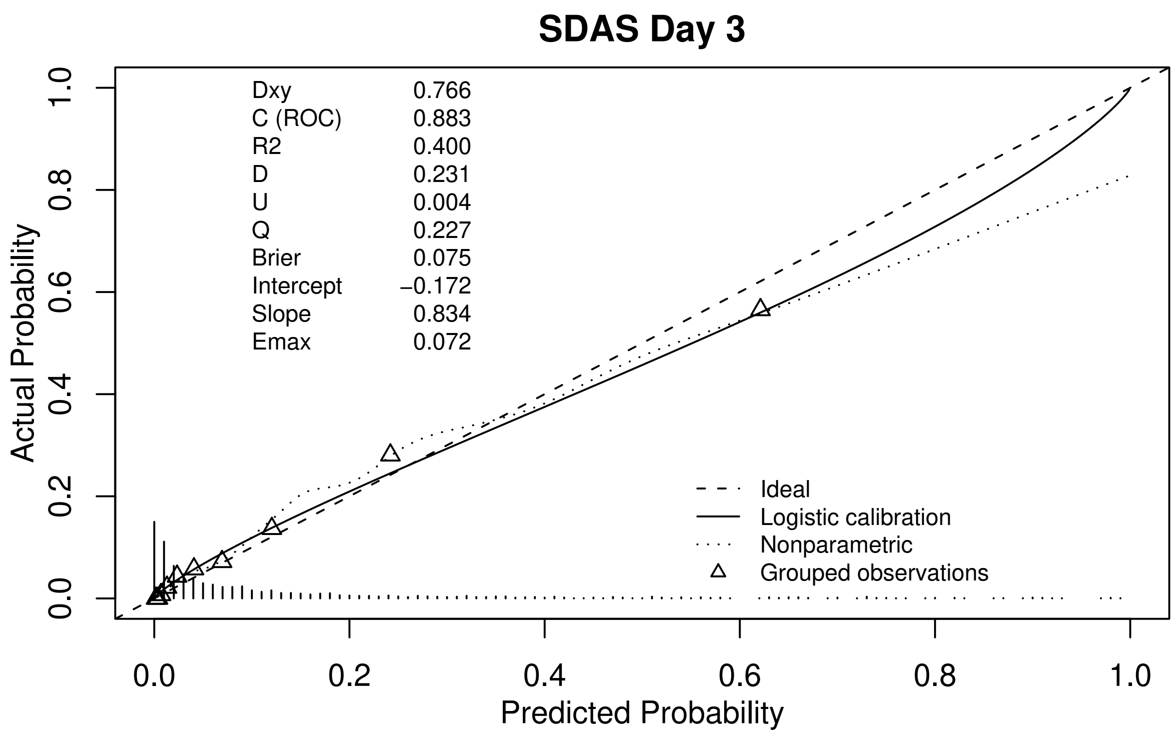


Figure 5-10: Calibration plots for `SDAS` day 2 and `DAS2` (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.



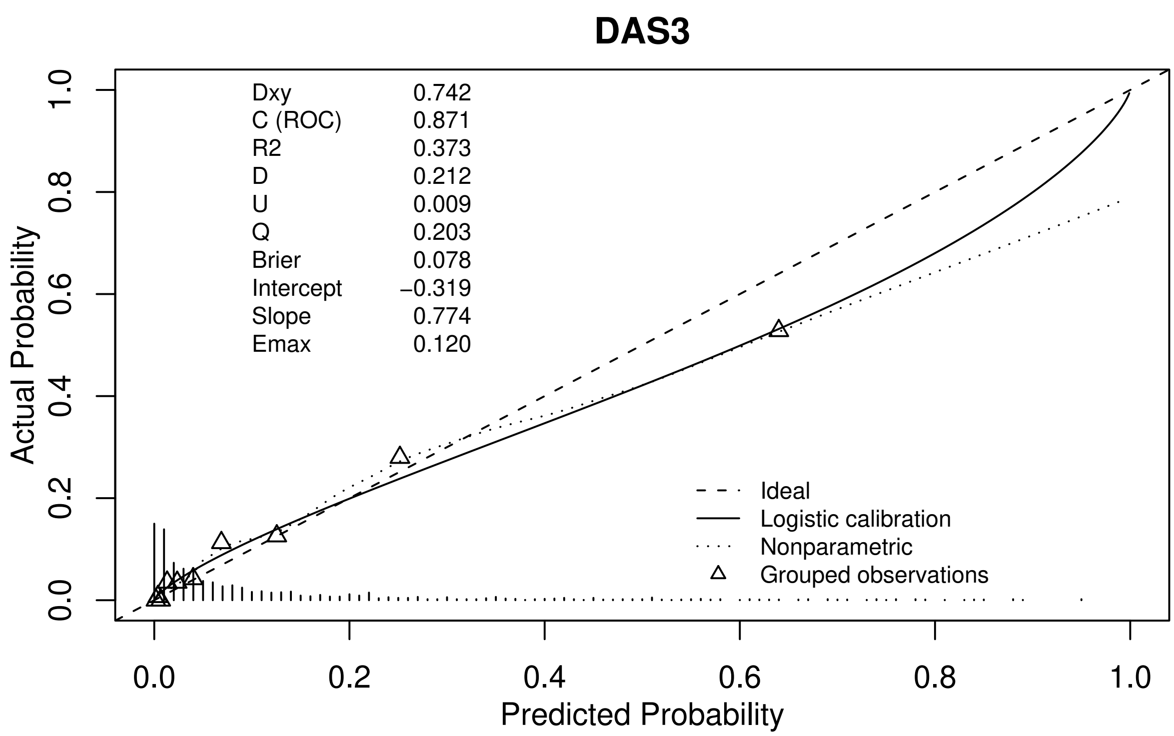


Figure 5-11: Calibration plots for `SDAS` day 3 and `DAS3` (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

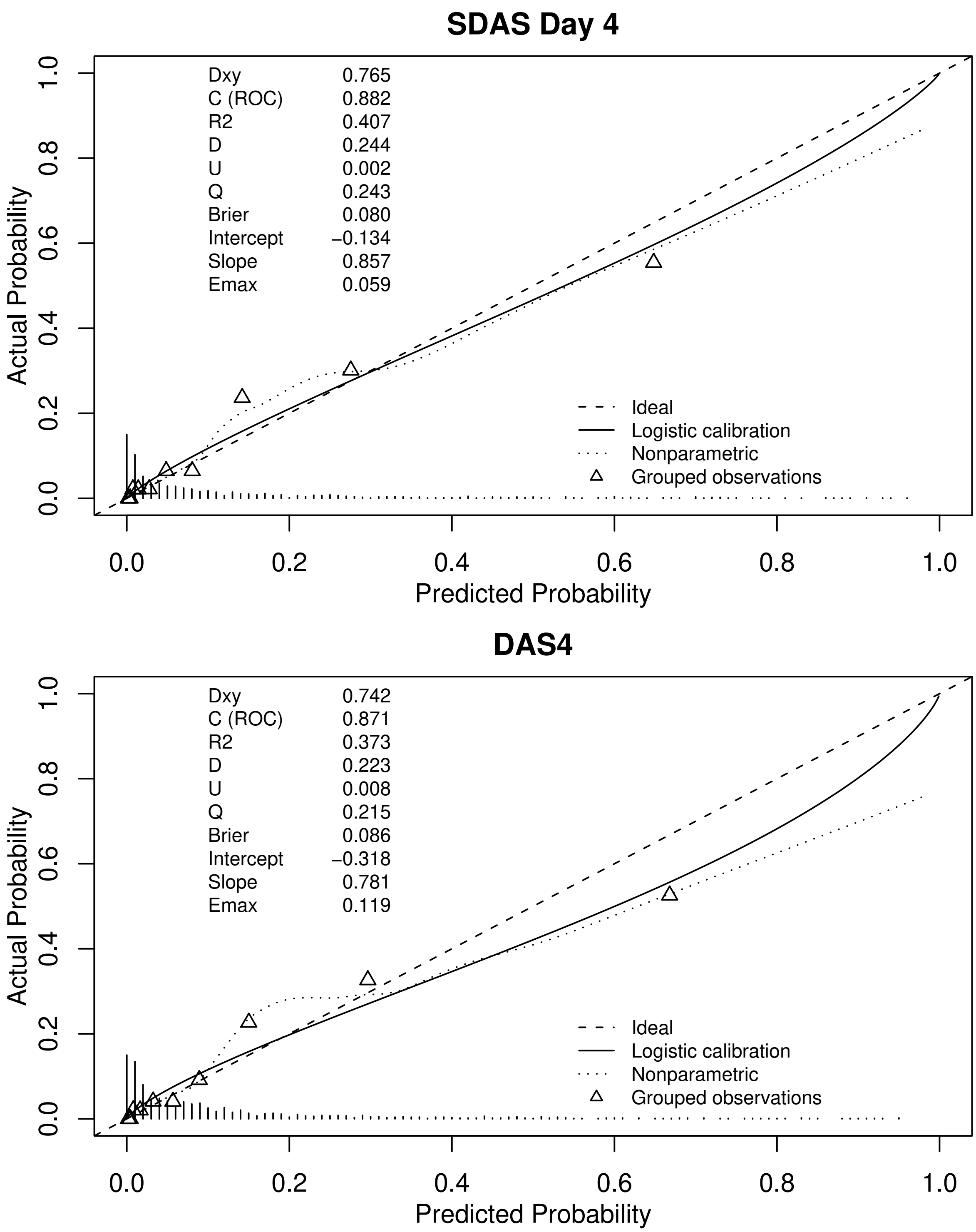


Figure 5-12: Calibration plots for `SDAS` day 4 and `DAS4` (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

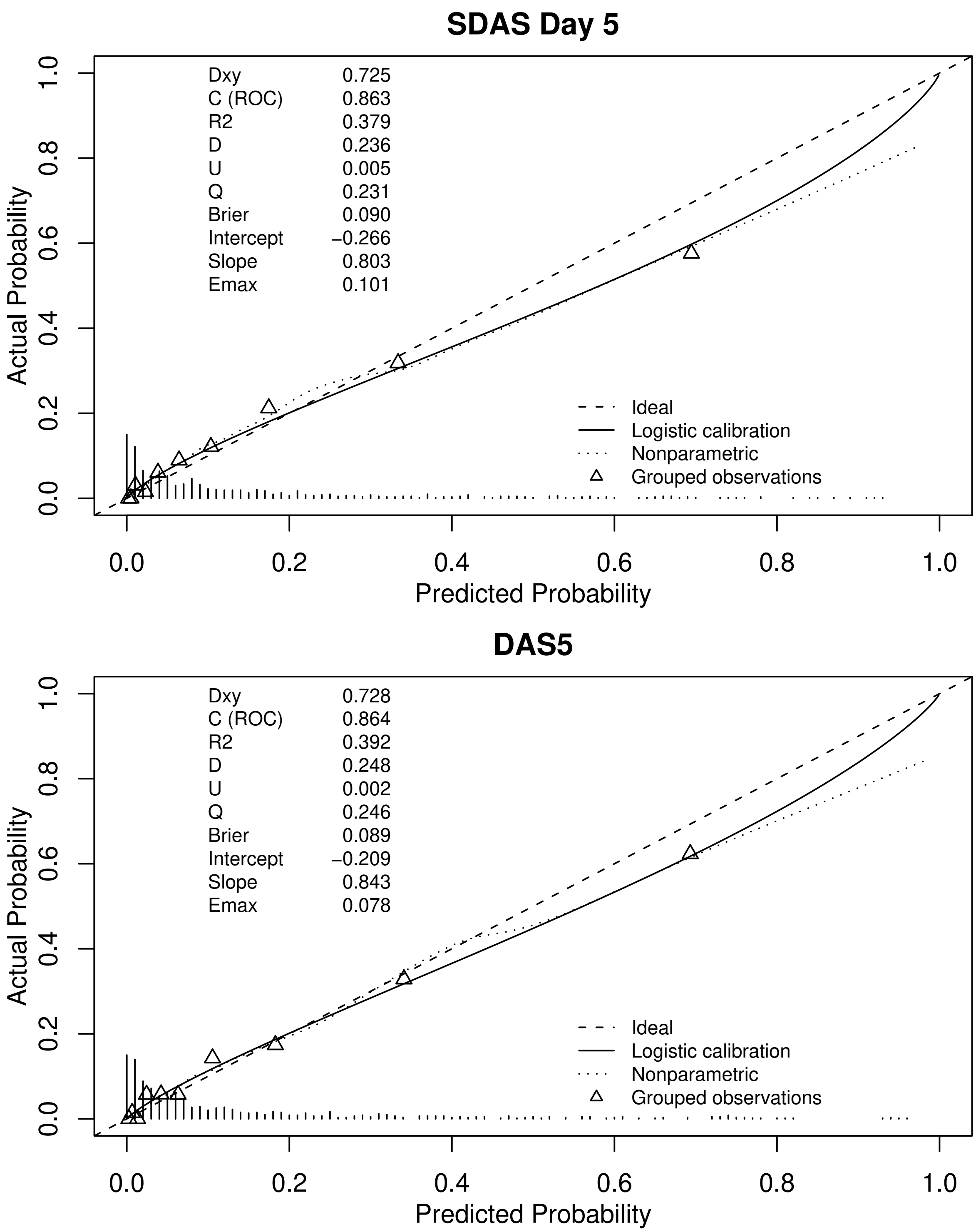


Figure 5-13: Calibration plots for `SDAS` day 5 and `DAS5` (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

5.3 RAS: Real-time Acuity Score

5.3.1 RAS Model

The third model type that I examined was a real-time acuity score (`RAS`) that used all observations available in the final dataset (see Table 5.2). After filtering, univariate analysis, and collinearity analysis, 200 variables remained for model training. Using five-fold cross-validation, five initial models were trained. Using backward elimination, Figure 5-14 shows the performance for new models, on their respective validation folds, as the AIC threshold was increased (and variables were dropped from the models). The top 60 variables from each of the best four models (the model for fold 3, with its weak validation performance, was excluded) were combined to train a final model using all of the development data.[2] The AUC performance of models built using this final feature set, as the number of inputs changes, is shown in Figure 5-15.

From Figure 5-15, it appears that the improvement in performance by using more than 50 variables is negligible. Using the top 50 variables, a number of manual refinements were made. First, as with the `SDAS` and `DAS`n models, Hematologic Malignancy, Metastatic Carcinoma, AIDS, and WBC count were manually added to the model (they each had a significant contribution). Next, a number of redundant features were eliminated: (1) `Dopamine_sqrt`, (2) `Dobutamine_perKg_sqrt`, (3) `24hUrOut_sqrt`, (4) `SBPmThreshCnt_sqrt`, (5) `SBPmThreshCntF_sqrt`, (6) `totIV`, (7) `24hBal`, (8) `SpO2CritEvnts.24h_la`, (9) `impairedSkin`, and (10) `LOSBalrdv`. After these variables were removed, the `ventLen_sqrt` was also removed due to a low significance (Wald Z = -6.09). The final model is described in Model 5.8. These adjustments had a negligible effect on the performance of the `RAS` model on the development data.

`RAS` Validation (Development Data)

The ROC curves for the development data are shown in Figure 5-16 on page 102. With `RAS` a number of summary functions are available to aggregate multiple predictions. Figure 5-16 shows three performance measures: (1) using *all* of the predictions ("All Patient Predictions"), (2) using the *mean* prediction over an entire patient's stay, and (3) using the *mean* daily prediction for a given day (days 1 through 7).

Table 5.16 lists the H and C statistics for each of the performance measures shown in Figure 5-16.

To further assess the fit, I also examined goodness of fit statistics from bootstrapping with 150 samples (with replacement) on the development data. The results are shown in Table 5.17.

[2]The cross-validation folds were helpful in understanding the risk of overfitting. Fold 3 from Figure 5-14, for example, fit the training data quite well but generalized to the validation data poorly.

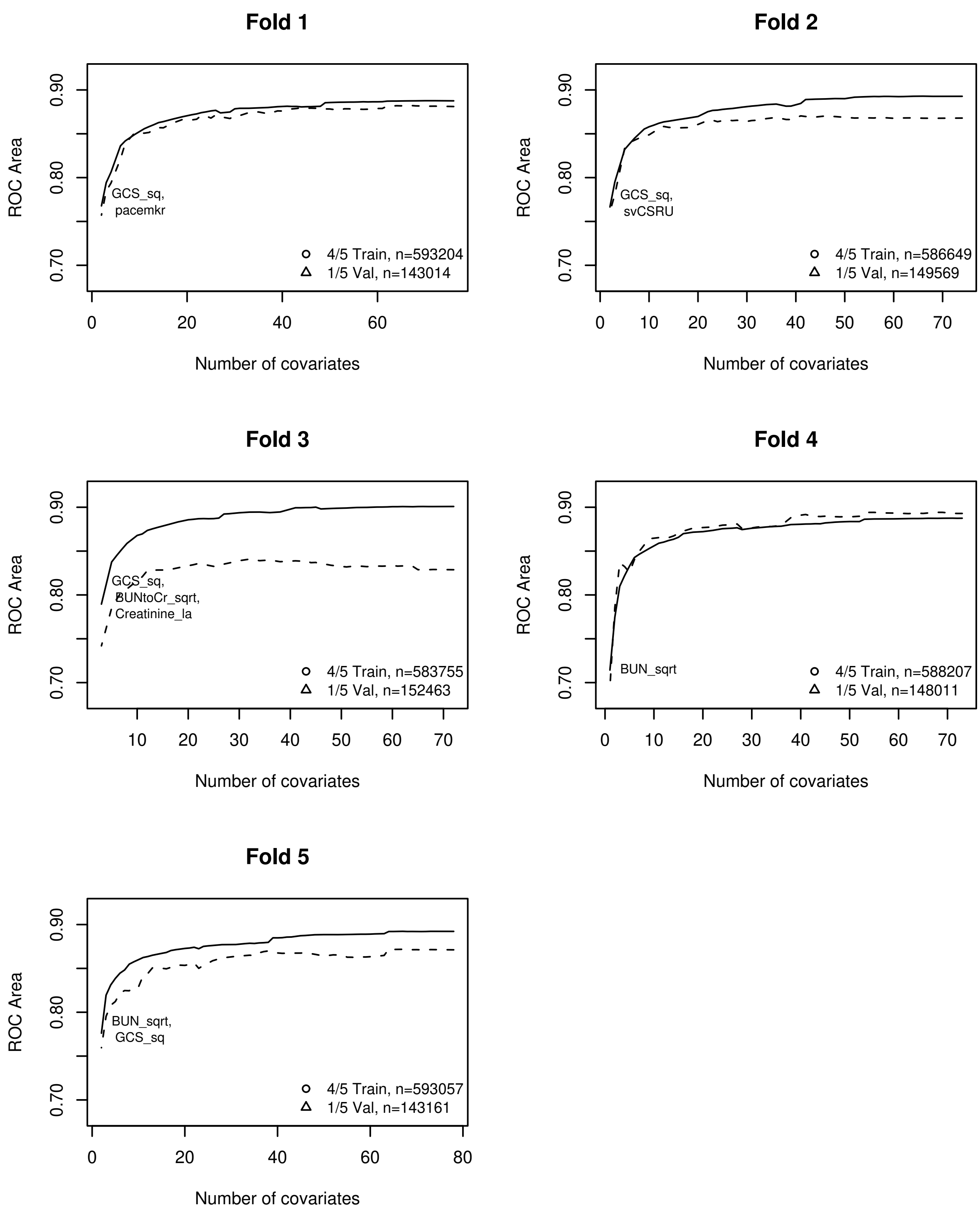


Figure 5-14: `RAS` model selection. Sensitivity to number of covariates on each cross-validation fold

Model 5.8 Final RAS model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
528850	2e-07	148516.3	43	0	0.885	0.769

Gamma	Tau-a	R2	Brier
0.771	0.19	0.436	0.084

	Coef	S.E.	Wald Z	P
GCS_sq	-6.448e-03	8.410e-05	-76.68	0
Intercept	-3.981e+00	6.471e-02	-61.51	0
svCSRU	-1.198e+00	2.095e-02	-57.19	0
pacemkr	-9.094e-01	2.019e-02	-45.04	0
alloutput_la	-9.355e-02	2.437e-03	-38.39	0
pressD01	-4.026e-01	1.140e-02	-35.32	0
Platelets_Slope_1680	-4.809e+00	1.670e-01	-28.80	0
Insulin	-3.085e-01	1.382e-02	-22.33	0
CV_HRrng_sqrt	-5.498e-02	2.974e-03	-18.49	0
GCSrng_am	-4.579e-02	2.499e-03	-18.32	0
MetCarcinoma	1.583e-01	3.246e-02	4.88	0
SBPm.oor120.t	1.292e-03	1.897e-04	6.81	0
PulsePres_i	6.265e+00	6.049e-01	10.36	0
admitWt_i	1.837e+01	1.421e+00	12.93	0
AIDS	5.078e-01	3.726e-02	13.63	0
RikerSAS_lam	1.215e-01	7.975e-03	15.23	0
Sandostatin_am	1.207e-02	6.423e-04	18.80	0
ShockIdx	4.788e-01	2.472e-02	19.37	0
temp_lam	9.002e-02	4.414e-03	20.40	0
cumPressorTime_am	9.326e-05	4.383e-06	21.28	0
INRrng	1.066e-01	4.958e-03	21.49	0
hrmVA_sqrt	7.107e-01	3.282e-02	21.66	0
CO2_am	3.167e-02	1.415e-03	22.38	0
WBC_am	2.045e-02	9.110e-04	22.44	0
Nondepolarizing_agent	7.739e-01	3.393e-02	22.81	0
PAPmeanM	3.213e-01	1.315e-02	24.44	0
Platelets_am	1.590e-03	6.318e-05	25.17	0
RESP	2.352e-02	8.055e-04	29.21	0
Lasix_perKg_lam	2.022e-01	6.741e-03	29.99	0
urineByHr.oor120.t	3.248e-03	1.072e-04	30.28	0
Antiarrhythmic_agent	5.823e-01	1.919e-02	30.35	0
Na_am	4.691e-02	1.530e-03	30.66	0
HemMalig	6.915e-01	2.240e-02	30.87	0
beta.Blocking_agent	1.037e+00	3.350e-02	30.96	0
SpO2.oor120.t_sqrt	9.728e-02	2.990e-03	32.54	0
PaO2toFiO2	2.257e-01	6.185e-03	36.49	0
hospTime_sqrt	5.970e-03	1.553e-04	38.44	0
index	1.022e-04	2.519e-06	40.57	0
jaundiceSkin	1.606e+00	3.770e-02	42.60	0
pressorSum.std_sqrt	5.373e-01	1.094e-02	49.11	0
Creatinine_la	4.255e-01	8.458e-03	50.31	0
INR_la	1.010e+00	1.626e-02	62.11	0
BUNtoCr_sqrt	3.065e-01	4.414e-03	69.45	0
Age_sq	2.640e-04	2.892e-06	91.28	0

Table 5.16: RAS Hosmer-Lemeshow calibration (development data)

Day(s)	Summary Func.	H	p ($d.f.$)	C	p ($d.f.$)	n
all	none	341.6	0 (8)	371.6	0.000 (8)	528850
all	mean	44.4	1.75e-07 (7)	51.45	7.49e-09 (7)	5977
1	mean	16.7	0.019 (7)	10.1	0.185 (7)	5719
2	mean	4.42	0.620 (6)	7.25	0.404 (7)	4975
3	mean	6.60	0.359 (6)	12.3	0.091 (7)	3363
4	mean	18.4	0.005 (6)	16.2	0.013 (6)	2248
5	mean	12.5	0.052 (6)	14.3	0.026 (6)	1604
6	mean	15.2	0.019 (6)	17.6	0.007 (6)	1188
7	mean	17.6	0.007 (6)	16.2	0.013 (6)	938

Table 5.17: RAS bootstrapped goodness of fit statistics (development data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.7694	0.7694	0.7693	0.0001	0.7693	150
R^2	0.4358	0.4359	0.4357	0.0002	0.4356	150
Intercept	0.0000	0.0000	-0.0004	0.0004	-0.0004	150
Slope	1.0000	1.0000	0.9998	0.0002	0.9998	150
E_{max}	0.0000	0.0000	0.0001	0.0001	0.0001	150
D	0.2808	0.2809	0.2807	0.0002	0.2806	150
U	0.0000	0.0000	0.0000	0.0000	0.0000	150
Q	0.2808	0.2809	0.2807	0.0002	0.2806	150
B	0.0838	0.0839	0.0838	0.0000	0.0838	150

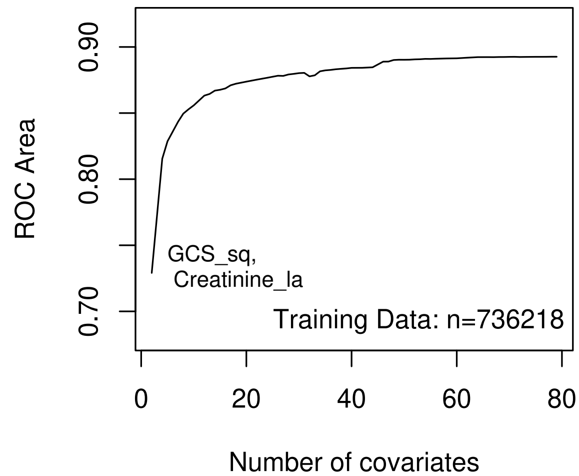


Figure 5-15: `RAS` model selection (all development data)

5.3.2 `RAS` Validation

The ROC curves for the `RAS` model on the validation data are provided in Figure 5-17. The H and C statistics for these performance evaluations are listed in Table 5.18. Details of the H statistic calculation using the mean prediction for each patient (second row in Table 5.18) are provided in Table 5.19.

Finally, calibration plots for `RAS`, using all predictions, the mean prediction, and the mean prediction from day 1, are provided in Figure 5-18 on page 105. The calibration plots for the mean probabilities on days 2, 3, 4, and 5 are provided in

Table 5.18: `RAS` Hosmer-Lemeshow calibration (validation data)

Day(s)	Summary Func.	H	p (*d.f.*)	C	p (*d.f.*)	n
all	none	2369	0 (10)	3284	0 (10)	218883
all	mean	13.2	0.067 (7)	17.3	0.027 (8)	2534
1	mean	22.4	0.004 (8)	14.3	0.073 (8)	2428
2	mean	19.2	0.008 (7)	29.0	3.22e-04 (8)	2083
3	mean	13.1	0.041 (6)	17.3	0.028 (8)	1378
4	mean	10.4	0.108 (6)	11.1	0.195 (8)	925
5	mean	8.63	0.196 (6)	11.6	0.116 (7)	652
6	mean	6.30	0.390 (6)	8.60	0.377 (8)	495
7	mean	23.3	7.07e-04 (6)	24.5	9.33e-04 (7)	401

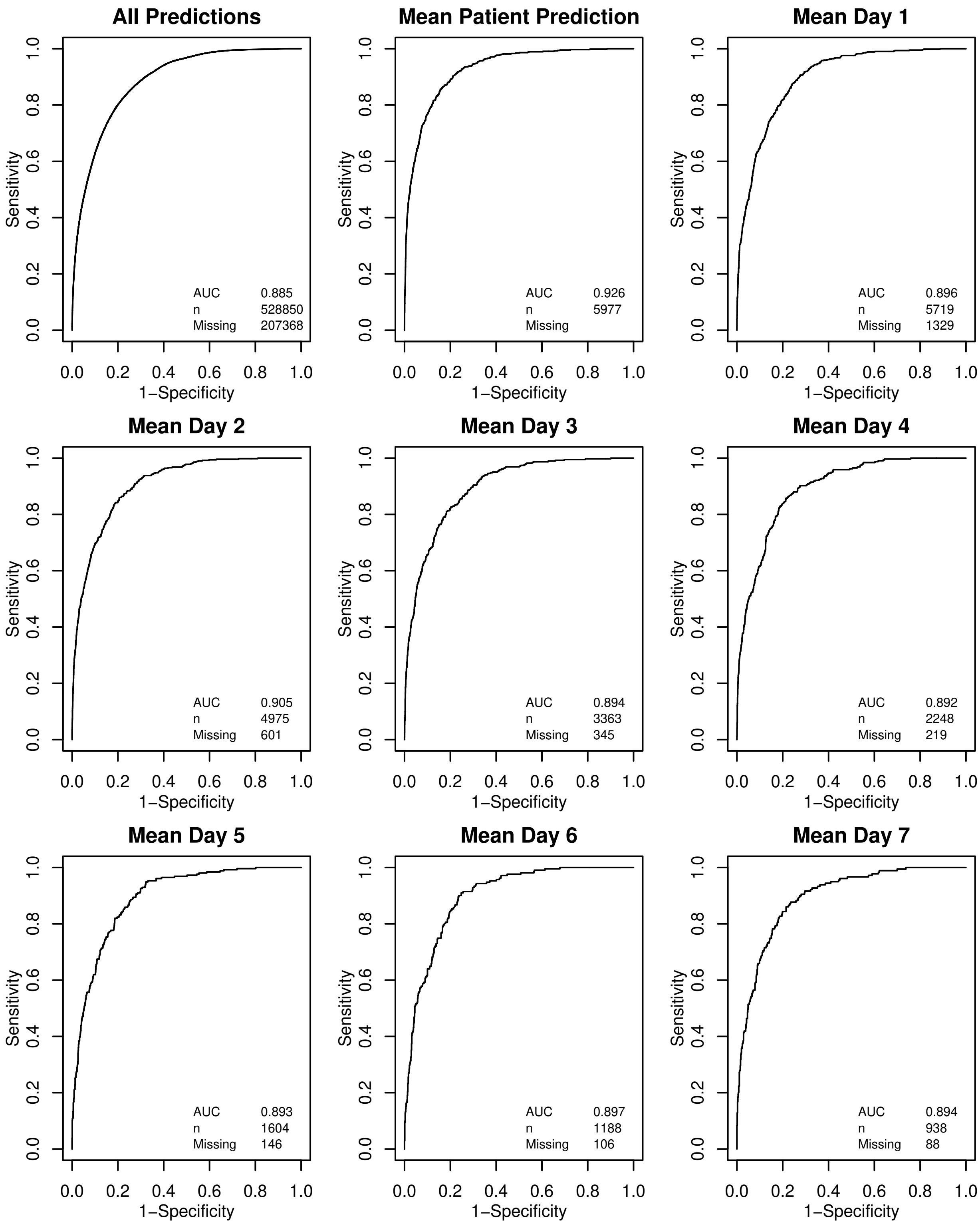


Figure 5-16: `RAS` ROC curves (development data)

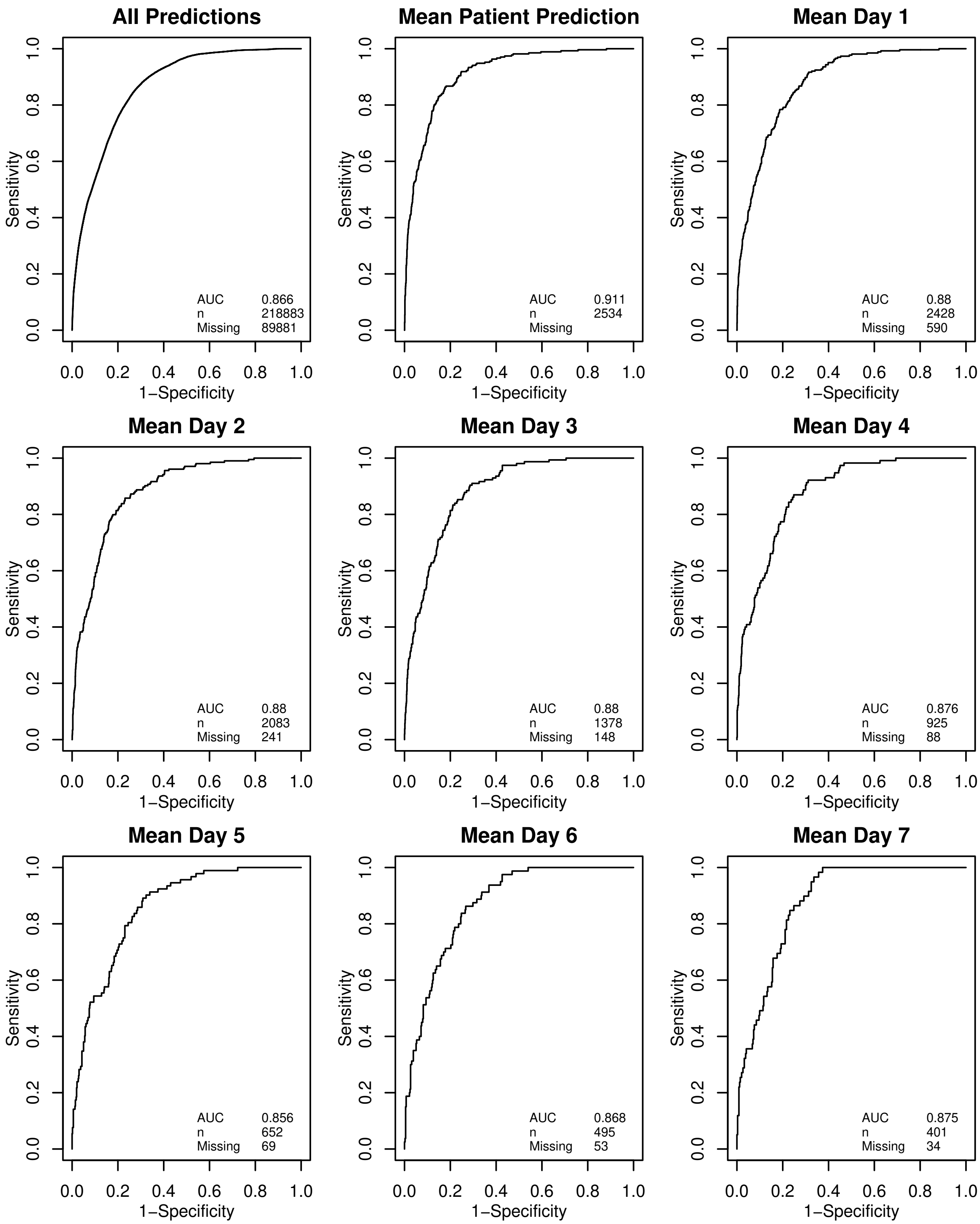


Figure 5-17: RAS ROC curves (validation data)

Table 5.19: `RAS` H statistic deciles of risk using mean prediction for each patient (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-4	[0.000493,0.01802)	0.008	4	7.7	1010	1006.3	1014
5	[0.018023,0.02904)	0.023	3	5.8	250	247.2	253
6	[0.029036,0.04700)	0.038	7	9.6	247	244.4	254
7	[0.046996,0.08465)	0.065	15	16.3	238	236.7	253
8	[0.084650,0.14648)	0.111	27	28.2	227	225.8	254
9	[0.146479,0.29485)	0.208	71	52.7	182	200.3	253
10	[0.294848,0.98298]	0.536	144	135.6	109	117.4	253

$\chi^2 = 13.22,\ d.f. = 7;\ p = 0.067$

Figures 5-19 and 5-20 on pages 106 and 107. As previously done with `SDAS` and `DAS`n, the predictions from `RAS` are uncorrected; one could expect slightly improved calibration performance if the corrected predictions, using the slope and intercept values found in Table 5.17, were used.

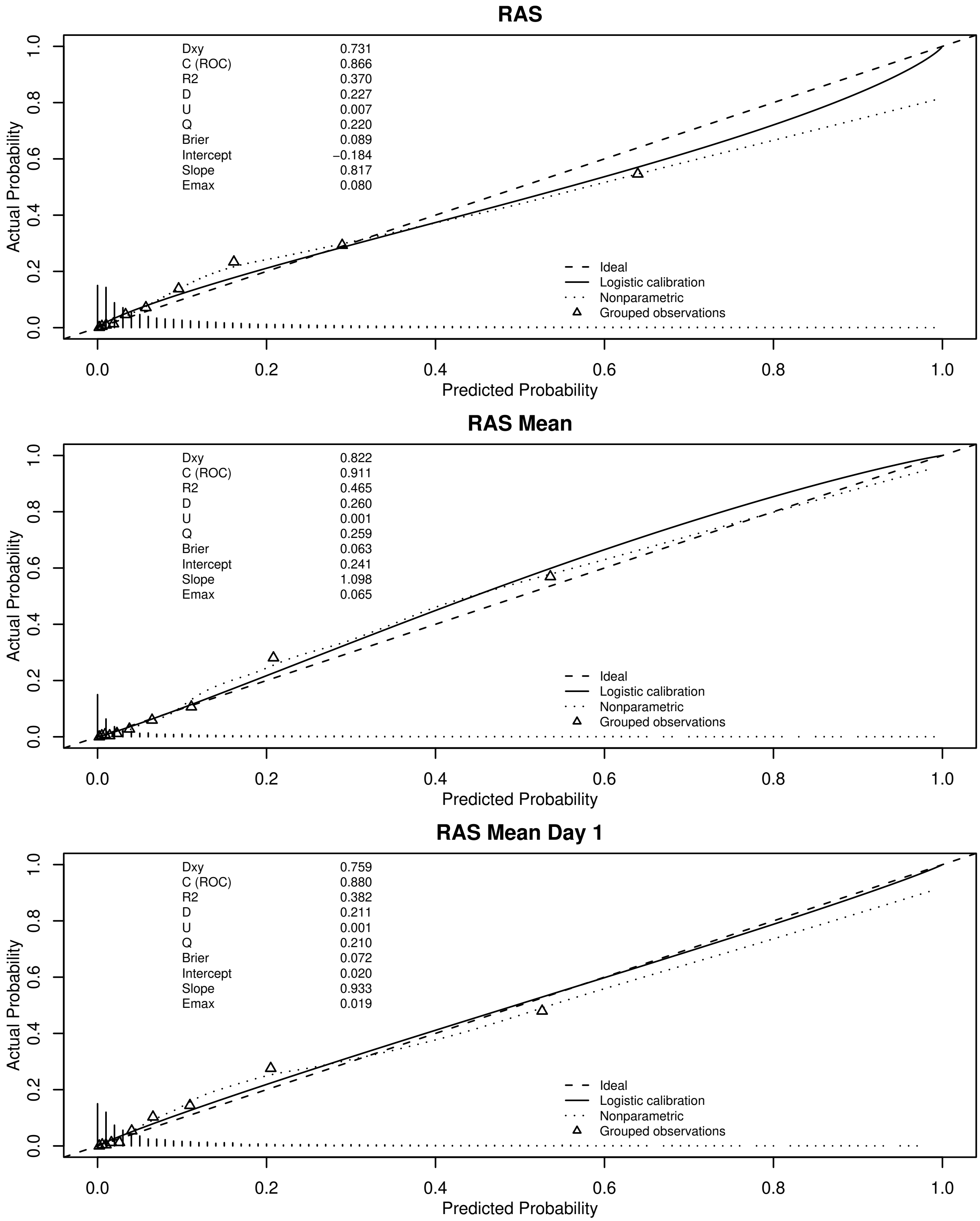


Figure 5-18: RAS calibration plots (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

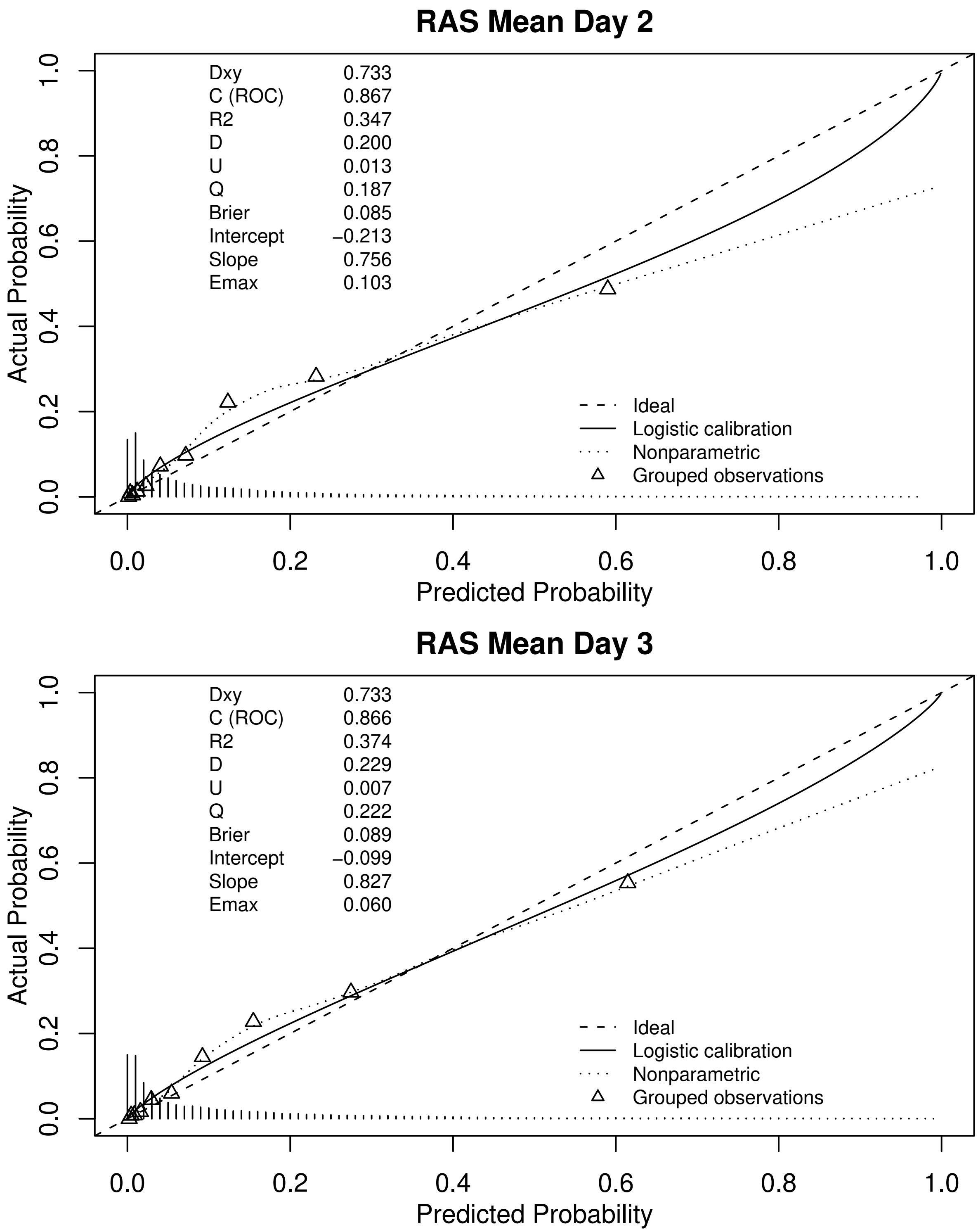
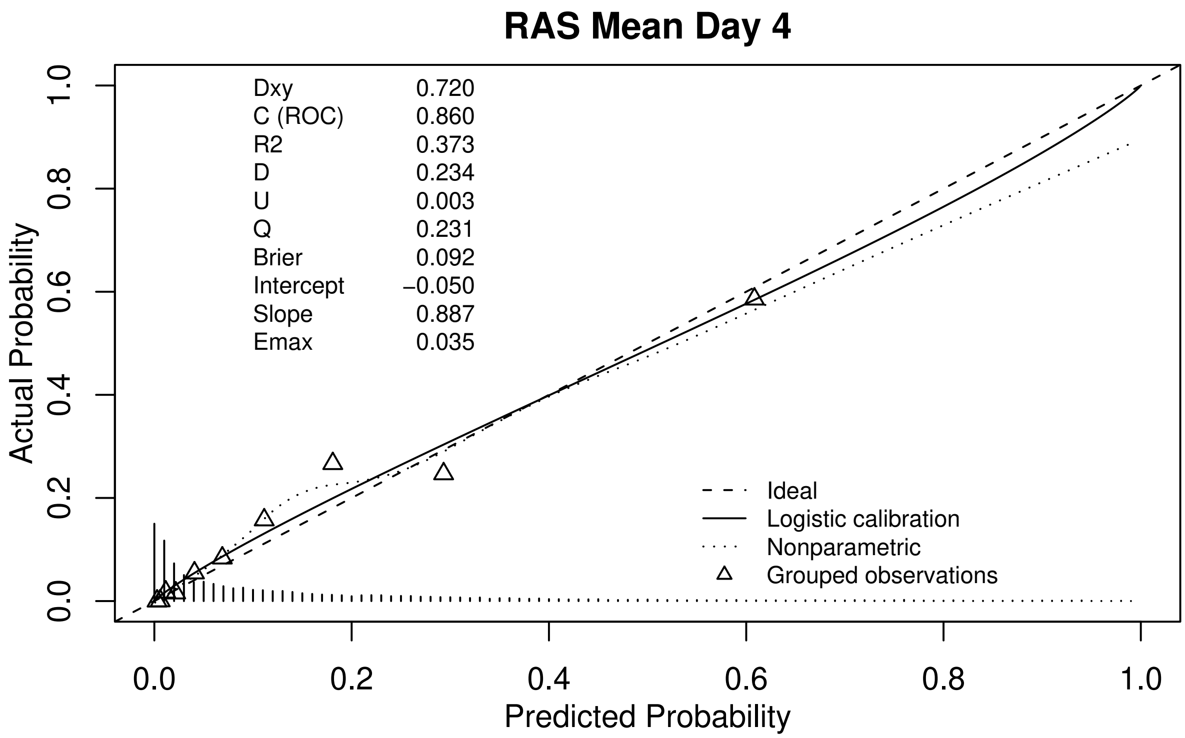


Figure 5-19: `RAS` calibration plots, days 2 and 3 (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.



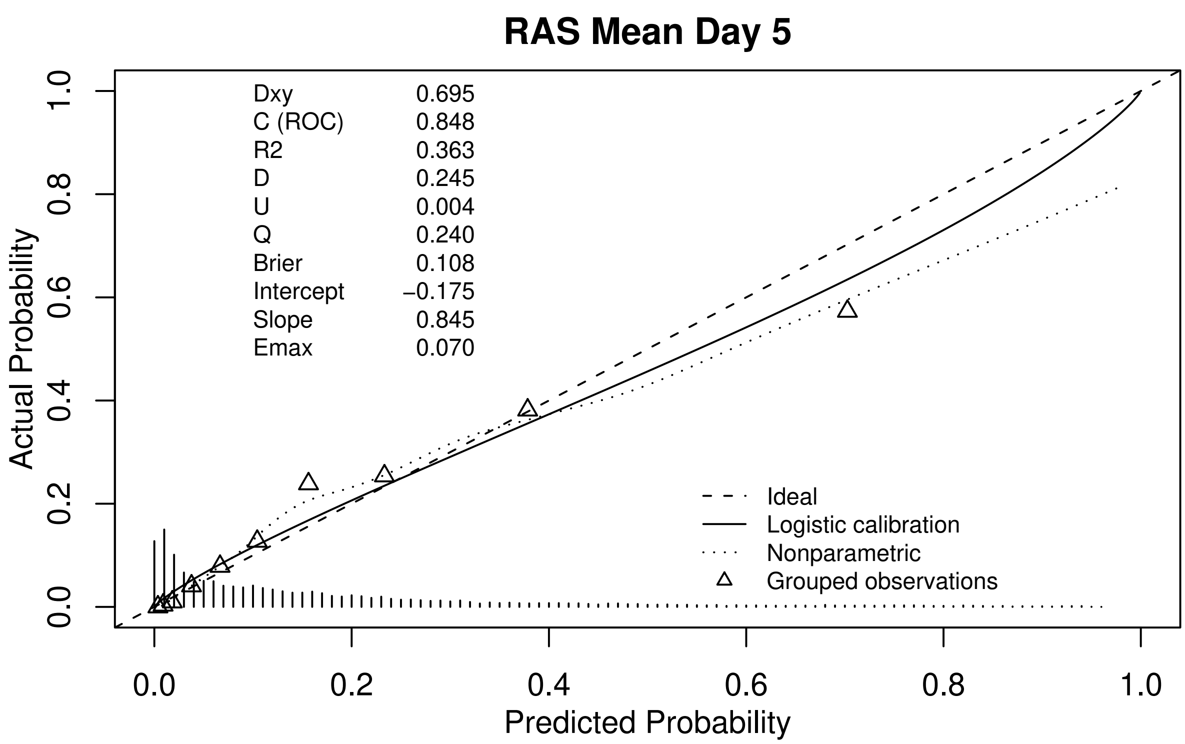


Figure 5-20: `RAS` calibration plots, days 4 and 5 (validation data). The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

5.4 SAPSII: Comparison Model

For comparison purposes, a customized SAPS II mortality prediction (described in Chapter 4) was used. We refer to our customized SAPS II score as $\texttt{SAPSII}_a$.

The $\texttt{SAPSII}_a$ model for each day can be described by looking at the $X\beta$ terms in the logistic regression equation,

$$P(Y=1) = \frac{1}{1+\exp(-X\beta)}.$$

For each day, $n \in \{1,2,3,4,5\}$, the $X\beta_n$ terms are given below:

$$\begin{aligned}
X\hat{\beta}_1 &= -4.331 + 0.0422\,\mathrm{SAPSII} + 0.301\ln(\mathrm{SAPSII}+1) - 2.714\,\mathrm{CSRU} \\
&\quad +0.204\,\mathrm{MICU} \\
X\hat{\beta}_2 &= -11.113 + 0.00627\,\mathrm{SAPSII} + 2.579\ln(\mathrm{SAPSII}+1) - 1.921\,\mathrm{CSRU} \\
&\quad +0.353\,\mathrm{MICU} \\
X\hat{\beta}_3 &= -9.808 + 0.0143\,\mathrm{SAPSII} + 2.203\ln(\mathrm{SAPSII}+1) - 2.051\,\mathrm{CSRU} \\
&\quad +0.241\,\mathrm{MICU} \\
X\hat{\beta}_4 &= -7.832 + 0.0379\,\mathrm{SAPSII} + 1.468\ln(\mathrm{SAPSII}+1) - 2.123\,\mathrm{CSRU} \\
&\quad +0.0657\,\mathrm{MICU} \\
X\hat{\beta}_5 &= -9.209 + 0.0201\,\mathrm{SAPSII} + 2.066\ln(\mathrm{SAPSII}+1) - 1.918\,\mathrm{CSRU} \\
&\quad +0.00737\,\mathrm{MICU}.
\end{aligned}$$

In contrast, the $X\beta$ terms first published by Le Gall et al. [44] were,

$$\mathrm{logit} = \mathrm{X}\beta = -7.7631 + 0.0737\,\mathrm{SAPSII} + 0.9971\ln(\mathrm{SAPSII}+1).$$

By looking at the equations, the `CSRU` and `MICU` variables seem to have captured the *type of admission* variable that they were intended to capture. The type of admission field in SAPS II has three different values. These values, along with the expected proxy equivalents, are as follows: (1) 0 points for Scheduled Surgery (`CSRU`=1), (2) 6 points for Medical Admission (`MICU`=1), and (3) 7 points for Unscheduled Surgery (`CSRU`=0 and `MICU`=0). When `CSRU`=1, the logit is reduced considerably which is consistent with the scheduled surgery input. On the other hand, when `MICU`=1, the risk is increased by a small amount. This is not expected, and indicates that in the case where both `CSRU` and `MICU` are false (0), the patient falls in a risk bin that is not quite as severe as might be explained by the unscheduled surgery contribution. At the same time, however, the relative influence of the `CSRU` variable is much higher than the type of admission variable in SAPS II. In my customized version of SAPS II, the `CSRU` variable is the most important feature observed with a Wald Z score that

Table 5.20: SAPSII$_a$ calibration statistics

	Development Data					Validation Data				
Day	H	p (*d.f.*)	C	p (*d.f.*)	n	H	p (*d.f.*)	C	p (*d.f.*)	n
1	11.2	0.345 (10)	8.01	0.432 (8)	6008	15.2	0.085 (9)	8.04	0.329 (7)	2598
2	3.88	0.868 (8)	4.02	0.855 (8)	5247	9.61	0.212 (7)	6.06	0.532 (7)	2207
3	5.18	0.627 (8)	0.513	0.999 (7)	3512	5.98	0.542 (7)	2.53	0.925 (7)	1440
4	3.90	0.866 (8)	6.79	0.560 (8)	2321	3.53	0.831 (7)	7.49	0.278 (6)	941
5	7.40	0.494 (8)	7.17	0.518 (8)	1620	7.21	0.408 (7)	8.07	0.326 (7)	668

Table 5.21: `SAPSII`$_a$ bootstrapped goodness of fit statistics, day 1 (dev data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.593	0.592	0.592	0.001	0.592	150
R^2	0.238	0.239	0.237	0.002	0.237	150
Intercept	0.000	0.000	-0.005	0.005	-0.005	150
Slope	1.000	1.000	0.997	0.003	0.997	150
E_{max}	0.000	0.000	0.001	0.001	0.001	150
D	0.133	0.134	0.133	0.001	0.132	150
U	0.000	0.000	0.000	0.000	0.000	150
Q	0.134	0.134	0.133	0.001	0.132	150
B	0.090	0.090	0.091	0.000	0.091	150

is large for all five models (more than twice as large as the contribution from SAPSII or ln(SAPSII+1)). The Wald Z score for the `MICU` variable is much less significant. The contribution from `MICU` is strongest on day 2 where its Wald Z score ($Z = 3.25$) is not far behind the better of the two SAPSII inputs, ln(SAPSII+1) ($Z = 4.85$). Considering the improvement offered by this customization, the customized SAPS II that I used for my comparison is expected to perform favorably to the original SAPS II with the type of admission variable, but we are unable to validate this expectation with our available data.

Figure 5-21 shows the ROC curves for the five `SAPSII`$_a$ models. The figure shows the performance of `SAPSII`$_a$ on both the development data and the validation data. The development data was not used to define the weights of the individual SAPS II components but it was used to fit the weights for each logistic regression equation. The cases where a `SAPSII`$_a$ prediction was unavailable came from episodes where a patient was mechanically ventilated but no value for `PaO2:FiO2` could be found. The Hosmer-Lemeshow calibration statistics for `SAPSII`$_a$ are listed in Table 5.20.

As done with my previous models, bootstrapping with 150 samples (with replacement) was performed on the development data to validate the goodness of fit for the `SAPSII`$_a$ models. Tables 5.21, 5.22, 5.23, 5.24, and 5.25 show these results for days 1 through 5.

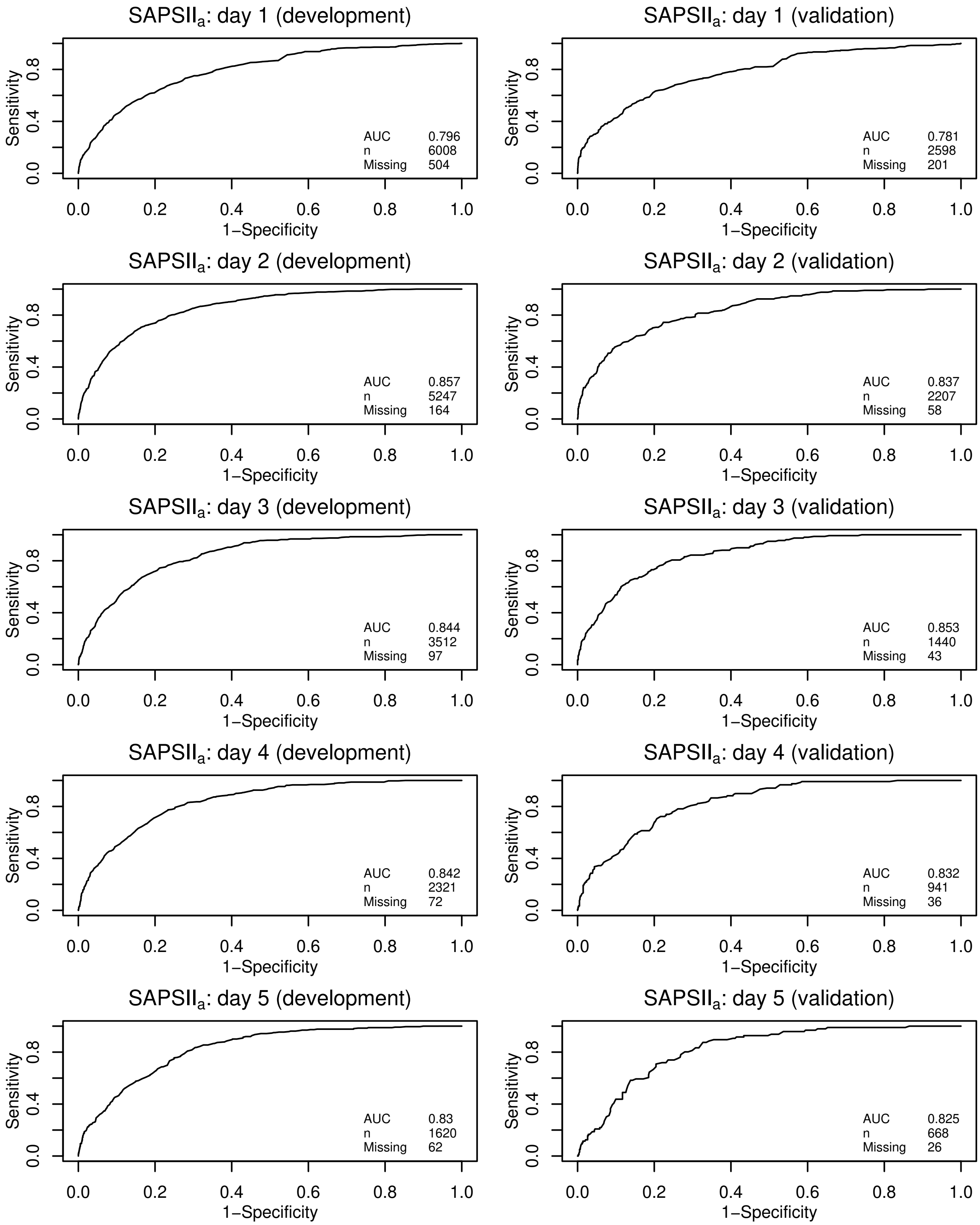


Figure 5-21: $\texttt{SAPSII}_a$ ROC curves

Table 5.22: $\texttt{SAPSII}_a$ bootstrapped goodness of fit statistics, day 2 (dev data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.714	0.714	0.713	0.001	0.713	150
R^2	0.328	0.329	0.327	0.002	0.327	150
Intercept	0.000	0.000	-0.012	0.012	-0.012	150
Slope	1.000	1.000	0.996	0.004	0.996	150
E_{max}	0.000	0.000	0.003	0.003	0.003	150
D	0.171	0.172	0.171	0.001	0.170	150
U	0.000	0.000	0.000	0.000	0.000	150
Q	0.172	0.172	0.171	0.001	0.170	150
B	0.072	0.071	0.072	0.000	0.072	150

Table 5.23: $\texttt{SAPSII}_a$ bootstrapped goodness of fit statistics, day 3 (dev data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.689	0.693	0.689	0.005	0.684	150
R^2	0.313	0.317	0.311	0.006	0.307	150
Intercept	0.000	0.000	-0.024	0.024	-0.024	150
Slope	1.000	1.000	0.984	0.016	0.984	150
E_{max}	0.000	0.000	0.008	0.008	0.008	150
D	0.174	0.177	0.173	0.004	0.171	150
U	-0.001	-0.001	0.000	-0.001	0.000	150
Q	0.175	0.178	0.173	0.004	0.170	150
B	0.082	0.082	0.083	-0.001	0.083	150

Table 5.24: $\texttt{SAPSII}_a$ bootstrapped goodness of fit statistics, day 4 (dev data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.683	0.686	0.682	0.004	0.679	150
R^2	0.326	0.329	0.324	0.006	0.320	150
Intercept	0.000	0.000	-0.010	0.010	-0.010	150
Slope	1.000	1.000	0.985	0.015	0.985	150
E_{max}	0.000	0.000	0.005	0.005	0.005	150
D	0.198	0.201	0.197	0.003	0.195	150
U	-0.001	-0.001	0.000	-0.001	0.000	150
Q	0.199	0.201	0.197	0.004	0.195	150
B	0.094	0.093	0.094	-0.001	0.095	150

Table 5.25: $\texttt{SAPSII}_a$ bootstrapped goodness of fit statistics, day 5 (dev data)

Index	Original Index	Training Sample	Test Sample	Optimism	Corrected Index	Samples
D_{xy}	0.660	0.665	0.659	0.006	0.654	150
R^2	0.321	0.327	0.318	0.009	0.312	150
Intercept	0.000	0.000	-0.032	0.032	-0.032	150
Slope	1.000	1.000	0.977	0.023	0.977	150
E_{max}	0.000	0.000	0.011	0.011	0.011	150
D	0.208	0.213	0.206	0.007	0.202	150
U	-0.001	-0.001	0.000	-0.001	0.000	150
Q	0.210	0.214	0.206	0.008	0.201	150
B	0.107	0.106	0.107	-0.001	0.108	150

The calibration plots for $\texttt{SAPSII}_a$ on the validation data (days 1 through 5) are shown in Figures 5-22 and 5-23. The uncorrected predictions from my models were used in these plots.

Finally, the performance of the SAPS II score without the `CSRU` and `MICU` additions was evaluated. With this configuration, the performance was considerably worse. For example, day 1 validation performance dropped from AUC=0.781 to AUC=0.661. The calibration on day 1 without the service-type inputs also suffered precipitously, dropping from $p = 0.085$ (*d.f.*=9) to $p = 0.004$ (*d.f.*=10). On day 3, where the two service-type inputs had much less relative importance, the performance decrease was less, with the AUC dropping from AUC=0.853 to AUC=0.819. The calibration for day 3 changed from $p = 0.542$ (*d.f.*=7) in the original model to $p = 0.146$ (*d.f.*=8) without the service-type inputs.

To put these results in context, a model trained with *only* `CSRU` and `MICU` as covariates resulted in an AUC of 0.673 on the validation data for day 1. Using 3 risk bins (only three unique probability outputs are available with the two binary inputs because a patient does not receive the `CSRU` and `MICU` services simultaneously), the calibration of the model was also good, with $p = 0.740$ (*d.f.*=3). In essence, the performance of SAPS II (without type of admission) on day 1 was inferior to a model that uses only the `CSRU` and `MICU` service inputs.

5.5 Direct Model Comparisons

A number of direct comparisons between the models described above were performed. First, we compared the models on each of the first 5 days in the ICU, with a decreasing patient count on each day as some patients leave or expire. This was done by only including patients with valid predictions available from each model for a specific day. A second type of comparison involved looking at the performance of the models

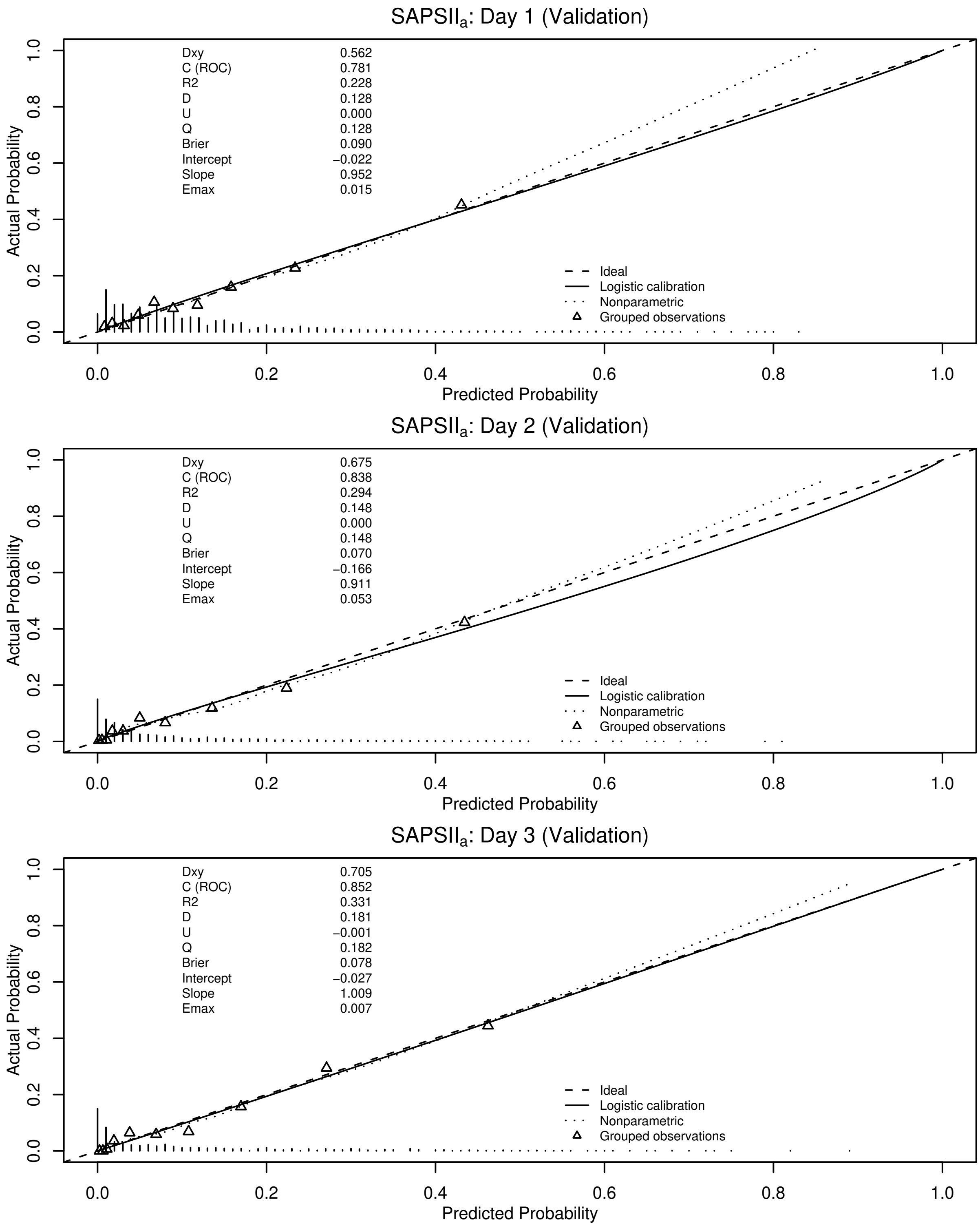
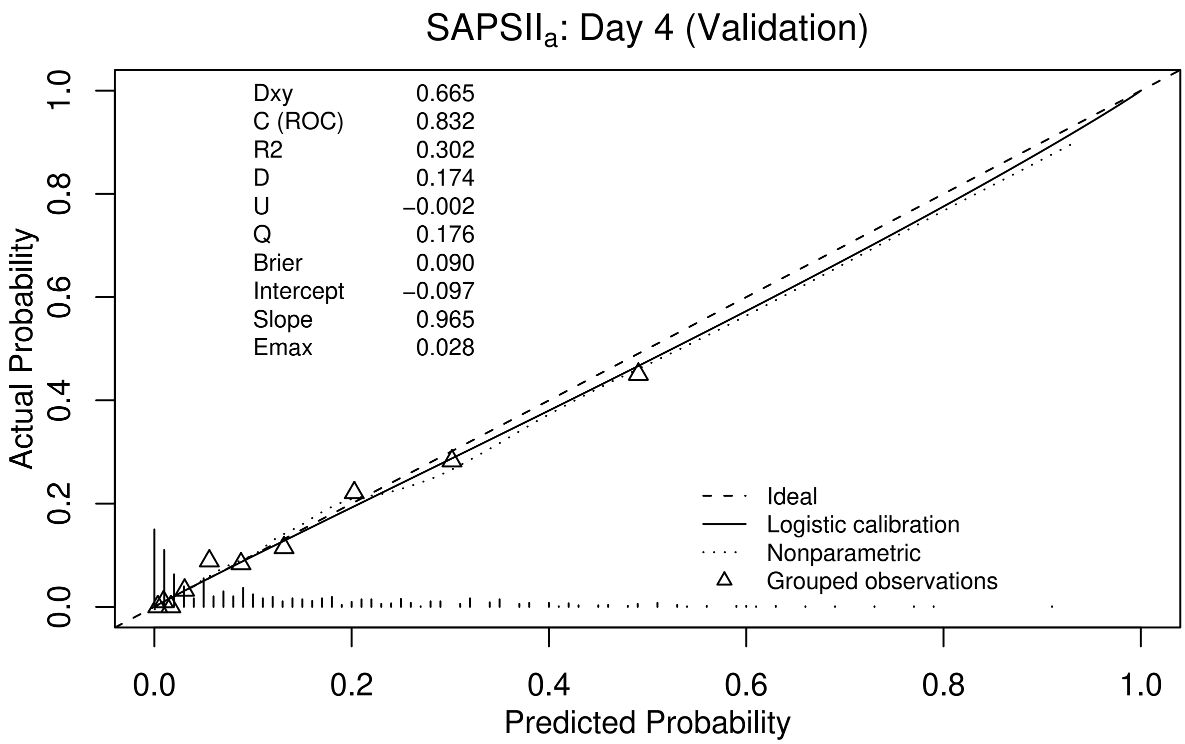


Figure 5-22: `SAPSII`$_a$ calibration plots, days 1 through 3. The relative frequencies for each predicted probability are indicated by the bars along the x-axis.



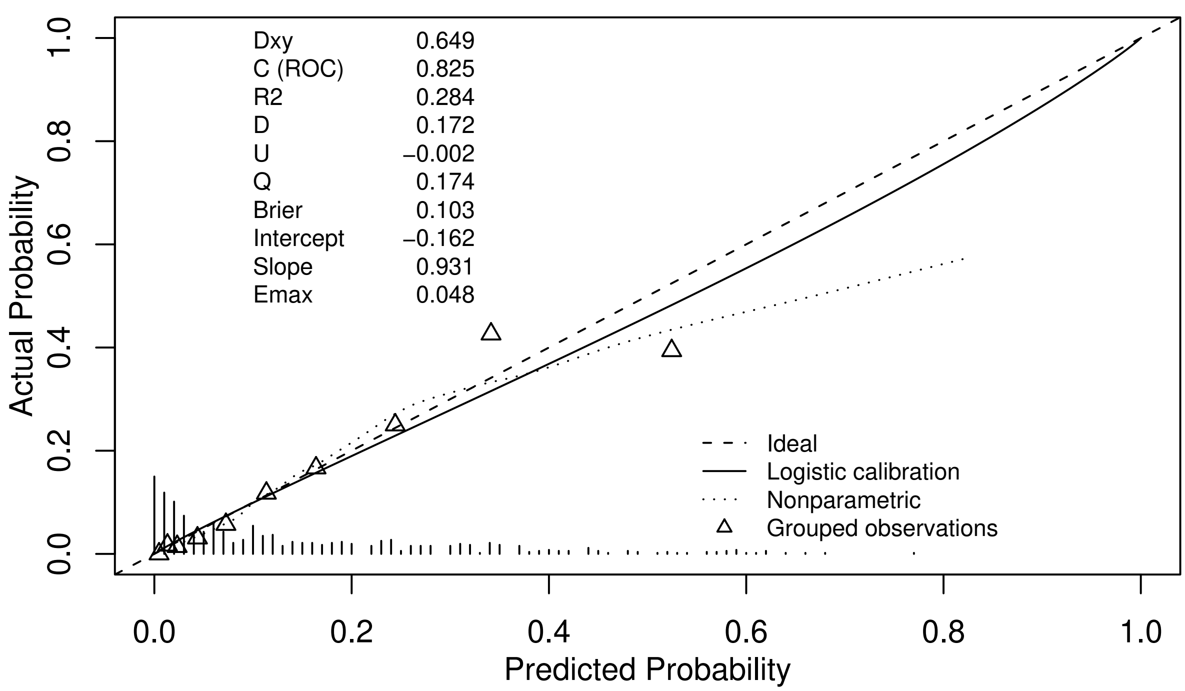


Figure 5-23: `SAPSII`$_a$ calibration plots, days 4 and 5. The relative frequencies for each predicted probability are indicated by the bars along the x-axis.

Table 5.26: SAPS II customization AUC performance (validation data). *no type of admission information. **only three different risk predictions were available using the two binary variables

	$\texttt{SAPSII}_a$		SAPS II*		CSRU and MICU	
Day	AUC	H	AUC	H	AUC	H^{**}
1	0.781	0.085 (9)	0.661	0.004 (10)	0.673	0.740 (3)
2	0.837	0.212 (7)	0.807	0.408 (8)	0.687	0.570 (3)
3	0.853	0.542 (7)	0.819	0.146 (8)	0.703	0.669 (3)
4	0.832	0.831 (7)	0.785	0.341 (8)	0.707	0.427 (3)
5	0.825	0.408 (7)	0.775	0.515 (8)	0.677	0.615 (3)

on patients that stayed in the ICU at least five days and had predictions available from each model for *every* day. For the first comparison, the AUC and calibration statistics were examined using the development data and the validation data. For the second comparison, I only report the AUC performance over time for the two datasets without calibration information.[3] In order to better understand the differences in AUC between models, I also provide a number of significance values derived using DeLong's method for comparing AUC areas [42].

The number of predictions available for the validation data using each model over the first five ICU days is provided in Table 5.27. This table also indicates the size of the intersection between all of these patients.

Table 5.27: Model coverage (i.e., number of patients with predictions) on validation data

Day	SDAS	DASn	RAS	$\texttt{SAPSII}_a$	Intersection
1	2434	2710	2428	2598	1954
2	2108	2177	2083	2207	1849
3	1389	1427	1378	1440	1245
4	928	976	925	941	836
5	662	696	652	668	596

Using the development data, Figure 5-24 shows the AUC for the four model types over the first five ICU days. The Hosmer-Lemeshow H statistics for these daily comparisons are provided in Table 5.28. Similarly, using the validation data, Figure 5-25 shows the AUC for the four model types over the first five ICU days. The Hosmer-Lemeshow H statistics for these daily comparisons are provided in Table 5.29.

Finally, Figures 5-26 and 5-26 plot the AUC for the first 5 ICU days when the validation and development patients were limited to patients who stay in the ICU for at least 5 days.

[3]The Hosmer-Lemeshow test was difficult to calculate for this smaller set of patients. In order to get reasonable expected frequencies, it would have been necessary to collapse most of the deciles.

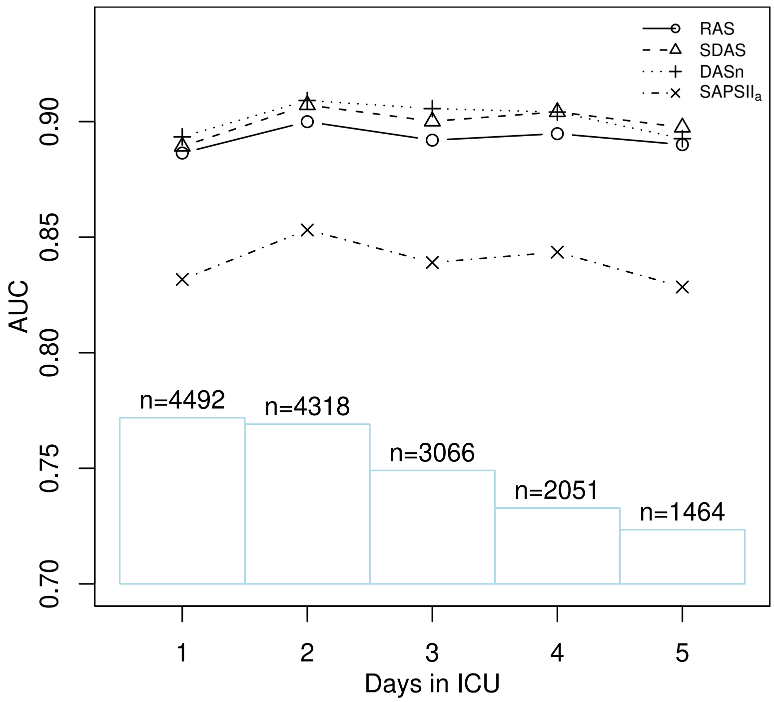


Figure 5-24: AUC versus day, first 5 ICU days (development data)

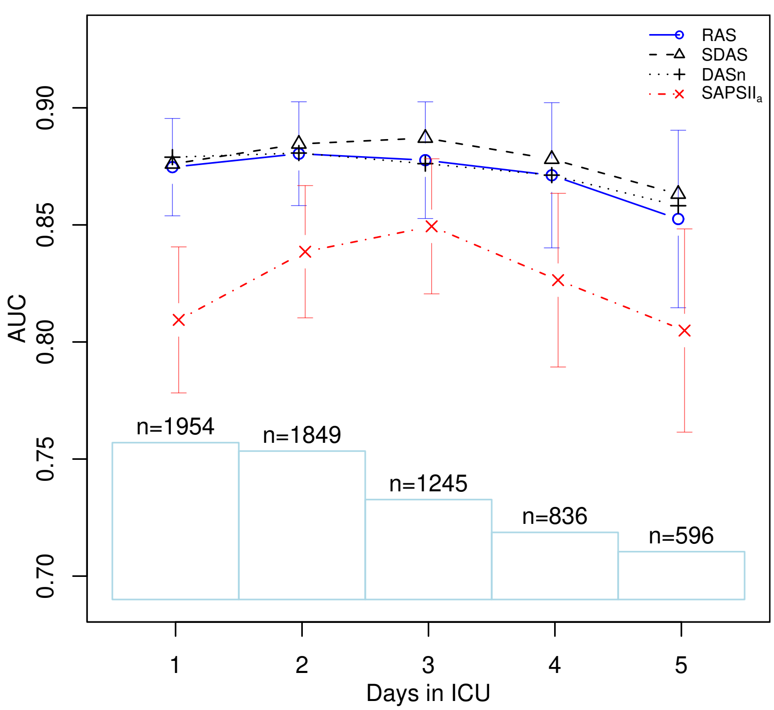


Figure 5-25: AUC versus day, first 5 ICU days (validation data). The 95% confidence intervals are shown for the `RAS` and `SAPSII`$_a$ performances.

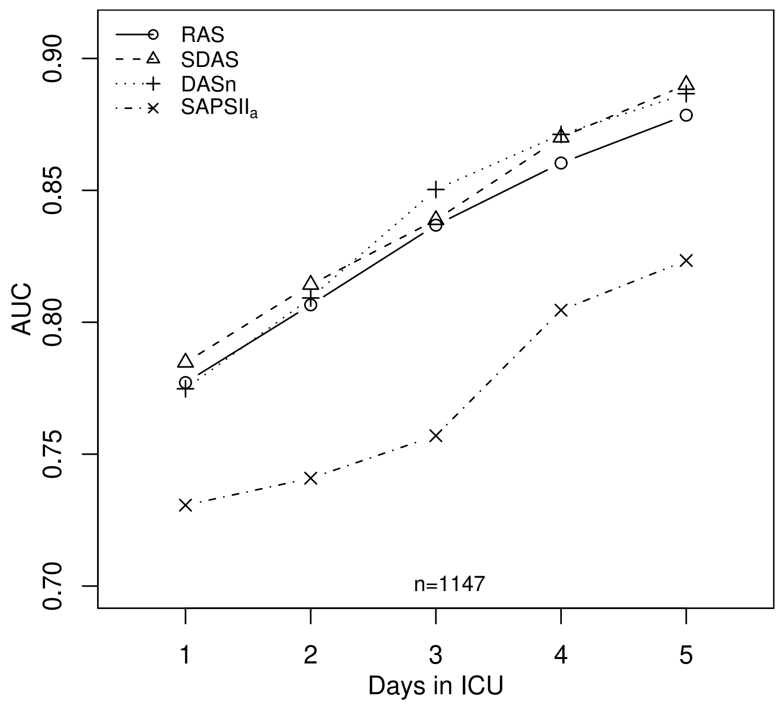


Figure 5-26: AUC versus day, patients with ICU stays $\geq$ 5 days (development data)

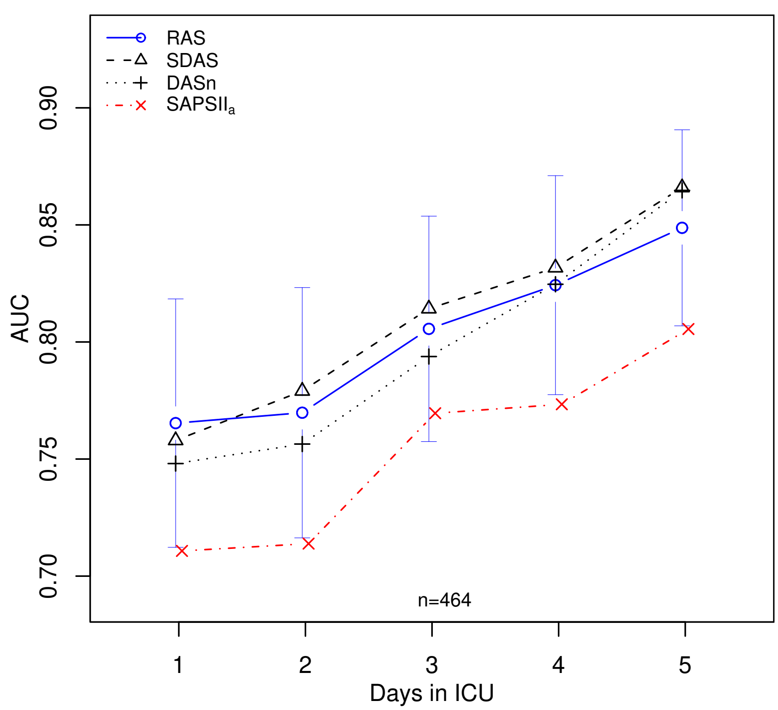


Figure 5-27: AUC versus day, patients with ICU stays $\geq$ 5 days (validation data). The 95% confidence intervals are shown for the `RAS` performances.

Table 5.28: Calibration statistics for daily model comparisons (development data)

	RAS		SDAS		DASn		SAPSII$_a$		
Day	H	p (*d.f.*)	H	p (*d.f.*)	H	p (*d.f.*)	H	p (*d.f.*)	n
1	13.9	0.031 (6)	13.6	0.035 (6)	8.31	0.216 (6)	34.6	0.000 (9)	4492
2	5.00	0.544 (6)	5.36	0.498 (6)	7.05	0.317 (6)	4.04	0.854 (8)	4318
3	4.29	0.637 (6)	5.59	0.471 (6)	8.34	0.139 (5)	5.31	0.724 (8)	3066
4	20.3	0.002 (6)	7.61	0.179 (5)	7.29	0.200 (5)	7.17	0.724 (8)	2051
5	9.64	0.141 (6)	6.23	0.285 (5)	2.20	0.821 (5)	6.33	0.610 (8)	1464

Table 5.29: Calibration statistics for daily model comparisons (validation data)

	RAS		SDAS		DASn		SAPSII$_a$		
Day	H	p (*d.f.*)	H	p (*d.f.*)	H	p (*d.f.*)	H	p (*d.f.*)	n
1	22.4	0.002 (7)	14.4	0.045 (7)	24.8	0.001 (7)	13.8	0.130 (9)	1954
2	17.9	0.006 (6)	15.3	0.018 (6)	17.5	0.008 (6)	6.90	0.439 (7)	1849
3	12.0	0.063 (6)	5.01	0.542 (6)	11.0	0.087 (6)	4.13	0.765 (7)	1245
4	10.7	0.097 (6)	10.4	0.110 (6)	11.9	0.064 (6)	3.01	0.884 (7)	836
5	9.58	0.143 (6)	7.55	0.183 (5)	6.65	0.248 (5)	8.70	0.275 (7)	596

Using DeLong's method, I made a number of comparisons between the AUC values in Figures 5-24, 5-25, 5-26, and 5-27. These comparisons focused on the relative performance of RAS, so I compared the RAS AUC performance against the SAPSII$_a$ AUC performance and against the best performing daily model (SDAS or DASn). The results are summarized in Tables 5.30 and 5.31.

Table 5.30: RAS DeLong AUC significance tests (days 1 through 5)

	Development Data		Validation Data	
Day	p (vs SAPSII$_a$)	p (vs best model)	p (vs SAPSII$_a$)	p (vs best model)
1	0.0000	0.2899	0.0000	0.6888
2	0.0000	0.1505	0.0005	0.6986
3	0.0000	0.0788	0.0437	0.4692
4	0.0000	0.3216	0.0128	0.6761
5	0.0000	0.5122	0.0312	0.6032

5.6 Discussion

This chapter presented three types of mortality models based on the methodology described in Chapter 4. For comparison, a customized SAPS II model was also presented. While the differences between my models were generally small, there were

Table 5.31: RAS DeLong AUC significance tests (5+ day patients)

	Development Data		Validation Data	
Day	p (vs $\texttt{SAPSII}_a$)	p (vs best model)	p (vs $\texttt{SAPSII}_a$)	p (vs best model)
1	0.0102	0.6596	0.0654	0.8988
2	0.0002	0.6411	0.0565	0.7384
3	0.0000	0.3754	0.1857	0.7405
4	0.0005	0.4480	0.0572	0.7661
5	0.0003	0.3910	0.0860	0.4582

a number of noteworthy attributes for each individual model type that I will discuss in order.

5.6.1 SDAS

Many of the 35 variables ultimately included in the SDAS model were expected. There were, however, a number of interesting exceptions. In addition to the important features included in SAPS II (e.g., GCS, Age, PaO2:FiO2, etc), a number of new features were included, such as INR (tied for the second largest absolute Wald Z score) and BUN:Creatinine (forth largest absolute Wald Z score). Clinically, the INR variable serves as an indicator of blood coagulation and it is important in assessing liver functionality, atrial fibrillation, and stroke risk. The BUN-to-Creatinine ratio helps indicate pre-renal failure, dehydration, and gastrointestinal bleeding. A number of more computationally intensive variables were included by our variable selection method for SDAS. The most important of these was the SpO2.oor30.t variable that tracks the amount of time that SpO2 was out of range ($< 90\%$) within the past 30 minutes. Another important derived variable was the Platelets_Slope_28hr_min. An increased level of Platelets was found to decrease risk of death. At first it seemed that the inclusion of the Platelets features might be a result of confounding with the INR variable (platelets increase blood coagulation), but manual removal of INR resulted in a slight *increase* in significance for both Platelets_mean_i and Platelets_Slope_28hr_min.

It was also interesting to note the absence of several simple variables such as Heart Rate (HR) and Systolic BP (SBP). In addition to the individual variability of these variables, we have shown previously that they are generally undersampled and sometimes miss important episodes [32, 33]. In their place, three variables were included that summarized HR and SBP: (1) HR range (HRrng_max), (2) ShockIndx (HR divided by SBP), and (3) the ratio of mean blood pressure on pressors to mean blood pressure off pressors (MBPm.pr_min_am). Each of the HR and SBP summary variables, however, still had a minor role in the SDAS model.

Looking at the performance on the development data, the SDAS model had a strong fit. The AUC was 0.898 for all days and 0.890 for day 1. The H statistic for the development data was not strong, but from looking at the calibration statistics from in Table 5.6, the weakness appears to be from the choice of cut-points for the Hosmer-Lemeshow statistic. The statistics in Table 5.6 show that the overfitting was minimal and the corrected slope between predicted probability and actual probability was 0.986 (with a corrected intercept of -0.017) and the maximum calibration error, E_{max}, had a low value of 0.006.

On the validation data, the `SDAS` AUC was 0.876 for all days and 0.870 for day 1. The calibration, however, was weaker with an $H = 51.1$ (*d.f.*=9) and an $E_{max} = 0.094$ (all days), and an $H = 14.4$ (*d.f.*=7) and an $E_{max} = 0.046$ looking only at day 1. As Table 5.14 and 5-9 show, most all of this weakness was from the tenth decile, with a large probability range between 0.3729 and 0.9986. In this decile, the expected number of deaths was 535 while only 456 deaths were observed. The other deciles demonstrated reasonable calibration.

5.6.2 `DAS`n

The daily acuity model, `DAS`n, explored using a different logistic model for each ICU day. While the specific set of variables used for the model between days varied, the significant inputs remained similar with consistent importance placed on `GCS`, `INR`, `Age`, `BUN:Creatinine` and `SpO2.oor30.t`.

Some important variables only appeared in one or two of the models. `Amiodarone`, for example, only appeared in `DAS2` (but has the third largest Z score). This also provided an example of the stratifying effect that many interventions have: while amiodarone was being given to help patients, the act of administering it indicated that the patient was having heart arrhythmias (and was therefore at a higher risk for mortality). In this case `Amiodarone` was an independent predictor of mortality from the ordinal ventricular arrhythmia variable (`hrmVA`) which was included in `DAS1` and `DAS2`.

The performance of `DAS`n on the development data was quite strong. The AUC was ≥ 0.90 for the first three ICU days and only slightly less for days 4 and 5. The calibration was also consistently strong as shown by the H statistics in Table 5.7 and the bootstrapped statistics in Tables 5.8, 5.9, 5.10, 5.11, and 5.12.

On the validation data, the `DAS`n model did not perform as well. The AUC performance was reasonable with AUC ≥ 0.870 for the first four days, and AUC=0.864 for day 5. The calibration, however, was generally poor with the possible exception of `DAS5` ($H = 6.65$ with *d.f.*=5 and $E_{max} = 0.078$). Day 2 (`DAS2`), in contrast, had an E_{max} of 0.120 (see Figure 5-10). The weaker performance by `DAS`n indicates that whatever benefit a daily model had in identifying patient risks that change from day to day was likely lost from the limitation of training on a subset of the data and the overfitting of the model that resulted.

5.6.3 RAS

The real-time model, by looking at each unique observation, had a wealth of data to use for training and validation. The variables that were included in the final `RAS` model were mostly seen previously in the daily models discussed above. There were, however, some notable differences. For example, `Creatinine` was included in `RAS` while it did not appear in any of the previous models. Similarly, the `RAS` model also included medications/medication-categories previously unseen such as `Sandostatin`[4] (used to treat acromegaly and endocrine tumors) and `Nondepolarizing_agent` (neuromuscular blocking drugs used to cause paralysis).

On the development data, the performance of the `RAS` model was consistently in the AUC=0.88 to AUC=0.91 range. If the model was evaluated using every observation in the development data, the AUC was 0.885. This represented strong performance considering many of these observations relied on minimal prior information, e.g., an observation at the beginning of ICU day 1 had limited information for many of the evolving features. The daily mortality prediction problem had a more thorough approach to summarizing the past 24 hours of each variable, which should result in an easier prediction problem. If predictions were summarized for individual days, the `RAS` model did quite well with AUCs between 0.905 (day 2) and 0.892 (day 4). The easiest prediction task looked at the mean prediction for each patient's stay up through day 7 (the limit placed on our dataset). This resulted in an AUC of 0.926.

Using the Hosmer-Lemeshow statistic, calibration for the RAS model was poor when all observations were used but reasonable when daily mean predictions were employed. As was the case in interpreting the significance of individual covariates in the `RAS` model, however, the calibration significance according to the Hosmer-Lemeshow H statistic was likely misleading due to increase in observations (about a 2 orders of magnitude increase) between the real-time model and the daily model. The calibrations using the *mean* prediction for ICU day 2 or ICU day 3 were in fact strong ($H = 4.42$ and $H = 6.60$, respectively). The weakest calibration performance using daily mean predictions was on ICU day 7, with $H = 17.6$ (*d.f.*=6) and $C = 16.2$ (*d.f.*=6).

In fact, the bootstrapped statistics in Table 5.17 showed that the optimism was quite low for the `RAS` model. Looking at only four significant digits, the correction was minimal with a corrected intercept of -0.0004 and corrected slope of 0.9998. The small optimism estimates can likely be explained by the fact that, on average, each patient had over 100 unique observations present in the training data. Consequently, each sampling of the data likely had representation from each patient and therefore underestimated the bootstrapped optimism.

When the performance on the validation data was examined, the `RAS` did well in terms of discrimination, with an AUC of 0.866 for all observations, and an AUC of 0.880 using mean predictions for day 1. On most days, the mean prediction AUC was

[4]Sandostatin is a brand name for octreotide.

about 0.88, with a low of AUC=0.856 on day 5.

The calibration for `RAS` on the validation data was weaker. The Hosmer-Lemeshow tests, with $H = 2369$ and $C = 3284$ (see Table 5.18), indicated highly significant deviation from calibration. This was somewhat supported by the calibration plot in Figure 5-18, where the logistic calibration curve's slope is 0.817 with an intercept of -0.184 and $E_{max} = 0.080$. In comparison to `SDAS` (*slope*=0.820, *intercept*= -0.257 and $E_{max} = 0.094$) and `DAS1` (*slope*=0.866, *intercept*= -0.132 and $E_{max} = 0.066$), however, the overall calibration performance was quite similar (albeit imperfect). When a daily mean prediction for each patient was used, the calibration performance from the Hosmer-Lemeshow statistics and the calibration plots looked much better (in general). Furthermore, based on the significant calibration corrections suggested by Figure 5-18, the bootstrapped estimate of optimism for `RAS` using the development data was clearly inadequate.

5.6.4 `SAPS II`

The customized SAPS II score described above performed moderately in terms of discrimination. While worse than all of my models trained entirely from this patient population, it did manage to obtain an AUC of 0.853 on the validation data for day 3 and an AUC of 0.857 on the development data for day 2. The day 1 performance for both development data and validation data, however, was surprisingly poor with an AUC less than 0.80.

The weight given to the type of service inputs for my customized SAPS II score corroborated my claim that $\texttt{SAPSII}_a$ provides a reasonable representation of SAPS II for comparison purposes. The contribution from the SAPS II score, however, still remained rather weak. A `SDAS` model trained using the development data and only `GCS_max_sq`, `CSRU`, and `MICU` obtained an AUC of 0.782 on day 1 validation data, which was close to the AUC=0.781 that resulted from the more complicated (but less customized) $\texttt{SAPSII}_a$ on day 1.

In contrast to its discrimination performance, the calibration of $\texttt{SAPSII}_a$ was strong. Over the first 5 ICU days, the only case where either the H or C statistic fell below the $p = 0.10$ threshold was day 1 on the validation data ($p = 0.085$). The corrected (bootstrapped) `slope`, `intercept`, and E_{max} statistics for the development data indicated negligible overfitting (and often under-fitting). Using the validation data, the calibration plots also indicated strong calibration, with the possible exception of day 2 and day 5 (which still compared favorably to my models).

5.6.5 Direct Model Comparisons

Due to differences in input requirements, slight differences existed between the sets of validation patients used to validate the models discussed above. To better compare the models against each other, I used a matched set of patients (the intersection of patients with valid predictions from each model). Table 5.27 on page 115 shows that the `SDAS` and `RAS` models were consistently more constrained than the `DAS`n and `SAPSII`$_a$ models. The `DAS`n and `SAPSII`$_a$ models had similar coverage.

The discrimination performance for the models was generally strong. If one compares Figure 5-24 to Figure 5-25, it is clear that the AUC performance only dropped by about 3% between the development and validation data. The worst validation performance occurred on later ICU days.

Looking at the validation data, the `DAS1` model had the best AUC performance on day 1, but the `SDAS` model performed best on days 2 through 5. These differences, however, were marginal. For example, on a given day no significant difference was found between the AUC from the `RAS` model and the model with the best AUC performance. The `SAPSII`$_a$ AUC performance was consistently below the other models, but discriminated best on day 3 with an AUC of 0.849. Even on this day the `SAPSII`$_a$ AUC performance remained significantly below the `RAS` performance ($p = 0.0437$).

According to the H and C statistics, the calibration performance on the validation data generally improved with each day in the ICU. The calibration for `SAPSII`$_a$ was consistently stronger than the other models on both the development and validation data except for day 1 development data where $H = 34.6$ with *d.f.*=9. When all `SAPSII`$_a$ development patients were used (i.e., not a matched subset), the `SAPSII`$_a$ H statistic was much smaller with a value of 11.2 (*d.f.*=10).

When the set of comparison patients was further restricted to only include patients that were in the ICU for at least 5 days with a valid prediction from each model for each day, the validation AUC performance generally suffered. A strong positive trend was observed in AUC as the ICU day increased for both the development and validation patients.

5.6.6 Limitations and Future Work

One weakness in the comparison of the models that I provided in this chapter is the slight bias given to the most input-constrained model. The most input-constrained models were validated on a subset of patients that may closer match the set of patients used for development whereas the less input-constrained models may be validated on a slightly constrained population that might deviate slightly from the less constrained population used for development. While this bias likely exists, its effect appears to be quite small. The AUC performances in Figure 5-25 align closely with those found using the entire validation set available to each model. Using a principled method for systemically dealing with missing variables, such as variable imputation, would provide an even stronger comparison between models.

Another limitation for this study was the use of at most one transformation of a given variable. By combining multiple transformations, I could have supported more complex (e.g., parabolic-shaped) effects.

By including treatment features, the results that I present are open to additional scrutiny. One hopes that each patient received optimal care. The presence of powerful intravenous drugs, for example, reflect a significant underlying patient risk as understood by the caregiver. One way to possibly mitigate this would be to explore automatic prediction of the need to administer specific drugs, but this may be infeasible given the patient data collected.

Throughout my analysis in this chapter, another question that remains is the best way to compare models with different temporal resolutions. I chose to use the mean daily-prediction for the higher resolution models as a summary, but there exist many alternatives to this choice.

In comparing calibration between my models and the $\texttt{SAPSII}_a$ model, I was likely liberal on my choice to allow SAPS II an extra two degrees of freedom for the development data. While the score was not *derived* using the development data, the logistic regression coefficients were fitted to this data. Considering this, the p-values for H and C statistics are likely somewhat inflated for the $\texttt{SAPSII}_a$ development data performance.

As one final note, my results are based on the analysis of the patient population from only one hospital. As with other retrospective studies, the results need to be validated on an external population to be fully generalized.

5.6.7 Conclusions

This chapter presented three types of mortality models: a stationary daily acuity score (`SDAS`), a daily acuity score ($\texttt{DAS}n$), and a real-time acuity score (`RAS`). For comparison, a customized SAPS II model ($\texttt{SAPSII}_a$) was also fit to the same development data. In general, my models demonstrated strong AUC performance that was significantly better than $\texttt{SAPSII}_a$ and mixed calibration performance that was weaker than $\texttt{SAPSII}_a$.

Between my models, the real-time model, `RAS`, performed similarly to the daily models, `SDAS` and $\texttt{DAS}n$, on the validation data. This is significant because the `RAS` model's prediction — based on any individual observation and limited trend information — was presumably more difficult than the prediction task of a daily model (looking at daily aggregate data).

While AUC performance was consistently strong, the calibration performance was mixed. Like many severity of illness metrics, without further adjustment, the probability values provided from these models should be used only with an understanding of their limitations. Using the corrected slope and intercept from the calibration plot, the calibration of the predicted probabilities from the model can be largely corrected. Discrimination, on the other hand, is more difficult to improve [23].

If the individual inputs for my models are examined, a number of interesting variables can be seen. Without the simplicity constraint commonly placed on other severity of illness metrics, a number of computationally intensive variables, such as the amount of time with a low SpO_2 or the slope of platelet administration, were included.

Also in contrast to most other severity scores, a number of interventions were included in my acuity models. The impact of the caregiver interventions, however, was not as strong as one might have expected. One explanation for this difference may be variation in caregiver practice.

In conclusion, the results of this chapter indicate that real-time mortality models are indeed feasible. I showed strong discriminatory ability for a real-time mortality model (i.g., good risk ranking). When considering risk estimates for individual patients, however, it remains important to carefully consider the model calibration (i.e., adequacy of individual risk estimates). With these considerations, additional work is needed to address the clinical utility of such real-time predictions.

Chapter 6

Predicting Secondary Outcomes

In contrast to the previous chapter, which examined models that predicted mortality, this chapter examines models to predict acute events that occur during a patient's ICU stay. These secondary outcomes represent significant events within a patient's stay that are automatically identifiable within our data. Both event *onset* (e.g., septic shock) where the outcome is bad and event *resolution* (e.g., successful weaning of pressors) where the outcome is beneficial, are considered.

Specifically, I develop models for the following events: (1) weaning of pressors (`PWM` model), (2) weaning of pressors *and* survival (`PWLM` model), (3) removal of intraaortic balloon pump (IABP) (`BPWM` model), (4) onset of septic shock (`SSOM` model), and (5) kidney injury (`AKIM` model).

After describing each individual model, I provide a comparison between the model and the real-time acuity score (`RAS` model) developed in the previous chapter. I do this by looking at the ROC performance, the positive predictive value, the negative predictive value, and the context surrounding the events of interest. Comparing the performance of our general models against specific models allows one to understand the relationship between the general mortality model and models that are trained to predict specific secondary events.

6.1 `PWM`: Weaning of Pressors

Vasopressor and inotropic drugs play an important role in managing vascular resistance and cardiac output in critically ill patients. Throughout this thesis I will refer to vasopressor and inotropic drugs collectively as "pressors". Pressors allow caregivers to manipulate a patient's cardiovascular and respiratory systems. Three common hypotensive situations that necessitate such intervention are vasodilation due to sepsis, decreased cardiac output from cardiogenic shock, or hypovolemia due to hemorrhaging. Hypotension is often life threatening if not treated quickly as irreversible ischemic organ damage can result in a matter of minutes. Most pressors fall under the sympathomimetic (or adrenomimetic) agent category. Others stem from the phosphodiesterase inhibitor or antidiuretic hormone agonist groups.

Sympathomimetic agents are typically grouped by the adrenoreceptors that they act on. In general, the three adrenoreceptors and their primary roles are as follows: (1) α agonists increase peripheral vascular resistance and venous pressure; (2) β agonists stimulate the heart and increase cardiac output and, in the case of β_2 agonists, often *reduce* peripheral vascular resistance; and (3) *dopamine* agonists dilate splanchnic (visceral organ) blood vessels and renal blood vessels but are dose related. While many sympathomimetic agents act selectively on a subset of adrenoreceptors, none are perfectly selective [87]. In order of frequency and with the primary adrenoreceptor targets in parenthesis, the sympathomimetic agents that are commonly used in MIMIC II patients are as follows: (1) Neo-Synephrine[1] (selective $\alpha_1 > \alpha_2$); (2) Levophed[1] (selective α_1, α_2, and β_1); (3) dopamine (dopamine agonist[2]); (4) epinephrine (general α and β); and (5) dobutamine (selective $\beta_1 > \beta_2$).

Other pressors found in the MIMIC II data include phosphodiesterase inhibitors and antidiuretic hormone agonists. The two types of phosphodiesterase inhibitors include milrinone and, much less frequently, amrinone. Milrinone and amrinone limit the decomposition of cyclic adenosine monophosphate (cAMP) thereby increasing the cardiac intracellular calcium and creating an effect similar to β agonists. In addition, phosphodiesterase inhibitors cause vasodilation. As a secondary vasopressor, vasopressin is often administered to patients who do not respond adequately to other vasopressors [14]. Vasopressin, also referred to as antidiuretic hormone (ADH), is a hormone that regulates the body's water retention. Vasopressin causes the kidneys to retain fluid by increasing urine concentration and it also results in moderate vasoconstriction.

Due to the powerful influences that pressors exert on a patient's hemodynamic system, they are typically only used after an attempt to stabilize a patient with fluids. When pressors are needed, it is important to carefully monitor their administration as a variety of harmful side effects can occur. Side effects include excessive vasoconstriction, cardiac arrhythmias, myocardial infarction, pulmonary edema or hemorrhage, and, in the case of vasopressin, dangerous hyponatremia.

After a patient has been stabilized, the protocol for weaning him or her from pressors is typically an empirical choice made by the caregiver. Pressor weaning typically proceeds by titrating the infusion rate and adapting to the patient's response. While this might seem straightforward, the process is complicated by varying patient response to different pressor agents and the need to switch pressors or add multiple pressors to sustain adequate perfusion.

In this section I develop a model to predict the successful transition from pressor infusions to no pressor infusions. I refer to this model as the pressor wean model (`PWM`).

[1] We refer to phenylephrine and norepinephrine by the brand names Neo-Synephrine and Levophed (respectively) to maintain consistency with the MIMIC II database labels.

[2] At moderate doses dopamine activates β receptors and at high doses dopamine also activates α receptors.

6.1.1 Data and Patient Inclusion Criteria

Prior to model selection, the data were limited to patient episodes that satisfied our inclusion criteria. The inclusion criteria were specified to include patient episodes where the patient was receiving pressors and had been receiving pressors for at least two hours. As noted above, the following drugs were included: (1) Neo-Synephrine (phenylephrine), (2) Levophed (norepinephrine), (3) Epinephrine, (4) Dobutamine, (5) Milrinone, (6) Amrinone, (7) Dopamine, and (8) Vasopressin.

In the data, the median episode length for pressor infusions was about 12 hours. Figure 6-1 provides a histogram showing the distribution of episode lengths. In defining a "pressor episode", periods separated by up to 4 hours of no pressors were merged together. I used the median episode length as the early-warning window for predicting pressor weaning. The median episode length allowed for a warning period that covered multiple nursing shifts while at the same time was not overly influenced by the potentially obvious behavior at the end of a successful weaning attempt.

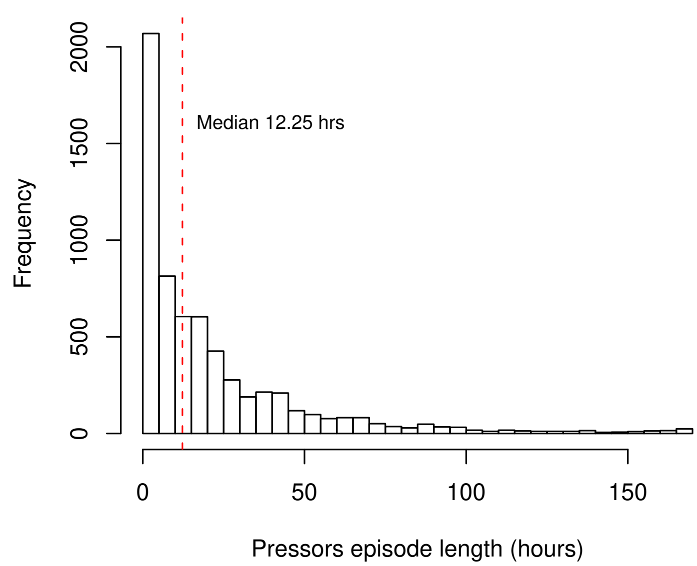


Figure 6-1: Pressor-infusion episode lengths

After annotating the final dataset, a number of instances were excluded where patients were not receiving pressors. Table 6.1 provides a summary of the included data used to develop the `PWM` model.

Table 6.1: PWM data

	Count
Included patients	3916
Included instances	215800
Weaned within 12 hours	56591
Not weaned within 12 hours	159209

6.1.2 Outcome

The outcome of interest for the PWM model is the successful discontinuation of all vasopressors and inotropic agents within 12 hours of the current point in a patient's stay. To qualify as discontinued, the patient must remain free of pressors for at least four consecutive hours.

To illustrate the pressor wean annotations, Figure 6-2 shows how an example patient (Subject_ID 2917) was annotated. The top plot shows the "pressor weaned" marks. To be weaned from pressors, the patient was required to be off all pressors and stay off pressors for at least 4 hours. The bottom plot in the figure shows the 12 hour warning annotations (i.e., the desired output from the trained model). Episodes where no annotations were made, due to absence of pressors or pressors for less than 2 hours, are marked with the dashed blue line. While the first pressor weaning for the patient in Figure 6-2 was temporary, it was long enough (≥ 4 hours) to be marked as a successful wean.

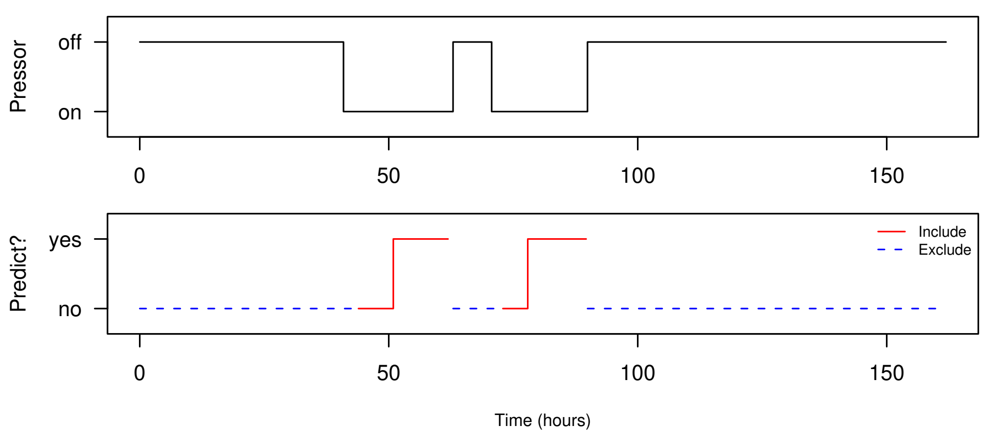


Figure 6-2: PWM example annotations for Subject_ID 2917

6.1.3 Model Development

To develop the PWM model, I follow the methodology described in Chapter 4. I first describe the model selection process and the resulting logistic regression model. Following the model selection description, I describe validation on the training (development) data.

Model Selection

Candidate variables were initially ranked against the outcome variable (successful weaning of pressors). Variables with a p-value greater than 0.05 were excluded. Furthermore, if multiple variables were strongly correlated (Spearman's rank correlation test > 0.8) the best univariate variable was retained. After the initial screening of the variables, variable selection for the PWM model was based on the best 40 variables from each of the top 4 of the 5 cross-validation folds (the individual cross validation plots are provided in Appendix F). When combined, the best 40 variables from the top 4 folds resulted in 72 candidate variables. Figure 6-3 shows the AUC that resulted from gradually increasing the AIC backward elimination threshold and greedily dropping additional variables.

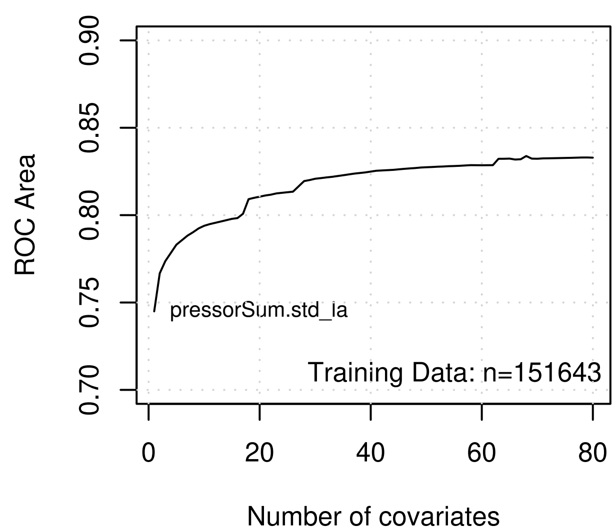


Figure 6-3: PWM model selection (all development data)

The abrupt drops in AUC performance in Figure 6-3 typically relate to removing variables that contained missing observations. For example, the drop between 19 covariates and 17 covariates resulted from removing totOut_am (total output deviation from mean) and X24hUrOut_am (24-hour urine output deviation from mean). These output variables contained a large number of missing values (more than 27000 of the 152082 training instances). By removing frequently absent variables, the model was less constrained and performance deteriorated. The missing values for `totOut_am` and `X24hUrOut_am` were not entirely random. Instead, they disproportionately affected the instances from ICU day 1, where the fluid output was infrequent and measurements were often unavailable. Given the critical role that fluid management plays in pressor decisions, the total output and 24-hour urine variables were included in my model despite their availability concerns. The final model was trained using the top 32 variables. No manual changes to the automatically derived model were deemed necessary, and the final model is described in Model 6.1.

In the description of Model 6.1, transformations applied to variables are denoted by the variable's suffix (such as "_i" for inverse). For example, the variable `pressorSum.std_la` should be interpreted as the logarithm of the absolute value of `pressorSum.std` (for a list of transformations and their abbreviations, see page 55 of Chapter 4). As explained in Chapter 3, the range of a variable up to the current point in a patient's stay is denoted by a suffix of "rng", and similarly the relative deviation of a variable from its evolving baseline is denoted by the suffix "rdv". To interpret model inputs, a positive coefficient should be understood as increasing the probability of a pressor wean (positive correlation) while a negative coefficient means a lower probability of a pressor wean (negative correlation).

Most of the variables included in Model 6.1 are reasonably clear and expected. A couple of the variables, however, may benefit from further description. For example, `cumPressorTime_am` — the difference between the cumulative amount of time that a patient has spent on pressors and the average total time that the population spends on pressors — increases the probability of a negative outcome.[3] That is, if a patient was recently started on pressors or has been on pressors for a long time he or she is less likely to be weaned in the following 12 hours.

In addition to `cumPressorTime_am`, a number of other important variables measure a variable's deviation from the mean (the `am` or `lam` (log `am`) transforms). The `CV_HR_Slope_1680_am` variable, for example, indicates that a large trend in the heart rate over the past 28 hours increases the probability for a successful wean.[4]

Development Validation

To validate the `PWM` model, I examine calibration performance and AUC performance. In addition, I also plot the positive predictive value (PPV) versus sensitivity and the negative predictive value (NPV) versus specificity. Table 6.2 shows the deciles used for the Hosmer-Lemeshow H statistic and Table 6.3 shows the deciles used for the Hosmer-Lemeshow C statistic. The classification performance of `PWM` on the training data is shown by the ROC curve in 6-4. In addition, plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-5

[3]On average, patients in the dataset spend a total of about 13 hours on pressors, including patients who never receive any pressors.

[4]The mean 28-hour heart rate slope for patients on pressors is about 0.00026 bpm/hr.

Model 6.1 Final PWM model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
102183	2e-06	26829.71	32	0	0.822	0.643
Gamma	Tau-a	R2	Brier			
0.644	0.238	0.344	0.139			

	Coef	S.E.	Wald Z	P
pressorSum.std_la	-7.815e-01	1.515e-02	-51.57	0
cumPressorTime_am	-2.580e-04	6.356e-06	-40.58	0
Milrinone_perKg	-3.554e+00	1.230e-01	-28.89	0
Sympathomimetic_agent	-8.035e-01	3.684e-02	-21.81	0
Intercept	-3.272e+00	1.703e-01	-19.21	0
Creatinine_sqrt	-4.703e-01	2.605e-02	-18.06	0
Neosynephrine_lam	-1.912e-01	1.210e-02	-15.80	0
SBPm.oor120.t_sqrt	-5.390e-02	3.514e-03	-15.34	0
iabp	-4.192e-01	2.806e-02	-14.94	0
totOut_am	-1.639e-04	1.121e-05	-14.63	0
Ativan_la	-5.985e-02	4.626e-03	-12.94	0
Levophed_perKg_lam	-1.464e-01	1.142e-02	-12.82	0
Art_PaCO2_am	-2.360e-02	1.963e-03	-12.03	0
Fentanyl_perKg_la	-3.461e-02	2.937e-03	-11.78	0
LactateM	-2.087e-01	1.843e-02	-11.33	0
ShockIdx	-5.089e-01	4.660e-02	-10.92	0
SpO2LowCntN_sqrt	-1.314e-01	1.237e-02	-10.62	0
INRrng_sqrt	-3.357e-01	3.262e-02	-10.29	0
Natrecor_la	-8.382e-02	8.305e-03	-10.09	0
SICU	-8.622e-01	8.576e-02	-10.05	0
PVC	-1.962e-01	2.050e-02	-9.57	0
AIDS	-9.189e-01	9.759e-02	-9.42	0
X24hUrOut_am	1.518e-04	1.574e-05	9.64	0
PTrng_sqrt	1.358e-01	1.388e-02	9.79	0
Fentanyl_Conc_i	9.733e-05	9.825e-06	9.91	0
pressD12	5.015e-01	4.819e-02	10.41	0
Sex	1.984e-01	1.812e-02	10.95	0
Integrelin_perKg_sq	4.428e-01	3.830e-02	11.56	0
CV_HR_Slope_1680_am	1.413e+01	1.180e+00	11.98	0
Doxacurium_sq	2.760e+00	2.010e-01	13.73	0
HCT_i	2.895e+01	1.967e+00	14.72	0
pressD01	6.084e-01	3.899e-02	15.60	0
GCS	4.479e-02	2.299e-03	19.48	0

Table 6.2: `PWM` Hosmer-Lemeshow H risk deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[3.59e-05,0.0207)	0.011	58	108.2	10161	10110.8	10219
2	[2.07e-02,0.0466)	0.033	235	334.1	9983	9883.9	10218
3	[4.66e-02,0.0812)	0.063	635	645.3	9583	9572.7	10218
4	[8.12e-02,0.1281)	0.104	1067	1061.2	9152	9157.8	10219
5	[1.28e-01,0.1827)	0.155	1648	1582.2	8570	8635.8	10218
6	[1.83e-01,0.2475)	0.214	2245	2187	7973	8031	10218
7	[2.47e-01,0.3292)	0.286	2910	2925.5	7309	7293.5	10219
8	[3.29e-01,0.4301)	0.377	4026	3855.2	6192	6362.8	10218
9	[4.30e-01,0.5772)	0.499	5218	5097.3	5000	5120.7	10218
10	[5.77e-01,0.9974]	0.704	6947	7193.1	3271	3024.9	10218

$\chi^2 = 105.76,\ d.f. = 8;\ p = 0.000$

Table 6.3: `PWM` Hosmer-Lemeshow C probability deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.042	1323	1484.7	33713	33551.3	35036
2	(0.1,0.2]	0.148	2895	2804.4	16078	16168.6	18973
3	(0.2,0.3]	0.247	3544	3514.7	10672	10701.3	14216
4	(0.3,0.4]	0.348	3785	3737.3	6947	6994.7	10732
5	(0.4,0.5]	0.447	3732	3617	4366	4481	8098
6	(0.5,0.6]	0.547	3485	3349.9	2635	2770.1	6120
7	(0.6,0.7]	0.65	2782	2825.2	1566	1522.8	4348
8	(0.7,0.8]	0.744	2096	2203.2	867	759.8	2963
9	(0.8,0.9]	0.842	1106	1194.2	313	224.8	1419
10	(0.9,1]	0.93	241	258.5	37	19.5	278

$\chi^2 = 121.81,\ d.f. = 8;\ p = 0.000$

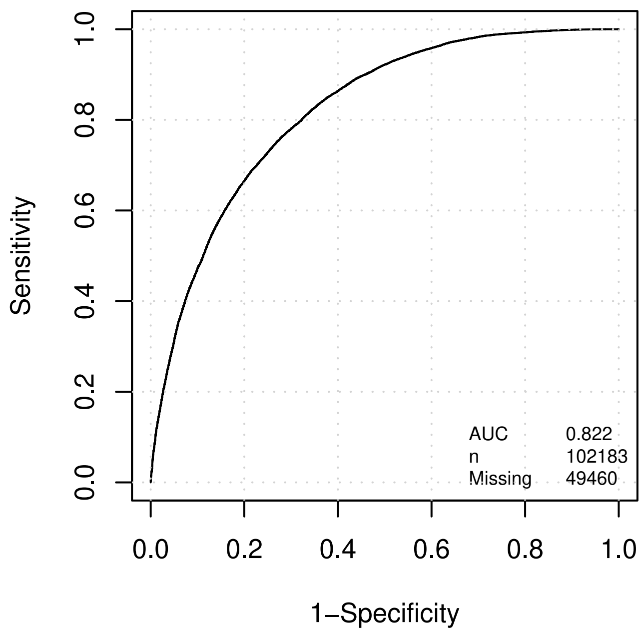


Figure 6-4: PWM ROC curve (development data). AUC = the area under the curve; n = the total number of valid predictions used to make the curve; $Missing$ = number of unavailable predictions from the model due to missing data.

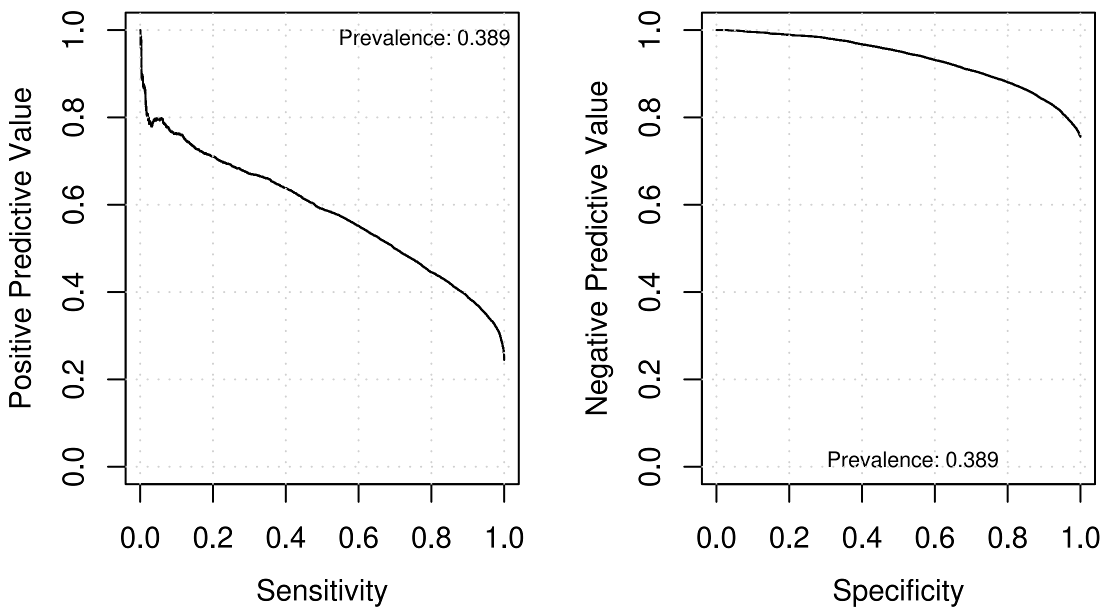


Figure 6-5: `PWM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (development data).

To graphically summarize the `PWM` predictions in the context of successful pressor weans, Figure 6-6 shows the aggregate behavior of `PWM` predictions over non-overlapping 3-hour intervals up to 18 hours before and after the event. The event (in this case, a "successful" pressor wean) is indicated by the dotted vertical line at $t = 0$ in the center. This figure separates the prediction summaries by patients who lived (left) and patients who died (right). The two missing bars following the center event are the result of the inclusion criteria and the event definition: a patient must have been receiving pressors for at least two hours previously to be included and a successful wean by my definition is followed by at least 4 hours with no pressors.

For each time interval, the distribution of model predictions over the interval is summarized by a box-and-whisker plot. The middle 50% of the data (i.e., the interquartile-range or IQR) are represented by the hollow vertical box. The lines extending from the box (i.e., the "whiskers") extend to the furthest estimate that is within $1.5 \times IQR$ of the top of the box and within $1.5 \times IQR$ of the bottom of the box. The solid horizontal bar in each box represents the median prediction value for the distribution over the respective interval. Finally, the number of predictions that fall into each interval is provided along the x-axis and the average overall estimate (e.g., the average of all `PWM` predictions for patients who lived) is shown along the y-axis. It is helpful to think of each interval as a conditional distribution (conditioned on time). By placing the conditional distributions in temporal order, insight into the aggregate behavior of the estimate can be obtained. By stratifying patients based on their final outcome, general differences in the model's prediction for the two classes of patients can be examined. Collectively, these visual representations allow one to visualize the distribution of model estimates as a function time for patients who lived and for patients who died.

A second pair of context plots is provided that examine all available predictions, including points that did not satisfy the inclusion criteria. These plots are shown in Figure 6-7. Without requiring the inclusion criteria to be satisfied, the left plot in Figure 6-6 includes estimates for the two intervals directly following the successful pressor wean.

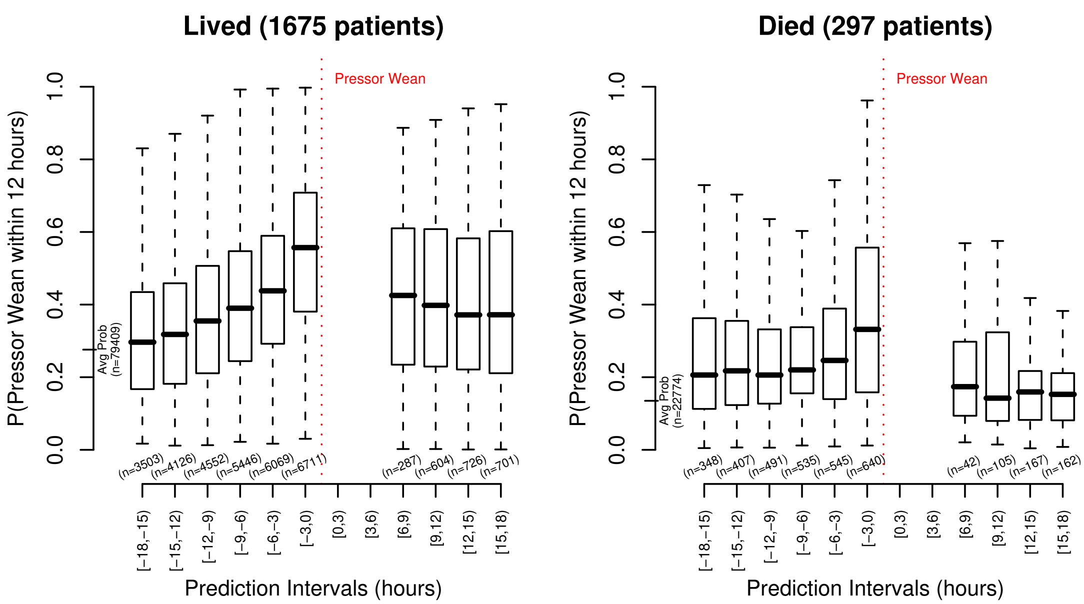


Figure 6-6: PWM prediction context surrounding successful pressor weans (development data). *Avg Prob*: the mean PWM probability from all patients who lived (left) and died (right).

Finally, as an illustration of predictions for an individual patient, Figure 6-8 shows the predictions for the patient shown earlier (in Figure 6-2) during the discussion of the annotation process.

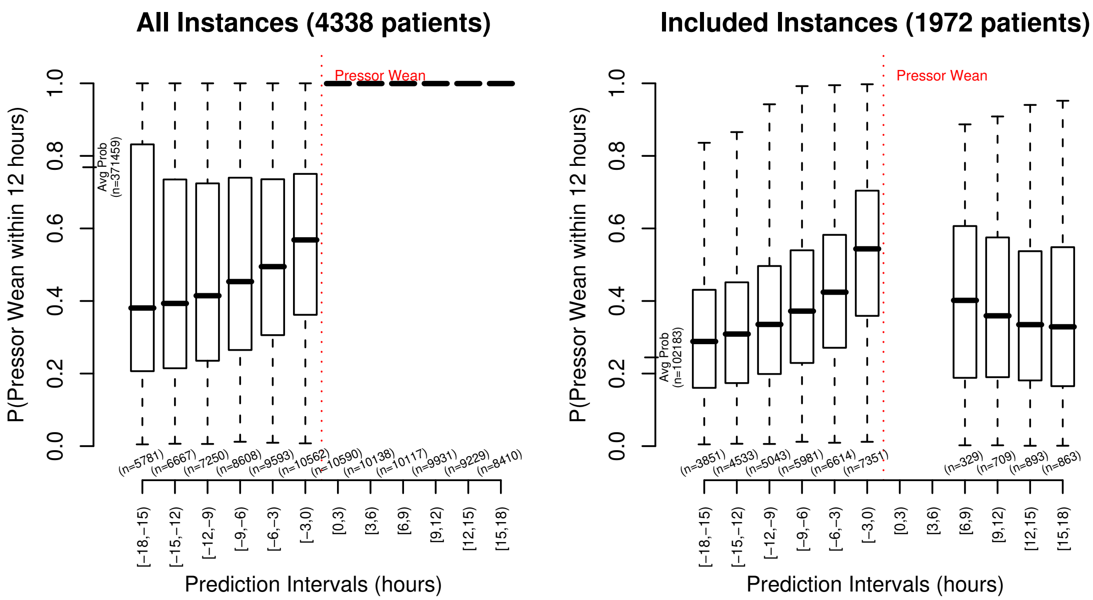


Figure 6-7: PWM prediction context surrounding successful pressor weans (development data). *Avg Prob*: the mean PWM probability from all patient instances (left) and valid instances (right).

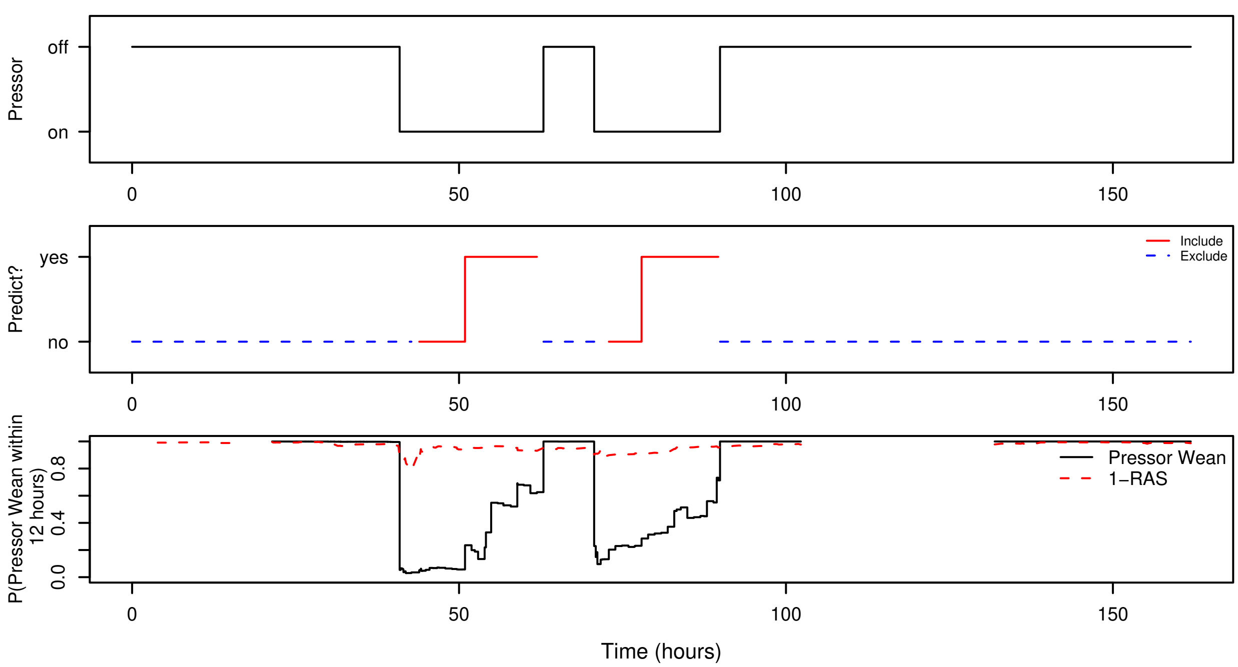


Figure 6-8: PWM annotations for Subject_ID 2917 with PWM and RAS predictions

Table 6.4: PWM Hosmer-Lemeshow H risk deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[8.85e-05,0.0242)	0.011	72	47.7	4296	4320.3	4368
2	[2.42e-02,0.0546)	0.039	149	170.4	4219	4197.6	4368
3	[5.46e-02,0.0909)	0.072	294	316.2	4074	4051.8	4368
4	[9.09e-02,0.1352)	0.112	479	488.2	3889	3879.8	4368
5	[1.35e-01,0.1941)	0.164	685	717.8	3683	3650.2	4368
6	[1.94e-01,0.2698)	0.231	986	1008.4	3382	3359.6	4368
7	[2.70e-01,0.3584)	0.313	1409	1366.9	2959	3001.1	4368
8	[3.58e-01,0.4692)	0.412	1716	1800.9	2652	2567.1	4368
9	[4.69e-01,0.6102)	0.537	2181	2344.7	2187	2023.3	4368
10	[6.10e-01,0.9592]	0.724	2823	3161.7	1545	1206.3	4368

$\chi^2 = 184.31,\ d.f. = 10;\ p = 0.000$

6.1.4 Model Validation

As a final step, I validate the PWM model on the separate validation data. To evaluate calibration, Table 6.4 and Table 6.5 provide the deciles used for the Hosmer-Lemeshow statistics. A plot of the calibration — actual probability versus estimated probability — is shown in Figure 6-9. The PWM classification performance is summarized by the ROC curve in Figure 6-10. If the performance evaluation is limited to exclude the warnings that occur within 6 hours of full weaning of the patient (i.e., only considering predictions between 6 and 12 hours before event instead of between 0 and 12 hours), the AUC performance drops to 0.78 (development data) and 0.76 (validation data). Excluding predictions that occur within 6 hours of a full pressor eliminates about 50% of the available predictions. For comparison purposes, Figure 6-10 includes a curve generated by my real-time general acuity model (the RAS model discussed in Chapter 5) applied to the same prediction task as PWM. The RAS predictions of pressor weaning are shown in Figure 6-10 as dotted blue lines. The comparison of a specialized model's predictions (PWM) against the output from the general acuity model (RAS) was motivated by the idea that a generic understanding of the patient's condition might be helpful in understanding other events. Strong correlation between mortality predictions from RAS and predictions of clinically significant outcomes from specialized models, such as PWM, may indicate that the specialized model is unnecessary.

Plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-11. The dotted blue lines show the performance obtained by using the RAS model output as a proxy to predict the same outcome as PWM.

Finally, as done previously with the development patients, the context surrounding successful pressor weans for the validation patients is examined. Figure 6-12 shows the context surrounding successful weans for patients who survived (left) and

Table 6.5: `PWM` Hosmer-Lemeshow C probability deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.045	604	631.3	13518	13490.7	14122
2	(0.1,0.2]	0.146	1142	1182.2	6947	6906.8	8089
3	(0.2,0.3]	0.248	1370	1379	4186	4177	5556
4	(0.3,0.4]	0.348	1607	1573	2916	2950	4523
5	(0.4,0.5]	0.448	1573	1670.4	2153	2055.6	3726
6	(0.5,0.6]	0.548	1503	1634	1477	1346	2980
7	(0.6,0.7]	0.645	1344	1488.2	963	818.8	2307
8	(0.7,0.8]	0.749	1084	1174.8	485	394.2	1569
9	(0.8,0.9]	0.84	463	556.1	199	105.9	662
10	(0.9,1]	0.917	104	133.9	42	12.1	146

$\chi^2 = 283.18,\ d.f. = 10;\ p = 0.000$

patients who died (right). In addition to the `PWM` predictions, the figure also shows *survival* predictions from the `RAS` model. Figure 6-13 shows the prediction context for all predictions, including ones that did not satisfy inclusion criteria (left), and the prediction context for patients that did satisfy the inclusion criteria (right).

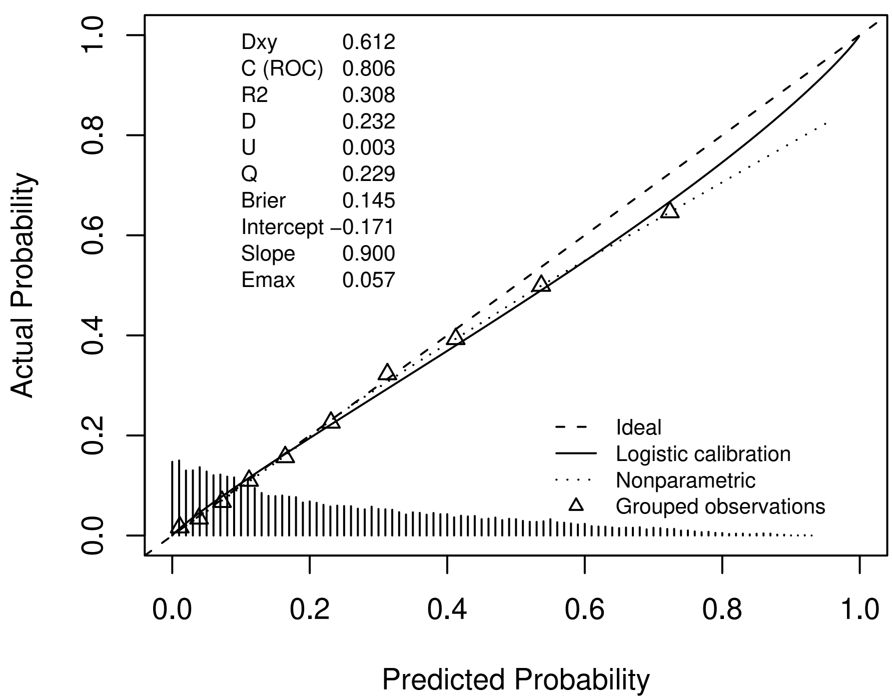


Figure 6-9: PWM calibration plot

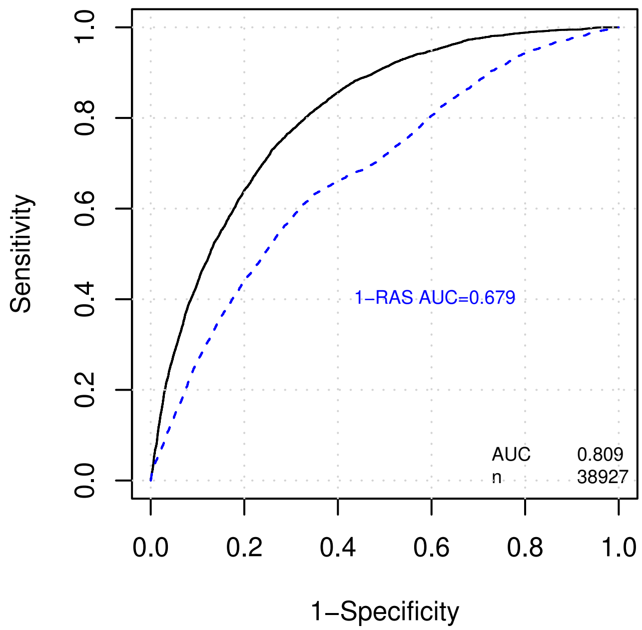


Figure 6-10: `PWM` ROC curve (validation data). AUC = the area under the curve; n = the total number of valid predictions used for curve.

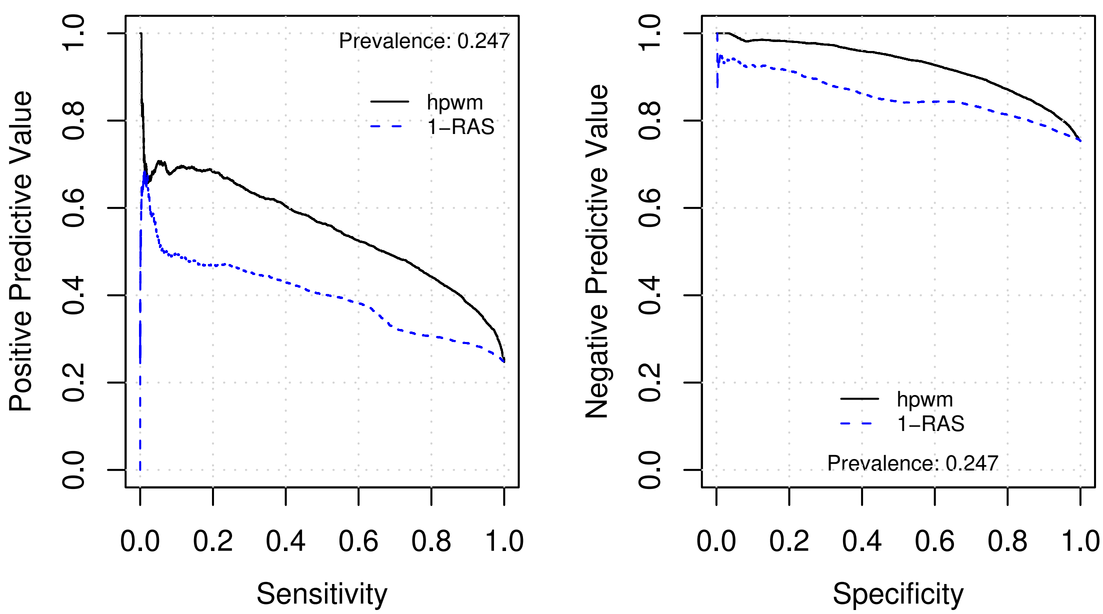


Figure 6-11: `PWM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (validation data).

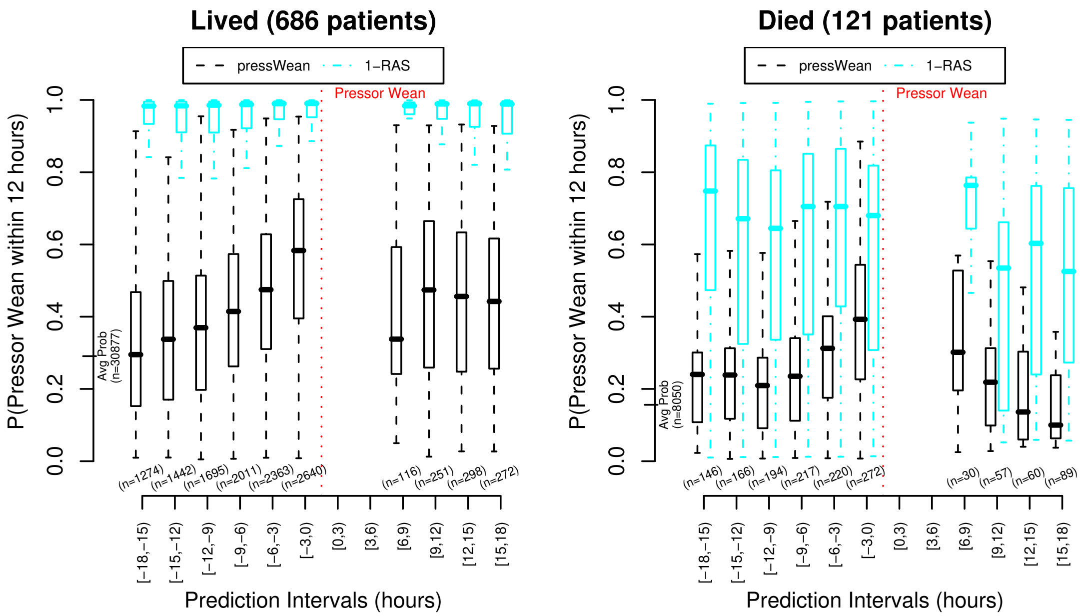


Figure 6-12: PWM prediction context surrounding successful pressor weans (validation data). *Avg Prob*: the mean PWM probability from all patients who lived (left) and died (right).

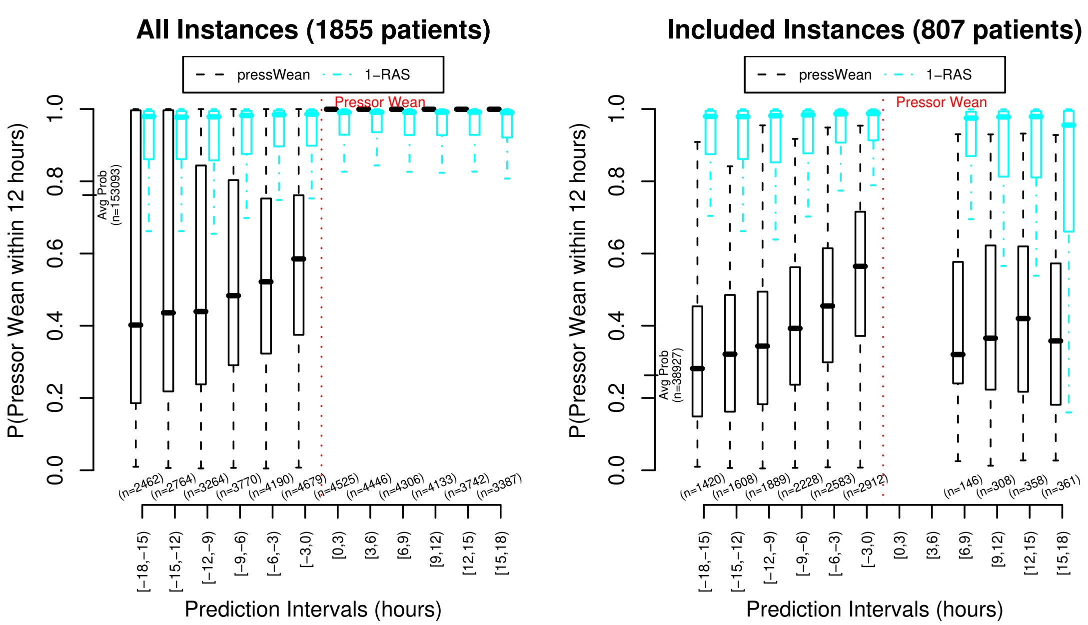


Figure 6-13: PWM prediction context surrounding successful pressor weans (validation data). *Avg Prob*: the mean PWM probability from all patient instances (left) and valid instances (right).

6.1.5 Discussion

The `PWM` model developed in this section performs well at discriminating between patient segments that precede successful pressor weaning by 12 hours or less and segments that precede pressor weaning by more than 12 hours. The AUC performance of about 0.81 is especially good considering the somewhat arbitrary choice for the warning window of 12 hours. While a 12-hour window includes cases where the prediction should be quite easy, such as cases where a patient is nearly weaned from pressors and goes entirely off in an hour, it also includes more challenging assessments such as cases where the patient is not fully weaned until 11 hours in the future. In fact, if the performance is limited to only look at warnings between 12 and 6 hours the AUC only drops to 0.78 (development data) and 0.76 (validation data).

The final model (Model 6.1) includes 32 inputs. Unsurprisingly, the two most significant inputs are the current level of pressors that the patient is receiving (`pressorSum.std_la`) and the total time that the patient has spent on pressors during his or her stay (`cumPressorTime_am`). If a patient is on a high dosage of pressors, they are likely to remain on pressors for more than 12 hours. In terms of pressor time, the deviation from the average cumulative time spent on pressors (the "am" transformation) indicates that patients on pressors for a short period of time or a long period of time are less likely to be weaned within 12 hours. The mean cumulative time that the patients in the dataset spend on pressors is about 13 hours.

The third input, `Milrinone_perKg`, is possibly more interesting. Milrinone is an inotropic agent that is given for acute heart failure and it has a rather long half-life of nearly 2.5 hours (dobutamine, in contrast, has a half-life of about 2 minutes). The long half-life for milrinone results in a prolonged weaning process. Furthermore, one of the common side effects for milrinone is an increase in ventricular ectopic activity. A variable indicating the presence of premature ventricular contractions (`PVC`) is included in the top 32 variables and also decreases the probability of a successful wean within 12 hours.

The drugs that increase the probability of a successful pressor wean within 12 hours include doxacurium (`Doxacurium_sq`), and Integrelin (`Integrelin_perKg_sq`). Doxacurium is a muscle relaxant given to patients during surgery or other procedures such as starting the patient on a mechanical ventilator. The presence of doxacurium likely indicates that the patient is receiving interventions in addition to pressors. Similarly, Integrelin is an antiplatelet agent that is often given to treat patients with acute myocardial ischemia who are receiving coronary angioplasty. Integrelin and doxacurium seem to serve as a proxy for why the patient received pressors and thereby provide insight into how long the pressors will likely be needed.

Since the outcome of interest is the weaning of pressors, it is expected that a number of the most predictive variables are therapy variables. A number of physiologic variables are also important. Some of the most significant physiologic variables include the Glasgow Coma Scale (`GCS`), creatinine (`Creatinine_sqrt`), amount of time that the systolic blood pressure was out of range during the past 2 hours (`SBPm.oor120.t_sqrt`), hematocrit (`HCT_i`), the total output from the patient (`totOut_am`), and the shock index (i.e., heart rate/systolic blood pressure) (`ShockIdx`). The interpretation for most of these physiologic variables is reasonably clear. A low hematocrit level, for example, is often indicative of hemorrhaging. The increased probability that the patient will be weaned within 12 hours that is associated with a low hematocrit can be explained by the fact that, once the hemorrhaging is addressed, the patient's need for pressors should diminish.

In general, the calibration for the `PWM` model was strong. While statistically significant values for the H and C statistics were found for the development and the validation data, the individual deciles appear to align well between the observed counts and the expected counts. As with the `RAS` model in the previous chapter, the large number of observations make small differences statistically significant and the differences are likely inflated as a result of multiple predictions per patient. The calibration plot in Figure 6-9 shows that calibration is generally good, and that with slight adjustment the predicted probabilities align quite well with the actual probabilities. The maximum error ($E_{\max}$) between the corrected curve and the actual probabilities is about 0.057.

The classification performance for `PWM` was also strong. The AUC on the separate validation data was found to be about 0.809 (Figure 6-10). This AUC was only 0.013 less than the AUC on the development data (Figure 6-4). The predictive value of the model was moderate as shown by the PPV versus sensitivity and the NPV versus specificity curves in Figure 6-11. For the included data points, the prevalence of successful weans within 12 hours was about 0.247 for the matched validation data (valid `PWM` and valid `RAS` estimates). By only sacrificing a small amount of sensitivity (e.g., 30%), a PPV of about 50% was obtainable. The maximum PPV from the model, however, was only about 70%. It is expected that the `RAS` estimates are less effective in terms of PPV than the `PWM` estimates, as the `RAS` score should reflect a number of dire conditions in addition to pressor dependence.

In the individual patient example provided in Figure 6-8, the `PWM` model appears to do a reasonable job of interpreting the weaning prospects for the patient. The gradually increasing probability until the time of the event (successful pressor wean) is expected given the discontinuity of the 12-hour warning that is used for fitting the model. If the context surrounding successful pressor weans is examined in aggregate, the trend noted in Figure 6-8 can be observed for the validation patients (Figure 6-12).

Furthermore, notable differences can be seen by comparing the predictions for the patients who lived to the predictions for the patients who died. The [3,0) hour warning interval in Figure 6-12, for example, has a mean prediction of about 0.59 for patients who lived and a mean of about 0.44 for patients who died ($p < 0.00001$). For patients that live, the median prediction demonstrates a linear increase from 18 hours prior to the event up until the point of the event. In contrast, for patients that died, the median prediction is lower and only starts to increase within 6 hours of the event. In general, the PWM model is much less sensitive at predicting weans for patients that ultimately die.

In the second context figure for the validation data, Figure 6-13, all instances with a prediction available (ignoring the inclusion criteria) are contrasted with the estimates that do satisfy the inclusion criteria. As one would expect, the model produces a very high probability output for instances that have no pressors present. The high probability for cases without pressors causes the left-most bars to extend much higher in the plot that looks at all instances, because many pressor episodes last less than nine hours. The number of instances that were excluded by the inclusion criteria can be found by calculating the difference between the interval counts between the two plots.

In comparison, the RAS model does significantly worse ($p < 0.00001$) than the PWM model at predicting the weaning of pressors within 12 hours. The RAS model provides some useful information when determining if the patient will be successfully weaned. When the PWM model is used to predict final patient outcome (instead of pressor weaning), it obtains an AUC of 0.713 on the development data and an AUC of 0.686 on the validation data. When the mean estimate for each patient is used, the AUC increases to 0.830 on the development data and 0.809 on the validation data. As one would expect, there appears to be a strong association between the ease of weaning a patient (PWM) and the risk of mortality (RAS). In fact, for the patients that died, the median 1-RAS score is higher for the [-18, -15), [-9, 6), and [-6, 3) hour intervals than it is for the [-3, 0) hour interval, which indicates an increasing mortality risk as the patient is weaned.

In conclusion, the PWM model does a good job of discriminating between which patient instances will be weaned from pressors within 12 hours and which patients instances will not be weaned within 12 hours. To make the pressor wean prediction, the most significant model inputs are treatments, but a number of important physiological inputs are included in the model as well. One confounder in the model construction is the inclusion of episodes where terminal patients were weaned from pressors. There is no way to tell if weaning the patient was *physiologically* justified.

Table 6.6: PWLM data

	Count
Included patients	3916
Included instances	215800
Weaned within 12 hours (and lived)	52076
Not weaned within 12 hours	163724

6.2 PWLM: Weaning of Pressors *and* Survival

In this section I present a slight variant of the PWM model. For this model I augmented the definition of a successful pressor wean used previously to also require patient survival. This model will be referred to as the pressor wean *and* live model (PWLM).

The motivation for the PWLM model is not to increase the prognostic value of the PWM model, but rather to further understand the limitations of the PWM model. By not counting patients that ultimately die in my definition of a successful wean, I do not penalize my model's performance for cases where patients perhaps should not have been weaned. Instead, the PWLM model only focuses on pressor weans that were clearly successful as measured by the patient's ultimate outcome. As noted in Chapter 3, periods with any limitation of support (e.g., comfort measures only) were excluded during the preparation of the dataset that all of my models are based on. Consequently, the PWLM model excludes the potentially obvious cases where a patient's support is removed as they are allowed to expire. The PWLM model only includes pressor weans during periods of full support (i.e., "full code").

6.2.1 Data and Patient Inclusion Criteria

The inclusion criteria used for the PWLM model were the same as those used for the PWM model: for an instance of a patient to be included, the patient was required to have been receiving pressors for at least 2 hours directly prior to the time of the instance.

Figure 6-14 provides histograms showing the distribution of pressor episode lengths for patients who lived (left) and patients who died (right). The median episode length for pressor infusions is about 12 hours for patients who lived. Again, in defining a "pressor episode", periods separated by up to 4 hours without pressors were merged together.

After annotating the final dataset, a number of patient instances were excluded because of no pressors. Table 6.6 provides a summary of the included data.

6.2.2 Outcome

The outcome of interest for the PWLM model is the successful discontinuation of vasopressors and/or inotropic agents (for at least four consecutive hours) occurring within

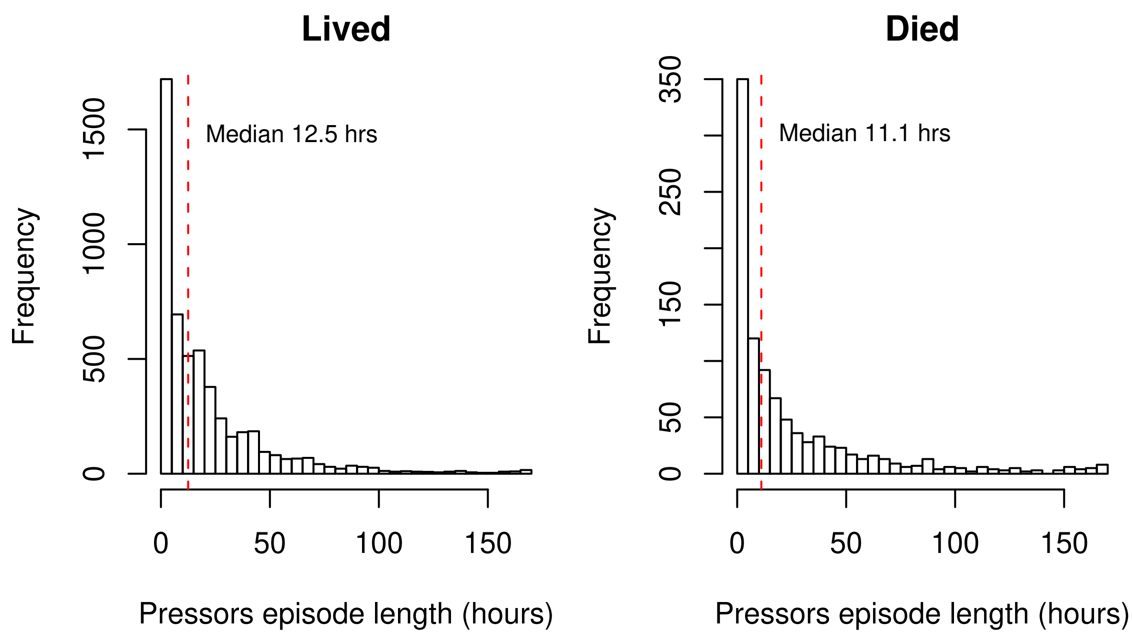


Figure 6-14: Pressor-infusion episode lengths

12 hours of a given point in a patient's stay *and* patient survival. The definition I use for survival is the same as the survival definition used for the `RAS` model previously — that is, the patient survived the ICU stay and at least 30 days in hospital or the patient was discharged from the hospital alive within 30 days.

Since `Subject_ID` 2917 (used previously to describe the annotation process for the `PWM` model) survived, the annotation for this patient remains unchanged (see Figure 6-15).

6.2.3 Model Development

Again, using the methodology described in Chapter 4, I first describe the `PWLM` model selection process and the resulting logistic regression model. After I describe the model selection, I explore the `PWLM` model's performance on the training (development) data.

Model Selection

Candidate variables were initially ranked against the outcome variable (successful weaning of pressors and survival). Variables with a p-value greater than 0.05 were excluded. Furthermore, if multiple variables were strongly correlated (Spearman's rank correlation test > 0.8) the best univariate variable was retained. After the initial screening of the variables, variable selection for the `PWLM` model was based

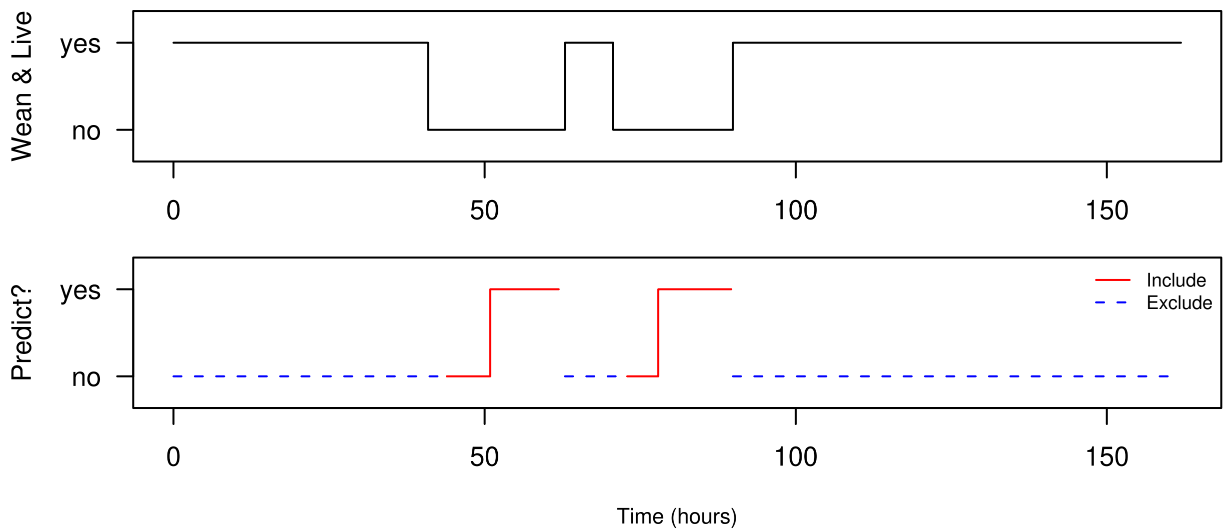


Figure 6-15: PWLM example annotations for Subject_ID 2917. Patient survived.

on the best 50 variables from each of the top 4 of the 5 cross-validation folds (the individual cross validation plots are provided in Appendix F). When combined, the best 50 variables from the top 4 folds resulted in 83 candidate variables. Figure 6-16 shows the AUC that results from gradually increasing the AIC backward elimination threshold and greedily dropping variables.

The top 32 variables were used to train the final model shown in Model 6.2. In the final model, of the six most predictive variables, five measure the deviation from normal (the _am transform).

Development Validation

To validate the PWLM model, I examine calibration performance and AUC performance. In addition, I also plot the PPV versus sensitivity and the NPV versus specificity. Table 6.7 shows the deciles used for the Hosmer-Lemeshow H statistic and Table 6.8 shows the deciles used for the Hosmer-Lemeshow C statistic. The classification performance of PWLM on the training data is shown by the ROC curve in 6-17. In addition, the PPV versus sensitivity plot and the NPV versus specificity plot are provided in Figure 6-18

Figure 6-19 shows the context surrounding successful pressor weans for all predictions, ignoring the inclusion criteria (left) and only patients that satisfied the inclusion criteria (right). Since successful pressor weans are defined by the PWLM model to only include patients that survive, no contrast is available between patients who lived and patients who died.

As an illustration of predictions for an individual patient, Figure 6-20 shows the predictions for the patient shown earlier (Figures 6-15 and 6-8).

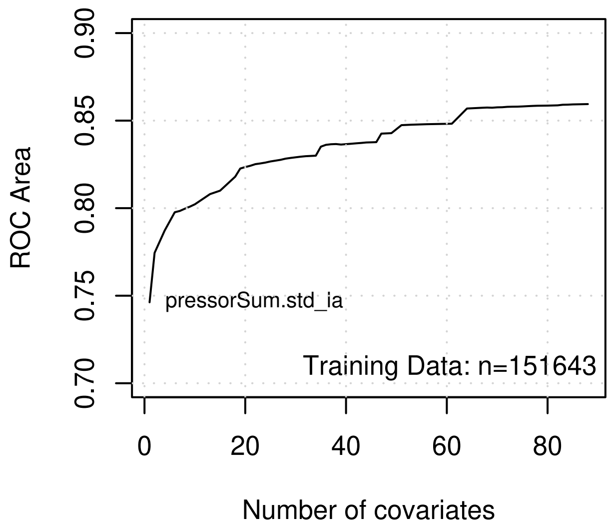


Figure 6-16: PWLM model selection (all development data)

Table 6.7: PWLM Hosmer-Lemeshow H risk deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[1.49e-06,0.0138)	0.006	49	70.5	11357	11335.5	11406
2	[1.38e-02,0.0347)	0.024	225	269.4	11181	11136.6	11406
3	[3.47e-02,0.0647)	0.049	519	558.4	10887	10847.6	11406
4	[6.47e-02,0.1047)	0.084	906	953.9	10499	10451.1	11405
5	[1.05e-01,0.1587)	0.131	1538	1489.1	9868	9916.9	11406
6	[1.59e-01,0.2243)	0.189	2239	2160.8	9167	9245.2	11406
7	[2.24e-01,0.3050)	0.263	2983	2996.7	8422	8408.3	11405
8	[3.05e-01,0.4067)	0.353	4198	4029.6	7208	7376.4	11406
9	[4.07e-01,0.5526)	0.474	5418	5401.1	5988	6004.9	11406
10	[5.53e-01,0.9634]	0.679	7604	7749.5	3801	3655.5	11405

$\chi^2 = 44.61,\ d.f. = 8;\ p = 0.000$

Model 6.2 Final PWLM model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
114057	2e-05	30004.52	32	0	0.83	0.659
Gamma	Tau-a	R2	Brier			
0.661	0.23	0.353	0.13			

	Coef	S.E.	Wald Z	P
Milrinone_perKg_am	-4.574e+00	1.296e-01	-35.29	0
cumPressorTime_am	-2.366e-04	7.593e-06	-31.17	0
vasopressorCnt_am	-7.406e-01	3.099e-02	-23.89	0
Neosynephrine_lam	-2.798e-01	1.280e-02	-21.85	0
Dobutamine_perKg	-2.315e-01	1.094e-02	-21.16	0
Levophed_perKg_lam	-2.441e-01	1.197e-02	-20.39	0
pressorSum.std_la	-4.105e-01	2.103e-02	-19.52	0
SBP_i	-1.095e+02	6.181e+00	-17.72	0
INRrng_sqrt	-5.661e-01	3.260e-02	-17.36	0
pressD24	-8.227e-01	4.941e-02	-16.65	0
Creatinine_sqrt	-4.585e-01	2.965e-02	-15.46	0
Art_PaCO2_am	-2.919e-02	2.064e-03	-14.15	0
vasopressorSum.std_lam	-1.315e-01	9.826e-03	-13.38	0
UrineEvnts.24h_sqrt	-6.624e-02	5.329e-03	-12.43	0
SICU	-1.187e+00	9.587e-02	-12.38	0
BUN_Slope_1680_lam	-1.609e-01	1.307e-02	-12.30	0
mechVent	-3.737e+03	3.097e+02	-12.07	0
VentMode_i	-3.737e-01	3.097e-02	-12.07	0
Art_pHrng	-1.296e+00	1.087e-01	-11.92	0
EctFreq_sqrt	-1.446e-01	1.397e-02	-10.35	0
DBPmrng_sqrt	-7.701e-02	7.804e-03	-9.87	0
BUNtoCr_sq	-1.824e-04	1.861e-05	-9.80	0
iabp	-2.580e-01	2.860e-02	-9.02	0
SBPmCritEvnts.24h_sqrt	-4.555e-02	5.245e-03	-8.68	0
Anticoagulant	-2.400e-01	3.103e-02	-7.73	0
Input_60rng_am	-2.788e-05	3.973e-06	-7.02	0
Dilaudid_sq	2.409e-01	2.688e-02	8.96	0
BUNrdv_sqrt	1.415e-01	1.310e-02	10.80	0
CV_HRrng_sqrt	6.669e-02	5.608e-03	11.89	0
Intercept	3.737e+03	3.097e+02	12.07	0
Ativan_perKg_i	6.852e-05	5.147e-06	13.31	0
GCS_sq	2.709e-03	1.677e-04	16.16	0
PTrng_sqrt	2.494e-01	1.359e-02	18.35	0

Table 6.8: PWLM Hosmer-Lemeshow C probability deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.039	1587	1738	42919	42768	44506
2	(0.1,0.2]	0.147	3073	2979.7	17151	17244.3	20224
3	(0.2,0.3]	0.248	3614	3592.1	10871	10892.9	14485
4	(0.3,0.4]	0.348	4108	3951.5	7261	7417.5	11369
5	(0.4,0.5]	0.447	3925	3856.1	4694	4762.9	8619
6	(0.5,0.6]	0.547	3261	3358.7	2875	2777.3	6136
7	(0.6,0.7]	0.648	2852	2765.6	1416	1502.4	4268
8	(0.7,0.8]	0.745	2250	2373.1	935	811.9	3185
9	(0.8,0.9]	0.836	939	993.5	249	194.5	1188
10	(0.9,1]	0.918	70	70.7	7	6.3	77

$\chi^2 = 86.36,\ d.f. = 8;\ p = 0.000$

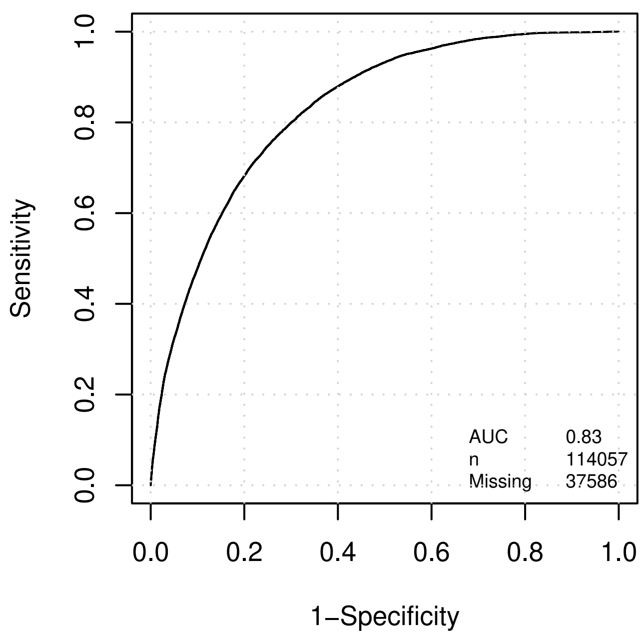


Figure 6-17: PWLM ROC curve (development data).

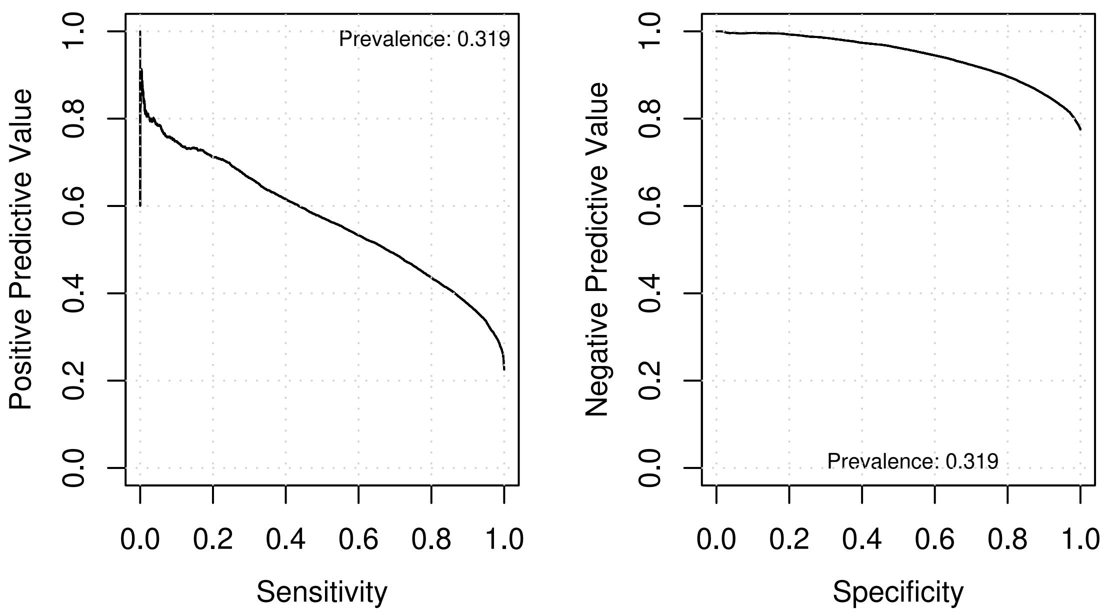


Figure 6-18: `PWLM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (development data).

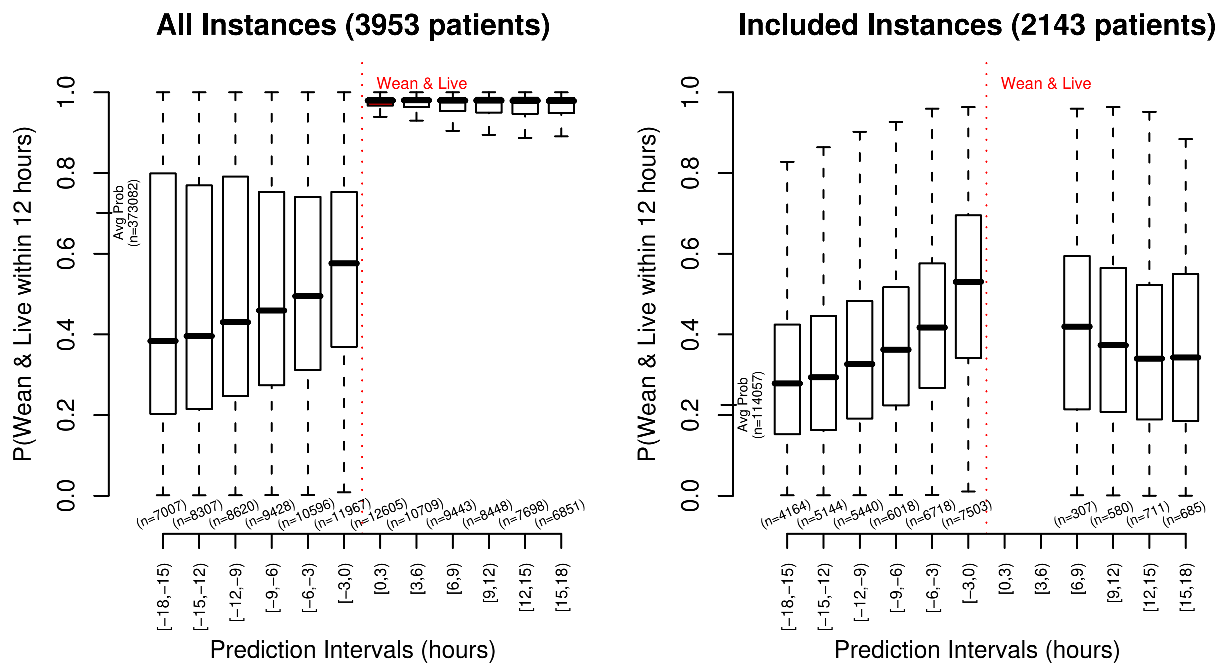


Figure 6-19: PWLM prediction context surrounding HDFR (development data). *Avg Prob*: the mean PWLM probability from all patient instances (left) and valid instances (right).

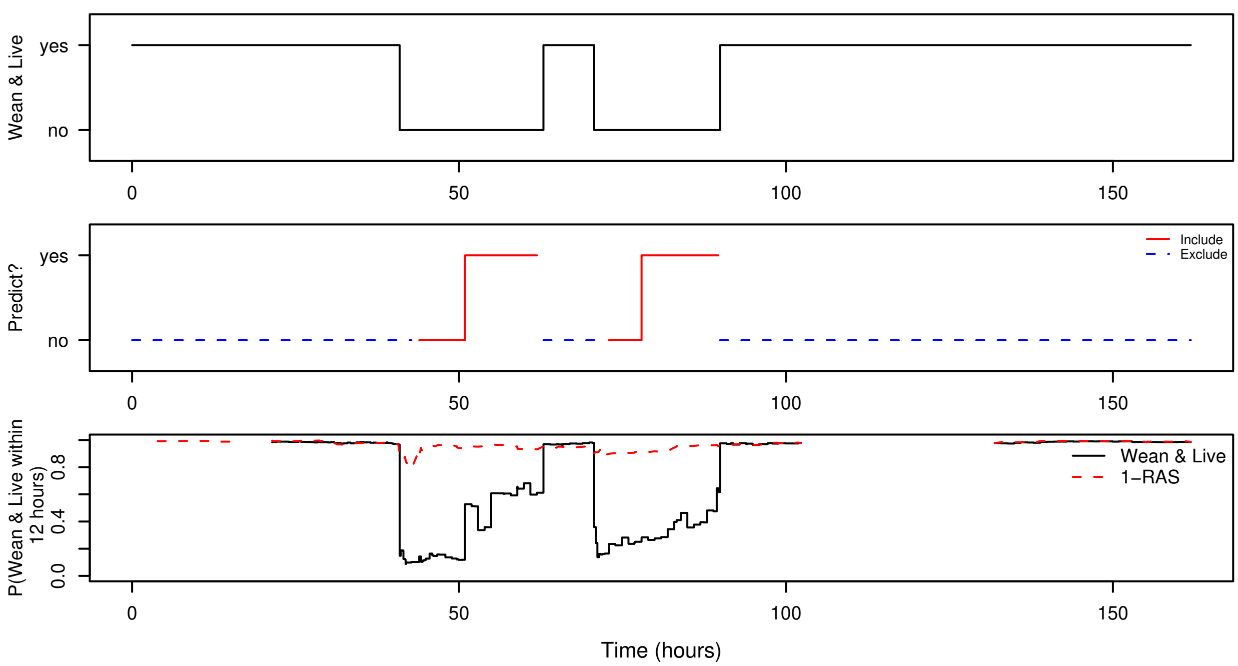


Figure 6-20: PWLM annotations for Subject_ID 2917 with PWLM and RAS predictions. Patient survived.

Table 6.9: PWLM Hosmer-Lemeshow H risk deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[2.48e-05,0.0140)	0.005	39	26.2	4882	4894.8	4921
2	[1.40e-02,0.0381)	0.026	94	127	4827	4794	4921
3	[3.81e-02,0.0725)	0.054	233	267.2	4688	4653.8	4921
4	[7.25e-02,0.1155)	0.093	429	458.5	4492	4462.5	4921
5	[1.16e-01,0.1794)	0.146	721	719.6	4200	4201.4	4921
6	[1.79e-01,0.2530)	0.216	1029	1060.7	3892	3860.3	4921
7	[2.53e-01,0.3393)	0.294	1274	1448.9	3647	3472.1	4921
8	[3.39e-01,0.4460)	0.39	1769	1919.2	3152	3001.8	4921
9	[4.46e-01,0.5865)	0.512	2312	2519.8	2609	2401.2	4921
10	[5.87e-01,0.9422]	0.704	3165	3461.4	1755	1458.6	4920

$\chi^2 = 192.85$, $d.f. = 10$; $p = 0.000$

6.2.4 Model Validation

As a final step, I validate the PWLM model on the separate validation data. To evaluate calibration, Table 6.9 and Table 6.10 provide the deciles used by the Hosmer-Lemeshow statistics. A plot of the calibration is shown in Figure 6-21. The PWLM classification performance is summarized by the ROC curve in Figure 6-22. For comparison purposes, Figure 6-22 includes a curve generated by the RAS model developed in the previous chapter applied to the same prediction task (dotted blue). If the performance evaluation is limited to exclude the warnings that occur within 6 hours of full weaning of the patient, the AUC performance drops to 0.83 for the development data and 0.825 for the validation data.

Plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-23. The dotted blue lines show the performance obtained by using the RAS model output as a proxy to predict the same outcome as PWLM.

Figure 6-24 shows the prediction context for all predictions, including ones that did not satisfy inclusion criteria (left), and the prediction context for patients that did satisfy the inclusion criteria (right). In addition to the PWLM predictions, the figure also shows *survival* predictions from the RAS model.

Table 6.10: PWLM Hosmer-Lemeshow C probability deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.039	612	702	17425	17335	18037
2	(0.1,0.2]	0.146	1187	1172.4	6834	6848.6	8021
3	(0.2,0.3]	0.249	1432	1565.1	4860	4726.9	6292
4	(0.3,0.4]	0.348	1586	1766.7	3488	3307.3	5074
5	(0.4,0.5]	0.448	1690	1805.6	2339	2223.4	4029
6	(0.5,0.6]	0.547	1584	1744.6	1605	1444.4	3189
7	(0.6,0.7]	0.646	1354	1472.1	926	807.9	2280
8	(0.7,0.8]	0.749	1091	1180.5	486	396.5	1577
9	(0.8,0.9]	0.839	501	559.1	165	106.9	666
10	(0.9,1]	0.916	28	40.3	16	3.7	44

$\chi^2 = 238.13,\ d.f. = 10;\ p = 0.000$

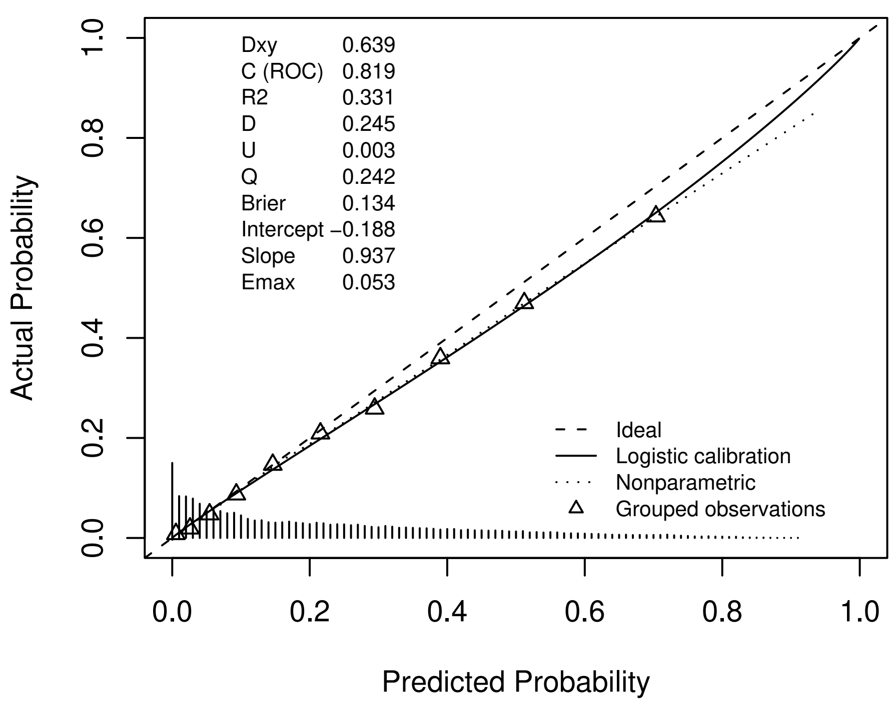


Figure 6-21: PWLM calibration plot

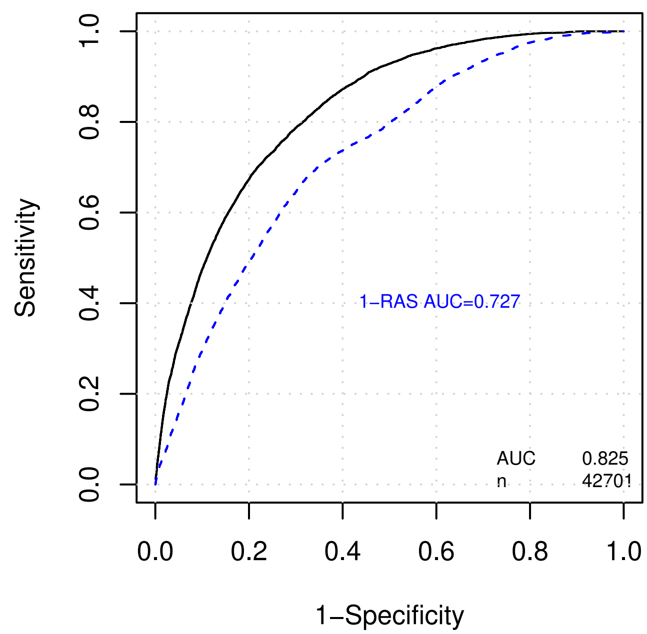


Figure 6-22: `PWLM` ROC curve (validation data).

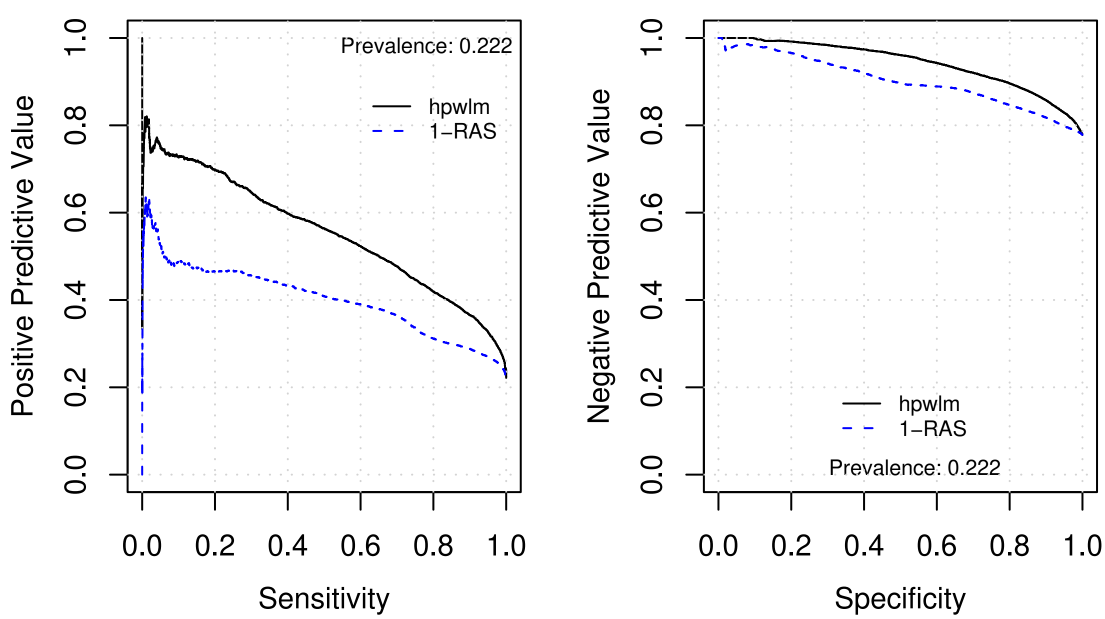


Figure 6-23: `PWLM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (validation data).

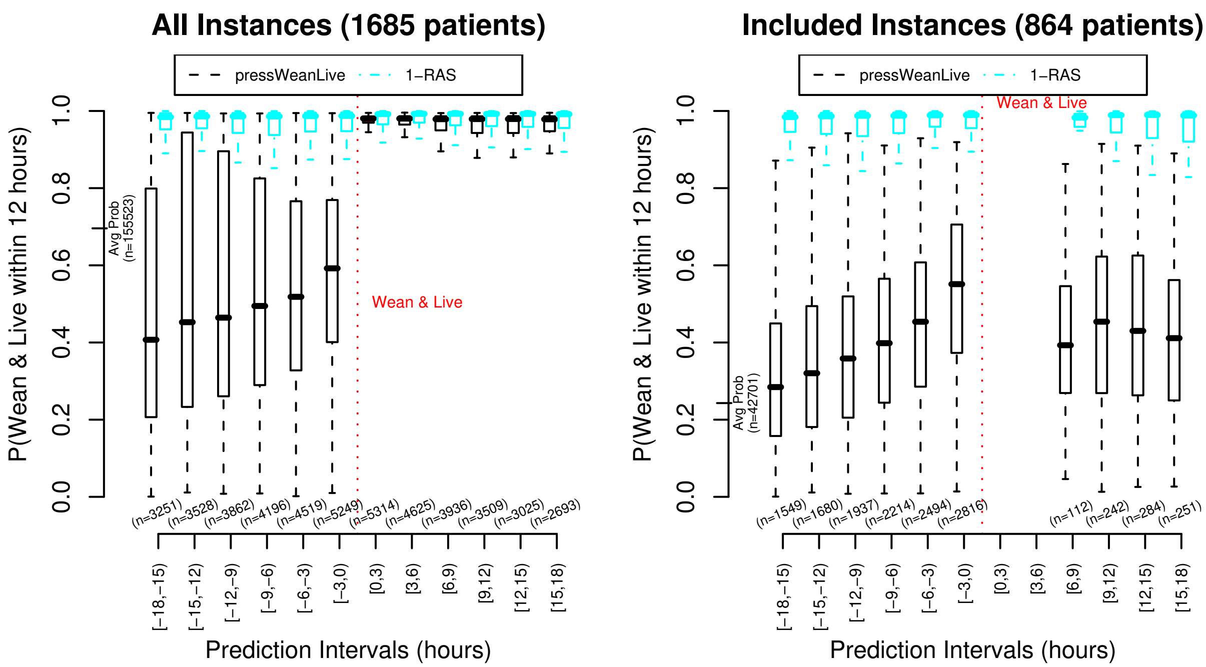


Figure 6-24: `PWLM` prediction context surrounding HDFR (validation data). *Avg Prob*: the mean `PWLM` probability from all patient instances (left) and valid instances (right).

6.2.5 Discussion

The PWLM model developed in this section is quite similar to the previous PWM model but adds the requirement of patient survival. In general, the PWLM model performs better than the PWM model. The PWLM model also performs better on the "hard" predictions as seen by the strong performance (AUC = 0.82 for validation data) when only the estimates between 12 and 6 hours before the successful wean are considered.

The final PWLM model (Model 6.2) includes 32 inputs. Most of the important inputs in the PWLM model also appeared in the PWM model. The coefficients for these important PWM model inputs, weighted for the different outcome, change significantly between the models. For example, in the PWLM model, the most significant input is Milrinone (`Milrinone_perKg_am`) while the standardized pressure measurement (`pressorSum.std_la`), easily the most significant input for the PWM model, is only the seventh most significant input in the PWLM model. This observation is consistent with findings by other researchers indicating that sensitivity to pressors is associated with a decreased mortality rate [53]. In the context of a multivariate model, the rankings of individual variables only explain part of the picture due to synergistic relationships between variables. The influence of the prothrombin time range (`PTrng_sqrt`) seems to be much higher in the PWLM model, but this is mainly caused by offsetting the influence of the INR range variable (`INRrng_sqrt`) which was not present in the PWM model. When the INR variable is manually removed, the PWLM performance decreases slightly and the coefficient and error for the prothrombin time closely match those found in the PWM model.

Notably absent from the PWLM model are the doxacurium and Integrelin inputs found in the PWM model. The PWLM model also contains the pH range (`Art_pHrng`), an indicator noting if the patient is on a mechanical ventilator (`mechVent`) and a categorical variable indicating the level of assistance from the mechanical ventilator (`VentMode_i`). The pH, ventilator, and ventilator mode variables are absent in the PWM model. These variables help the PWLM model identify the general severity of the patient's current condition. For example, if a patient is receiving mechanical ventilation the probability that the patient will be weaned from pressors within 12 hours and live decreases.

The calibration performance of the PWLM model is similar to that of the PWM model. On the development data, the H and C statistics are much better for the PWLM model than the PWM model, but on the validation data they are quite similar. The calibration plot in Figure 6-21 indicates that the PWLM model might be slightly better with a slope closer to 1 and a slightly smaller maximum error (0.053 vs 0.057).

At their respective prediction tasks, the PWLM model has a higher AUC than the PWM model. The difference in AUC indicates that predicting weaning within 12 hours *and* survival is easier than simply predicting weaning within 12 hours. The performance improvement is likely the result of cases where terminal patients were weaned (as a result of withdrawal of care) despite physiologic variables and treatment variables indicating that they would not survive. The absence of fentanyl in the PWLM model likely supports this conclusion.

Given the inclusion of survival in the outcome for the PWLM model, one would expect the RAS model to do reasonably well at predicting the PWLM outcome. As Figure 6-22 shows, this is indeed the case. The RAS predictions result in an AUC of 0.727 compared to the AUC of 0.825 for the specialized PWLM model. The significance level for the difference in performance, however, is still quite high ($p < 0.00001$).

In conclusion, by augmenting the outcome used in PWM (i.e., pressor wean within 12 hours) to require survival, I am able to improve upon the PWM model. The resulting PWLM model is not penalized for missing the "successful" weaning of pressors from patients who do not survive the ICU. As with the PWM model, the PWLM model relies heavily on treatments but also uses a variety of physiological measurements to make predictions. A further improvement to the pressor wean model might be to place stricter limits on the survival requirement used with PWLM to, for example, include pressor weans that were followed by at least 24 hours of survival (or discharged alive within 24 hours). Limiting the survival requirements would help to isolate cases where the patient expired for reasons that were not directly related to the current episode.

6.3 BPWM: Weaning of Intraaortic Balloon Pump

Intraaortic balloon pumps have been used for over 40 years to provide hemodynamic assistance to patients with heart failure. An IABP is an inflatable membrane that is surgically implanted in the descending thoracic aorta via the femoral artery. It functions by using gas to inflate a balloon in the aorta at the onset of cardiac diastole and deflate the balloon at the onset of systole. In [88] Trost et al. describe a variety of indications for IABP insertion, including:

- Cardiogenic shock
- Cardiogenic shock due to ventricular septal rupture or papillary muscle rupture, with resultant mitral regurgitation
- Intractable ventricular arrhythmias
- Post-MI angina or unstable angina refractory to medical therapy
- Heart failure refractory to medical therapy
- Hemodynamic support for "high-risk" catheterization and angioplasty
- Hemodynamic support for high-risk coronary artery bypass grafting
- Myocardial dysfunction from septic shock

Cardiac assistance via an IABP is considered short-term therapy. Before an IABP is initially inserted, the care team typically develops a plan to wean the device [28]. In many cases the endpoint is clear. For example, if the device is used to stabilize the patient until cardiac surgery, the device is removed post-operatively. In other cases the endpoint is more qualitative such as recovery from an myocardial infraction. After multiple days of support, the risk of infection is an significant concern. In many cases the device is removed as a result of other complications such as an ischemic leg or uncontrollable bleeding despite hemodynamic contraindication. While the presence of an IABP indicates serious heart impairment, the patient mix remains quite complex. Furthermore, most of the procedural context for these patients (e.g., high risk coronary artery bypass graft surgery) is lacking in the MIMIC II data under consideration and there is likely a significant amount of variance between cardiologist's decisions regarding when and how to wean [28]. Despite these concerns, there are presumably still differences between patients who need IABP assistance and those who no longer need IABP assistance.

The IABP frequency can typically be set between 1:1 (one inflation per cardiac cycle) and 1:8, as required by the patient's hemodynamic status. One important concern with inactive balloons is thrombosis, and the administration of heparin to keep thromboplastin time to 50 to 70 seconds is common [88]. As a result, it is commonly suggested to remove an inactive IABP device within 30 minutes to prevent thrombosis complications. For patients with an IABP it is important to carefully observe the hematocrit, hemoglobin and platelet counts as thrombocytopenia (shortage of platelets in the blood) may result from the device and/or the administration of heparin [88].

Despite the long history of the device and common usage[5], few studies specifically address the best protocol for IABP weaning. It is generally accepted practice to remove inotropic and vasoactive medications prior to IABP weaning [54]. Pressor drugs commonly include dopamine, dobutamine, and Levophed (norepinephrine). Two methods exist for weaning an IABP: (1) volume reduction and (2) frequency reduction. In volume reduction the volume of gas used to inflate the balloon is reduced in 20-25% increments. In contrast, frequency reduction gradually decreases the ratio of cardiac cycles that receive assistance thereby exposing the heart to a wide range of inter-beat afterloads. Clinical indicators suggested by Bolooki in 1984 for determining if a patient is ready to be weaned from and IABP include: (1) Mean blood pressure above 70 mmHg, (2) Systolic blood pressure or diastolic augmentation above 90 mmHg, (3) PCWP less than 18 mmHg, and (4) Cardiac index above 2.2 L/min/m^2 [4]. Ideally, as a patient improves, the diastolic pressor augmentation decreases as the patient's stroke volume increases and the IABP blood displacement is a smaller fraction of cardiac output.

Despite a dearth of formal studies, volume reduction appears to be the preferred method as the frequency reduction method is more abrupt [54]. But in most ICUs a combination of the two methods is common. Some have suggested volume reduction at time intervals between 15 and 30 minutes [34]. During a weaning process, the nursing staff closely watch patients for changes in urine output, temperature, sensorium, hemodynamic profile, and heart or lung sounds as well as any chest pain [2]. If the patient starts to decompensate as a result of the weaning process, the IABP is typically returned to its maximum assist setting. If indicated by the hemodynamic parameters, another weaning attempt can be made after a few hours.

In this section I develop a model to predict the successful removal of IABP support. I refer to this model as the IABP wean model (`BPWM`).

6.3.1 Data and Patient Inclusion Criteria

In the MIMIC II dataset under consideration, most periods of IABP therapy last about two days (median length of 38 hours). For particularly unstable patients, however, IABP therapy can last much longer. Figure 6-25 provides a histogram of IABP episode lengths. The figure shows some tendency for the therapy episodes to last a multiple 24 hours. Weaning attempts require close attention to the patient

[5]The 2005 National Hospital Discharge Survey estimated 40,000 patients received an IABP in 2005 [12].

Table 6.11: BPWM data

	Count
IABP patients	595
Included instances	49046
IABP weaned within 12 hours	12162
IABP not weaned with 12 hours	36884

and the timing is influenced by caregiver-workload.[6] The process of weaning from an IABP occurs gradually as the caregivers first remove inotropic and vasoactive drugs and then determine how the patient's heart responds to the increased afterload of less IABP assistance. The MIMIC II data contain the IABP frequency information for patients on an IABP (ratio of assisted beats to total number of beats). While values of 1:1, 1:2, 1:3, and 1:4 occur, the 1:4 frequency is quite rare and most patients are completely weaned directly from 1:2 or after a short time at 1:3. The median duration of the low frequency episodes (including 1:3 and 1:4) is about one hour. The median length for 1:2 episodes is about 10 hours. For the purposes of the BPWM model, we chose 12 hours for an early warning window.

Properly annotating the removal of the IABP was somewhat challenging. For most patients, the time of IABP removal was denoted with a null value in the ChartEvents entry. Not all patients, however, conformed to this pattern and null values do not always indicate IABP removal. Many of the IABP patients undergo surgeries that last several hours, during which no data is available. I used the following scheme to annotate patients with an IABP: periods of up to 8 hours between IABP chart entries were merged if no null value existed during the period. If a null existed, and preceded the next IABP entry by at least 3 hours, the null ended the current IABP episode and subsequent IABP entries were considered separate IABP episodes (re-insertions). In general, the IABP was charted quite frequently with an average duration between chart entries of about 40 minutes.

The only inclusion criterion that I required for the BPWM model training and validation was IABP assistance that lasted at least two hours. Table 6.11 provides a summary of the patients with an IABP that were used to develop the BPWM model.

6.3.2 Outcome

The outcome of interest for the BPWM is the successful removal of IABP assistance within 12 hours. Figure 6-26 demonstrates a typical patient that receives an IABP and is then weaned from it about 45 hours later.

[6]While patients are occasionally weaned during the night, most patients are weaned between 7:00 am and 7:00 pm, with the highest frequency occurring during late morning and early afternoon.

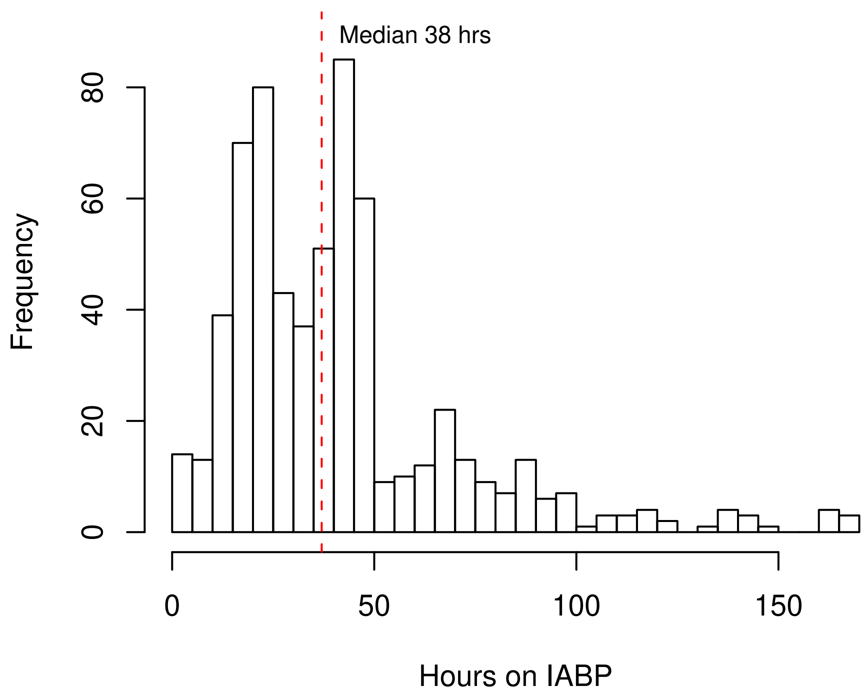


Figure 6-25: IABP episode lengths

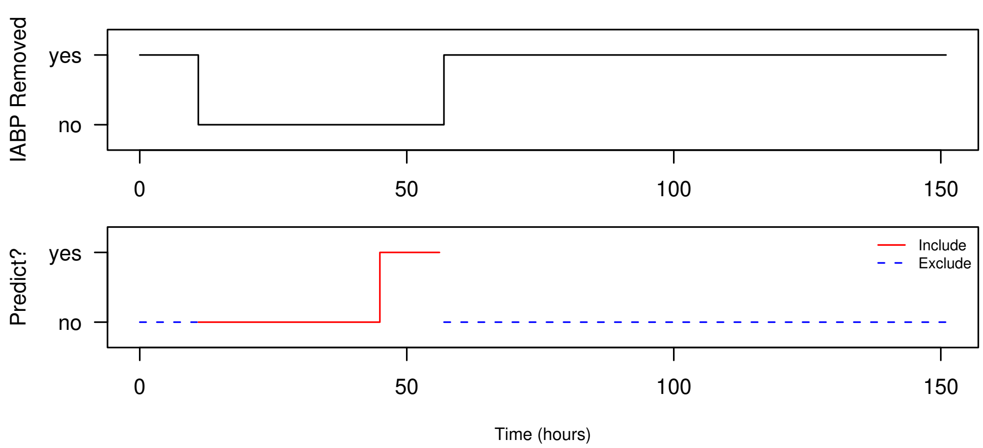


Figure 6-26: `BPWM` example annotations for `Subject_ID` 354

6.3.3 Model Development

To develop the predictive model for the IABP removal task described above, I again follow the methodology described in Chapter 4. The model selection process is described this section. I also provide an overview of the resulting logistic regression model and I describe the model's performance on the training data.

Model Selection

Candidate variables were initially ranked against the outcome variable (successful removal of IABP). Variables with a p-value greater than 0.05 were excluded. Furthermore, if multiple variables were strongly correlated (Spearman's rank correlation test > 0.8) the best univariate variable was retained. After the initial screening of the variables, variable selection for the BPWM model was based on the best 20 variables from each of the top 4 of the 5 cross-validation folds (the individual cross validation plots are provided in Appendix F). When combined, the best 30 variables from the top 4 folds resulted in 58 candidate variables. Figure 6-3 shows the AUC that results from gradually increasing the AIC backward elimination threshold and thereby forcing additional variables to be greedily dropped.

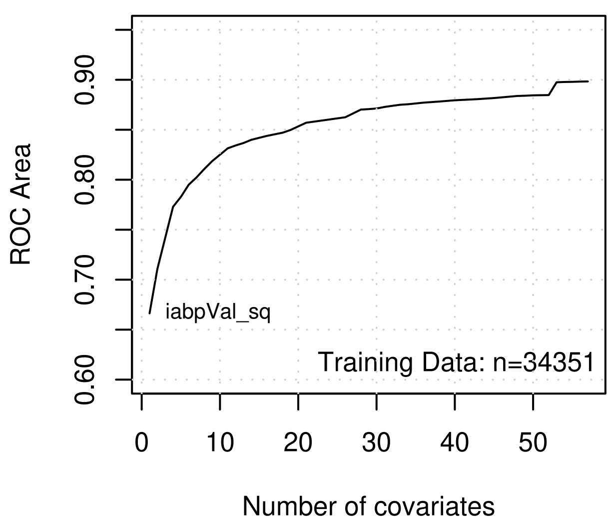


Figure 6-27: BPWM model selection (all development data)

The final model was trained using the top 31 variables obtained from cross validation. The details for the BPWM model are shown in Model 6.3.

Model 6.3 Final BPWM model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
18039	5e-07	7948.34	31	0	0.874	0.748
Gamma	Tau-a	R2	Brier			
0.749	0.322	0.501	0.129			

	Coef	S.E.	Wald Z	P
iabpVal_sq	-1.940e-01	5.452e-03	-35.58	0
Anticoagulant	-8.523e-01	4.965e-02	-17.17	0
CrdIndx_i	-4.104e+00	2.664e-01	-15.41	0
X24hBal	-2.054e-04	1.364e-05	-15.05	0
Milrinone_perKg_am	-2.273e+00	1.674e-01	-13.58	0
PAPsd_sqrt	-3.555e-01	2.843e-02	-12.50	0
ShockIdx_am	-1.856e+00	1.514e-01	-12.26	0
cumPressorTime_lam	-2.475e-01	2.247e-02	-11.02	0
pressD24	-9.921e-01	1.035e-01	-9.58	0
FullCode	-9.099e-01	9.513e-02	-9.56	0
mechVent	-5.130e-01	5.417e-02	-9.47	0
alloutput_lam	-1.776e-01	1.951e-02	-9.10	0
Levophed_perKg_la	-7.665e-02	8.648e-03	-8.86	0
Glucose_la	-6.450e-01	7.534e-02	-8.56	0
BUN_Slope_1680	-5.215e+01	6.194e+00	-8.42	0
PTT_Slope_1680	-8.346e+00	1.014e+00	-8.23	0
Dobutamine_perKg	-1.264e-01	1.602e-02	-7.89	0
Input_60rng_sqrt	-1.054e-02	1.368e-03	-7.71	0
ventLen_lam	-1.790e-01	2.372e-02	-7.55	0
Dopamine_perKg_sqrt	-2.458e-01	3.335e-02	-7.37	0
Mg_Slope_1680	4.333e+02	5.525e+01	7.84	0
INRrng_i	1.083e-04	1.207e-05	8.97	0
pressD01	4.707e-01	4.972e-02	9.47	0
GCS_Slope_1680_sqrt	7.669e+00	7.790e-01	9.85	0
PT_i	3.396e+01	3.229e+00	10.52	0
SBP_Slope_1680	2.024e+01	1.858e+00	10.89	0
Vasopressin_i	1.174e-04	1.068e-05	10.99	0
INR_Slope_1680_sq	1.609e+05	1.339e+04	12.02	0
GCSrdv	9.518e-02	6.920e-03	13.75	0
Intercept	8.922e+00	5.759e-01	15.49	0
totIn_sqrt	2.509e-02	1.566e-03	16.02	0
LOSBalrng_sqrt	2.455e-02	9.026e-04	27.20	0

Table 6.12: BPWM Hosmer-Lemeshow H risk deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-2	[3.40e-05,0.0479)	0.021	83	77.1	3525	3530.9	3608
3	[4.79e-02,0.0855)	0.067	108	120.9	1696	1683.1	1804
4	[8.55e-02,0.1336)	0.109	176	196.2	1628	1607.8	1804
5	[1.34e-01,0.2087)	0.171	318	308.2	1486	1495.8	1804
6	[2.09e-01,0.3004)	0.252	491	453.8	1313	1350.2	1804
7	[3.00e-01,0.4248)	0.36	648	649.4	1156	1154.6	1804
8	[4.25e-01,0.6176)	0.518	939	934	865	870	1804
9	[6.18e-01,0.8148)	0.717	1238	1293.8	566	510.2	1804
10	[8.15e-01,0.9999]	0.906	1666	1633.5	137	169.5	1803

$\chi^2 = 24.15,\ d.f. = 7;\ p = 0.001$

Development Validation

To validate the BPWM model, I examine calibration performance and AUC performance. In addition, I also plot the positive predictive value (PPV) versus sensitivity and the negative predictive value (NPV) versus specificity. Table 6.12 shows the deciles used for the Hosmer-Lemeshow H statistic and Table 6.13 shows the deciles used for the Hosmer-Lemeshow C statistic. The classification performance of the BPWM model on the training data is shown by the ROC curve in 6-28. In addition, the PPV versus sensitivity and the NPV versus specificity are plotted in Figure 6-29

Figure 6-30 shows the context surrounding successful IABP weans for patients who lived (left) and patients who died (right). Similarly, Figure 6-31 shows the context surrounding successful IABP weans for all predictions, ignoring the inclusion criteria (left) and only patients that satisfied the inclusion criteria (right). The need to reinsert an IABP is rare. There do appear to be a few patients who require the IABP to be reinserted after weaning as shown by the small number of predictions after IABP wean events in Figure 6-30.[7]

As an illustration of predictions for an individual patient, Figure 6-32 shows the predictions for the patient used to demonstrate the annotation process (Figure 6-26).

[7]Many of the apparent re-insertions shown to the right of the IABP wean event are incorrect. A number of gaps greater than 6-hours exist even when no sign of an IABP wean is present in the nursing notes. In fact, many of the longer gaps correspond to times when the patient was moved between care units (e.g., CCU to CSRU). Manual review of individual patients could reduce such mislabeled cases.

Table 6.13: BPWM Hosmer-Lemeshow C probability deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.042	247	251.4	5744	5739.6	5991
2	(0.1,0.2]	0.145	387	407.9	2431	2410.1	2818
3	(0.2,0.3]	0.247	542	495.8	1469	1515.2	2011
4	(0.3,0.4]	0.35	518	527.1	990	980.9	1508
5	(0.4,0.5]	0.447	485	472.5	573	585.5	1058
6	(0.5,0.6]	0.549	478	477.3	391	391.7	869
7	(0.6,0.7]	0.65	578	607.4	357	327.6	935
8	(0.7,0.8]	0.75	627	660.8	254	220.2	881
9	(0.8,0.9]	0.848	844	824.3	128	147.7	972
10	(0.9,1]	0.946	961	942.5	35	53.5	996

$\chi^2 = 28.71,\ d.f. = 8;\ p = 0.000$

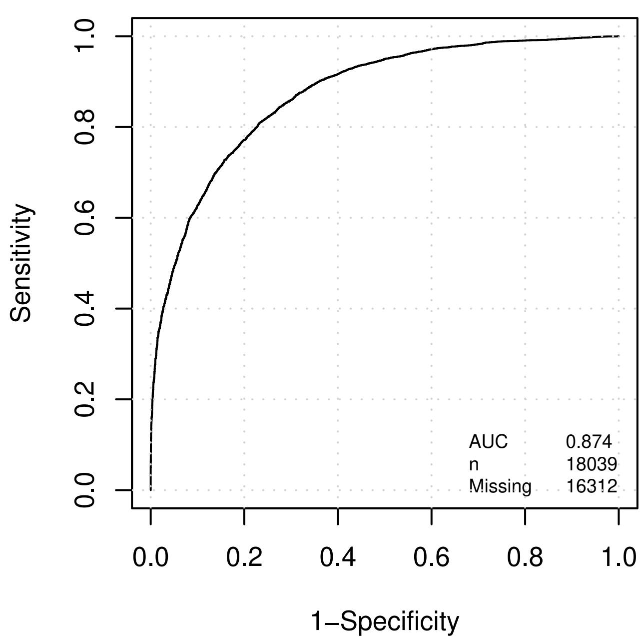


Figure 6-28: BPWM ROC curve (development data).

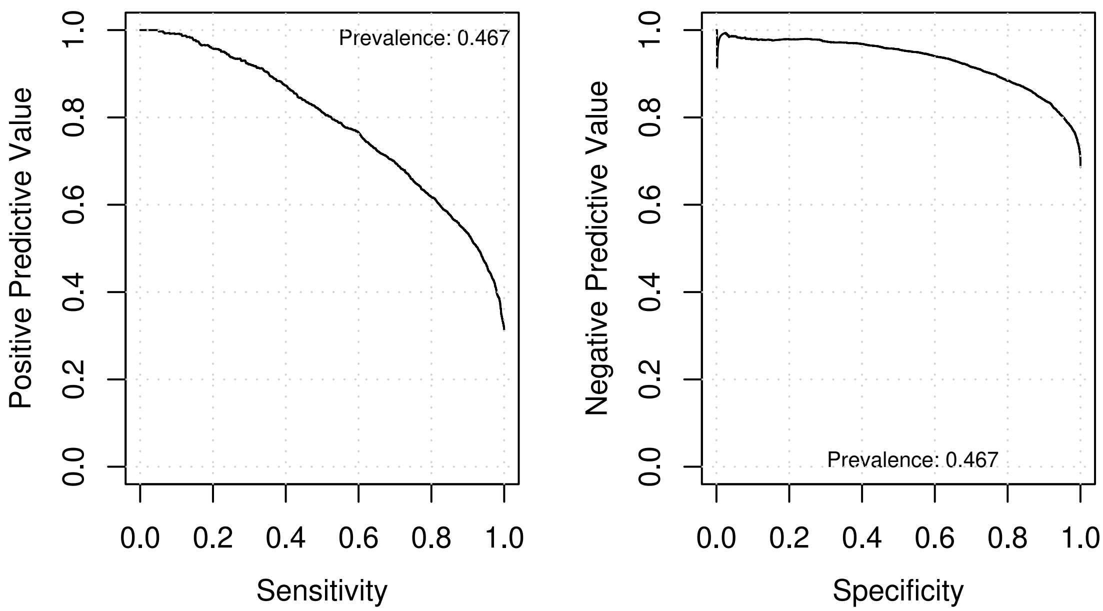


Figure 6-29: BPWM positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (development data).

6.3.4 Model Validation

As a final step, I validate the BPWM model on the separate validation data. To evaluate calibration, Table 6.14 and Table 6.15 provide the deciles used for the Hosmer-Lemeshow statistics. A plot of the calibration — actual probability versus estimated probability — is shown in Figure 6-33. The BPWM classification performance is summarized by the ROC curve in Figure 6-34. For comparison purposes, Figure 6-34 includes a curve generated by the RAS model developed in the previous chapter applied to the same prediction task (dotted blue).

Plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-35. The dotted blue lines show the performance obtained by using the RAS model output as a proxy to predict the same outcome as BPWM.

Finally, as done previously with the development patients, the context surrounding successful IABP wean is examined for the validation data. Figure 6-36 shows the context surrounding successful weans for patients who survived (left) and patients who died (right). In addition to the BPWM predictions, the figure also shows *survival* predictions from the RAS model. Figure 6-37 shows the prediction context for all predictions, including ones that did not satisfy inclusion criteria (left), and the prediction context for patients that did satisfy the inclusion criteria (right).

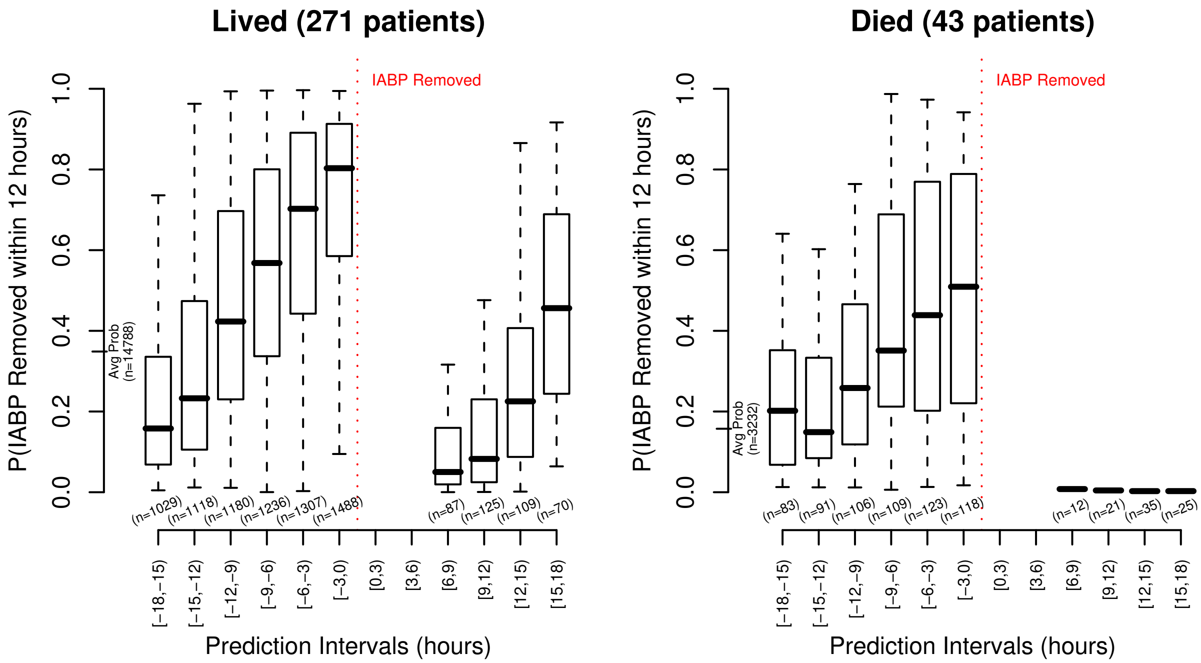


Figure 6-30: `BPWM` prediction context surrounding IABP removal (development data). *Avg Prob*: the mean `BPWM` probability from all patients who lived (left) and died (right).

Table 6.14: `BPWM` Hosmer-Lemeshow H risk deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-3	[1.81e-05,0.0865)	0.038	233	91.8	2198	2339.2	2431
4	[8.65e-02,0.1280)	0.107	105	87	706	724	811
5	[1.28e-01,0.1872)	0.156	136	126.2	674	683.8	810
6	[1.87e-01,0.2740)	0.229	210	185.4	600	624.6	810
7	[2.74e-01,0.4153)	0.334	271	270.7	540	540.3	811
8	[4.15e-01,0.6075)	0.509	416	412.4	394	397.6	810
9	[6.07e-01,0.8237)	0.71	562	575.5	248	234.5	810
10	[8.24e-01,0.9965]	0.915	720	741.3	90	68.7	810

$\chi^2 = 243.37,\ d.f. = 8;\ p = 0.000$

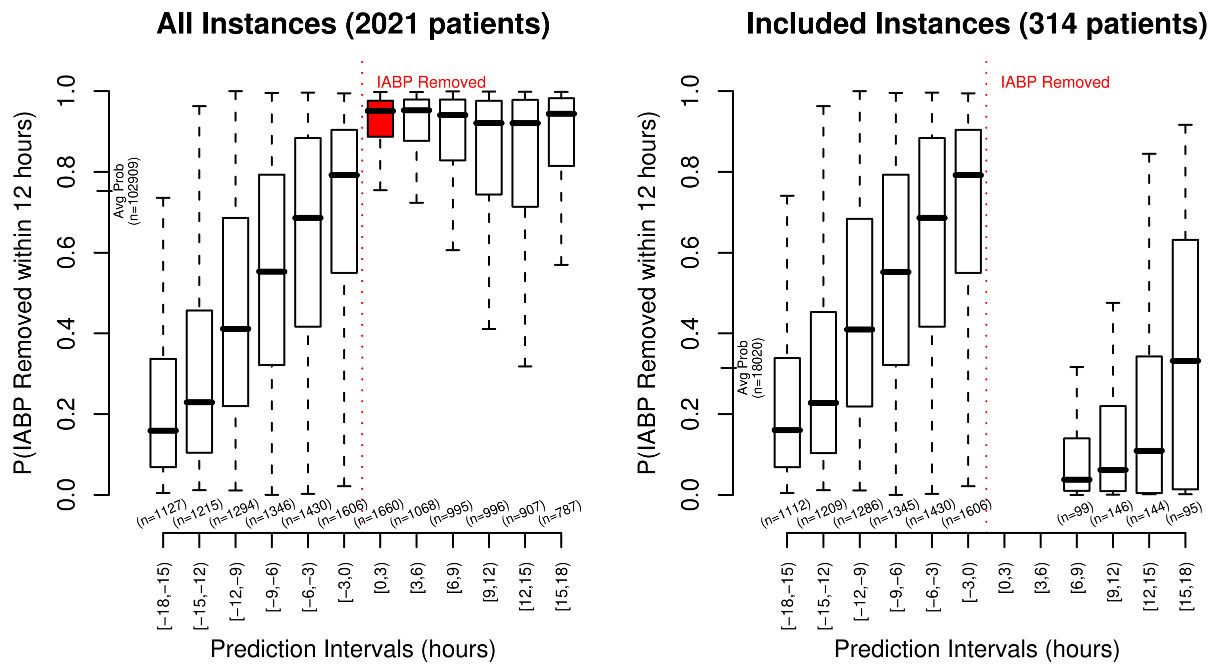


Figure 6-31: BPWM prediction context surrounding IABP removal (development data). *Avg Prob*: the mean BPWM probability from all patient instances (left) and valid instances (right).

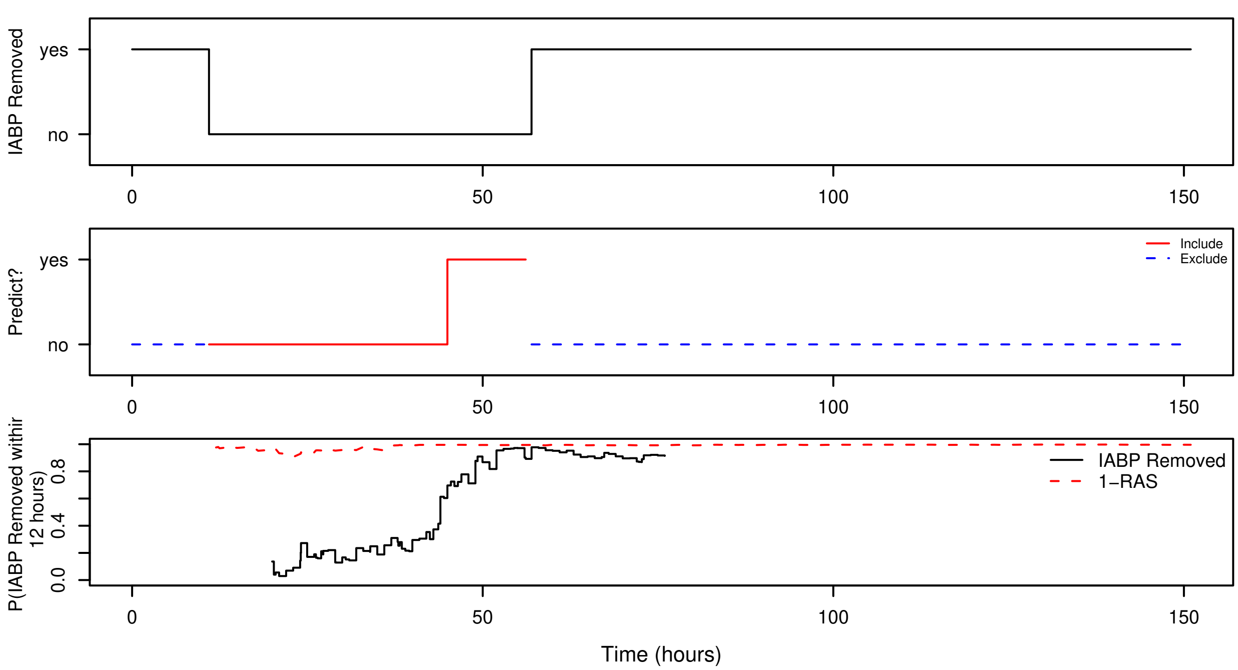


Figure 6-32: BPWM annotations for Subject_ID 354 with BPWM and RAS predictions

Table 6.15: BPWM Hosmer-Lemeshow C probability deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.043	259	114.7	2417	2561.3	2676
2	(0.1,0.2]	0.143	259	216.8	1254	1296.2	1513
3	(0.2,0.3]	0.248	223	218	655	660	878
4	(0.3,0.4]	0.343	191	187.7	356	359.3	547
5	(0.4,0.5]	0.45	189	188.3	230	230.7	419
6	(0.5,0.6]	0.548	229	231.8	194	191.2	423
7	(0.6,0.7]	0.649	269	262.8	136	142.2	405
8	(0.7,0.8]	0.749	243	260.8	105	87.2	348
9-10	(0.8,1]	0.905	791	809.4	103	84.6	894

$\chi^2 = 209.43,\ d.f. = 9;\ p = 0.000$

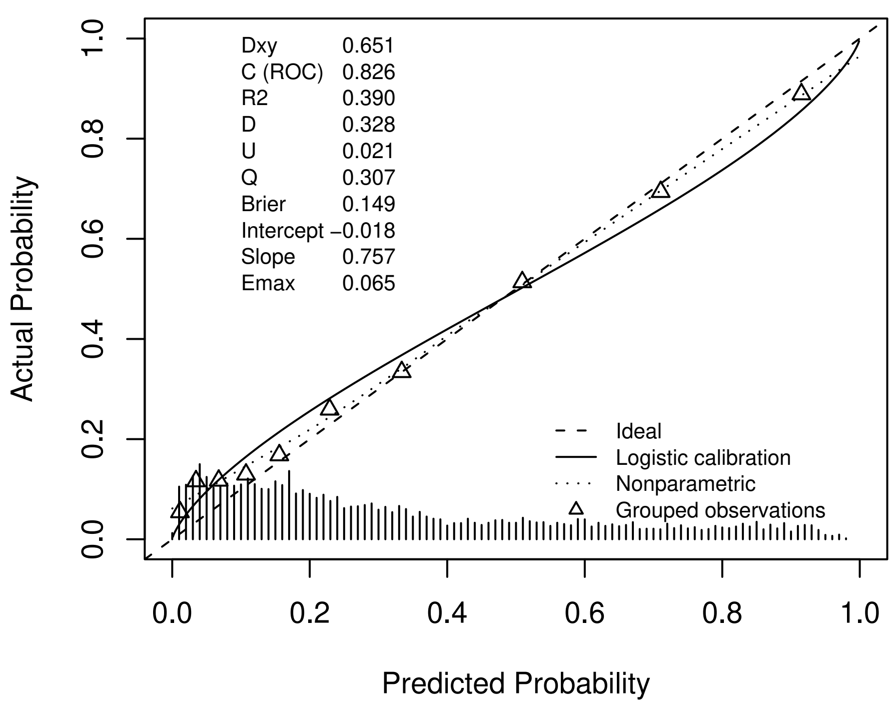


Figure 6-33: BPWM calibration plot

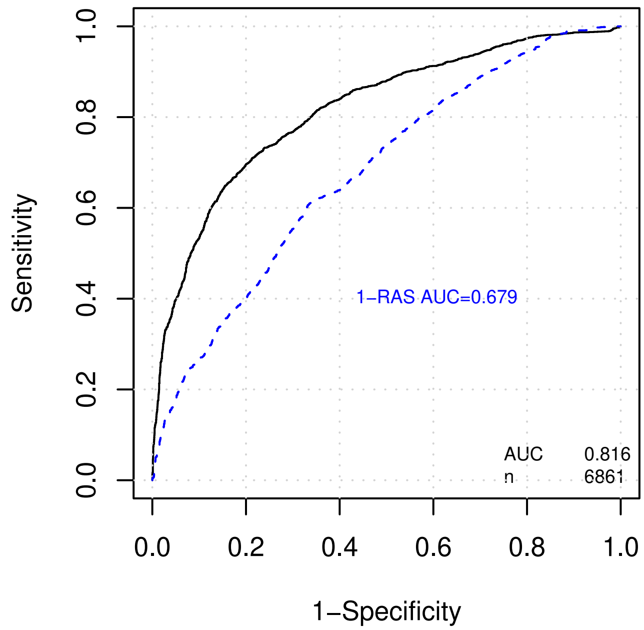


Figure 6-34: BPWM ROC curve (validation data).

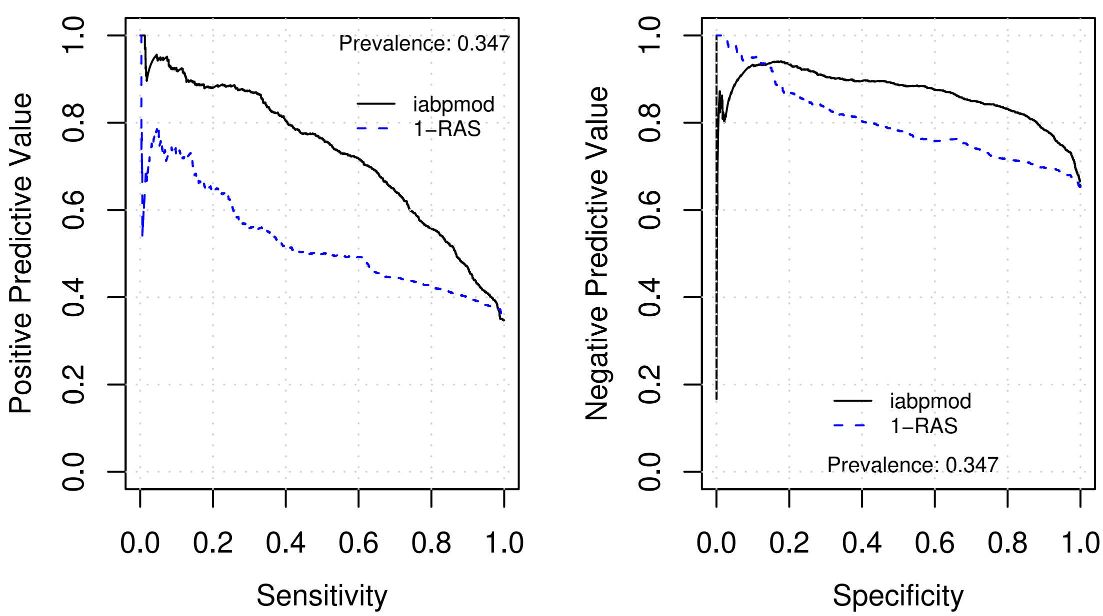


Figure 6-35: BPWM positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (validation data).

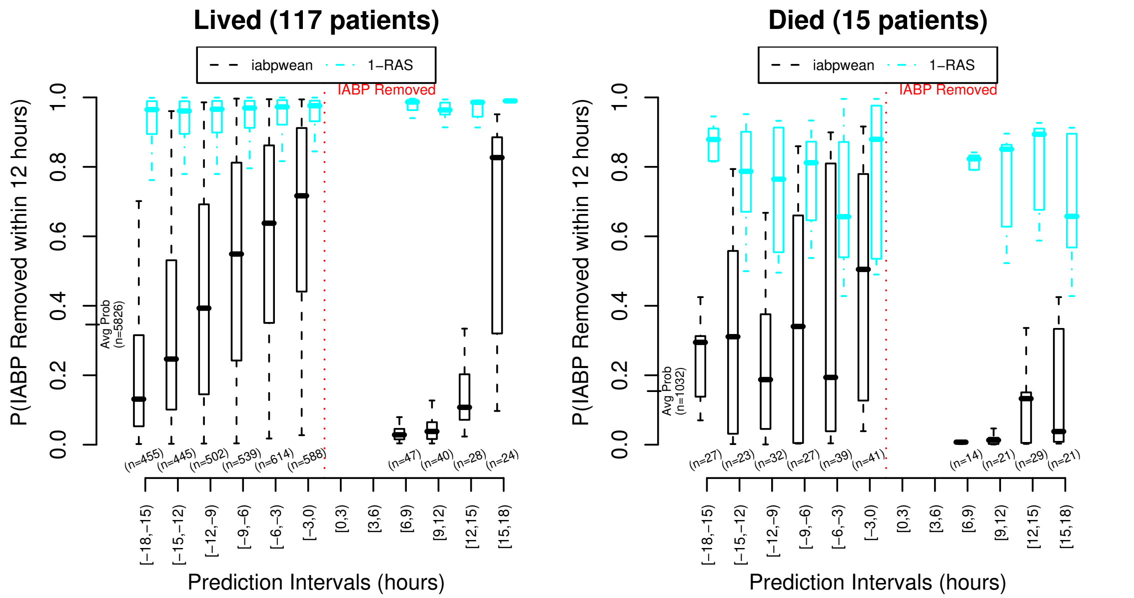


Figure 6-36: `BPWM` prediction context surrounding IABP removal (validation data). *Avg Prob*: the mean `BPWM` probability from all patients who lived (left) and died (right).

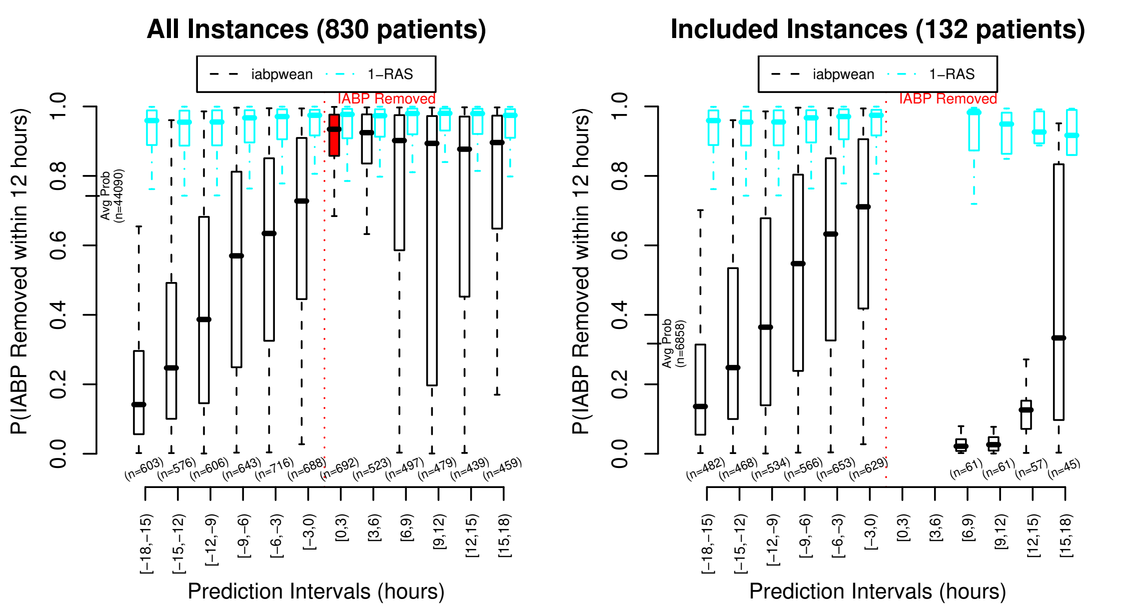


Figure 6-37: `BPWM` prediction context surrounding IABP removal (validation data). *Avg Prob*: the mean `BPWM` probability from all patient instances (left) and valid instances (right).

6.3.5 Discussion

The BPWM model developed in this section attempts to predict the successful removal of an intraaortic balloon pump (IABP) within 12 hours. As the process of withdrawing an IABP typically proceeds over a significant amount of time (e.g., 24 hours) and is typically accompanied by withdrawal of pressors, there are a number of indicators that IABP weaning is taking place. In addition, before the pump is physically removed, anticoagulants are typically given to the patient to lower the risk of an embolism or other related complications during removal. With standard steps called for by IABP protocols, the prediction task for the BPWM model should be reasonably straightforward. The final stages of the weaning process are often the most interesting as the ability of the patient's heart to pump without assistance is evaluated. A prediction window of 12 hours focuses on these final stages.

In the BPWM model the most important indicator for IABP weaning is the frequency of IABP assistance (`iabpVal_sq`). The IABP frequency variable increases as more IABP assistance is provided; assistance on every beat (1:1) is assigned an ordinal value of "4" whereas the lowest assist rate (1:4) is coded as "1" and no assistance is coded as "0". As expected, patients that are receiving frequent IABP assistance are less likely to be weaned in the following 12 hours. In contrast, less frequent assistance is a sign that a weaning attempt has started — during a weaning attempt, it is often necessary to revert back to higher assist frequencies if the patient responds poorly. Other highly predictive variables in the BPWM model include the length of stay fluid balance range (`LOSBalrng_sqrt`), the administration of anticoagulants (`Anticoagulant`), the total fluid input (`totIn_sqrt`), and the cardiac index (`CrdIndx_i`).

With the exception of vasopressin, the presence of pressors decreases the probability of an IABP wean. The total time that the patient has spent on pressors, `cumPressorTime_lam`, also decreases the probability that the patient will be weaned. The presence of pressors generally reflects instability in the patient and it is not surprising that spending a long time on pressors and receiving high pressor dosages correlate with continued IABP therapy.

The BPWM model also includes more long-term slope variables than previous models. The slope variables include the blood urea nitrogen 28-hour slope (`BUN_Slope_1680`), the partial thromboplastin time 28-hour slope (`PTT_Slope_1680`), the magnesium 28-hour slope (`Mg_Slope_1680`), the Glasgow Coma Scale 28-hour slope (`GCS_Slope_1680_sqrt`), the systolic blood pressure 28-hour slope (`SBP_Slope_1680`), and the international normalized ratio 28-hour slope (`INR_Slope_1680_sq`). The individual coefficients indicate that an increasing BUN or PTT decreases the probability of a successful IABP wean within 12 hours. In contrast, the coefficients for the Mg, GCS, SBP, and INR slope variables indicate that an increasing trend for these variables increases the probability of a successful IABP wean within 12 hours. The physiologic interpretation of the BPWM model trend variables generally reflect relative improvement in blood circulation or changes in blood coagulation. Trends in these variables are predictive of imminent IABP removal.

Even with the small number of IABP validation patients, the `BPWM` predictions from each of the three episodes preceding the IABP removed event are statistically significant between the lived and died cases ($p < 0.001$). The `BPWM` predictions are also statistically significant between consecutive intervals that precede IABP removal. In contrast, the `RAS` predictions do not generally show statistically significant differences between consecutive intervals preceding IABP removal.

In conclusion, the `BPWM` does well at tracking the progression of an IABP wean. Unlike many of the other models considered in this chapter, the protocol surrounding an IABP wean is fairly clear and the presence of an IABP limits our development and validation population to a specific category of patients. The range of insults necessitating cardiac assistance, however, yields a complex patient mix that raises questions about the direct clinical utility of this model. Furthermore, while the `BPWM` model contains a number of interesting relations, several recent advances in IABP alternatives, such as percutaneous left ventricular assist devices, may finally replace IABPs and render the `BPWM` discussed here largely irrelevant [28].

6.4 SSOM: Onset of Septic Shock

Sepsis occurs when a patient demonstrates a systemic inflammatory response to an infection. Common usage of the term *sepsis* varies, but recent efforts have tried to clarify the term. A consensus conference of the American College of Chest Physicians and the Society of Critical Care Medicine met in 1991 and subsequently published a report seeking to clarify the definition of sepsis, severe sepsis and septic shock [5]. They proposed the systemic inflammatory response syndrome (SIRS) as an essential component of sepsis, defining sepsis as the presence of SIRS with a confirmed infection. This includes a spectrum of more severe conditions, including (1) severe sepsis, or sepsis and evidence of end-organ dysfunction as a result of hypoperfusion and (2) septic shock, where the patient has severe sepsis and persistent hypotension despite adequate fluid resuscitation and resulting tissue hypoperfusion.

Severe sepsis and septic shock are grave conditions. Mortality rates ranging between 28% and 50% are commonly reported for severe sepsis and severe sepsis represents an increasing portion of hospitalizations and hospital mortality over the past decade [55, 93, 15].

In this section I develop a model to predict septic shock. This model will be referred to as the septic shock onset model (SSOM). Previous work by Shavdia [81] trained a classifier using a subset of MIMIC II patients with ICD-9 codes indicating septic shock. I use similar definitions, but do not place the same limitations on my training data. Instead, my work focuses solely on automatically annotated Systemic Inflammatory Response Syndrome (SIRS) episodes as determined by the charted MIMIC II data. I define septic shock as persistent hypotension despite fluid resuscitation (HDFR). My model seeks to predict the transition from SIRS *without* HDFR to SIRS *with* HDFR. Without confirmation that individual cases of HDFR were secondary to sepsis (i.e., a confirmed infection), other insults that result in SIRS and cause HDFR, which may not be associated with sepsis, are inevitably included by our broad definition.

6.4.1 Data and Patient Inclusion Criteria

For training and validating my model, I included patient episodes with SIRS but no evidence of HDFR. The definition of SIRS and HDFR are provided below.

For SIRS, I use the consensus definition developed by the American College of Chest Physicians and the Society of Critical Care Medicine conference in 1991 [5]. The SIRS definition requires the presence of at least two abnormalities among the variables described in Table 6.16. An abnormal value is indicated by a value less than the low threshold or greater than the high threshold. Following Shavdia's methodology, abnormal values were required to persist for at least 5 hours for inclusion and SIRS intervals within 6 hours of each other were merged together to form a single SIRS episode.

Table 6.16: SIRS variables and their normal ranges. A value is considered abnormal if it falls below the low threshold or above the high threshold. Two or more abnormal values indicates SIRS.

Variable	Low Thresh	High Thresh
Temperature	36 °C	38 °C
Heart Rate	-	90 bpm
Respiratory Rate	-	20 bpm
WBC Count	4000/μL	10000/μL

HDFR was indicated by sustained hypotension in the presence of substantial fluid input or, alternatively, increased pressor infusion without regard to blood pressure. Hypotension was defined as a systolic blood pressure falling below 90 mmHg for at least 30 minutes.[8] Substantial fluid input was defined as at least 600 mL of input over the period ranging from one hour prior to the start of the hypotensive episode through the midpoint of the hypotensive region. Pressor infusion was defined as an increase of at least 20% in vasopressors (Vasopressin, Neo-Synephrine, Levophed, Dopamine, Epinephrine) or inotropic agents (Dobutamine, Amrinone, Milrinone) between the average preceding three dosages (limited to 120 minutes) and the maximum subsequent dose within 18 hours.

To emphasize the *onset* of septic shock, only episodes of SIRS without *HDFR* were included. For each SIRS instance, if HDFR occurred within 12 hours the instance was labeled positive for onset of septic shock. No minimum interval was required between the onset of SIRS and the onset of HDFR. For example, if the start of a SIRS episode occurred concurrently with evidence of HDFR, no prediction annotations were made.

The prediction window of 12 hours was often not fully utilized as the classification of SIRS, on average, only preceded HDFR by 5.4 hours. The average HDFR warning length excluded a large number of cases where the classification of HDFR and SIRS occurred simultaneously. Figure 6-38 shows the distribution of lengths for the warning annotations. As the figure shows, many HDFR warning episodes were limited by SIRS periods that were less than 12 hours in length.

After annotating the final dataset, a number of instances were excluded that did not contain episodes of SIRS. Table 6.17 provides a summary of the included data.

6.4.2 Outcome

The outcome I explore was the progression from SIRS *without* HDFR to SIRS *with* HDFR within a 12 hour period. By focusing on periods up to 12 hours prior to HDFR, I emphasize trends and other patterns that might provide early warning of HDFR.

[8]While blood pressure is typically only documented hourly, documentation frequency typically increases (e.g., every 5 minutes) during hypotensive episodes (See Figure 3-1).

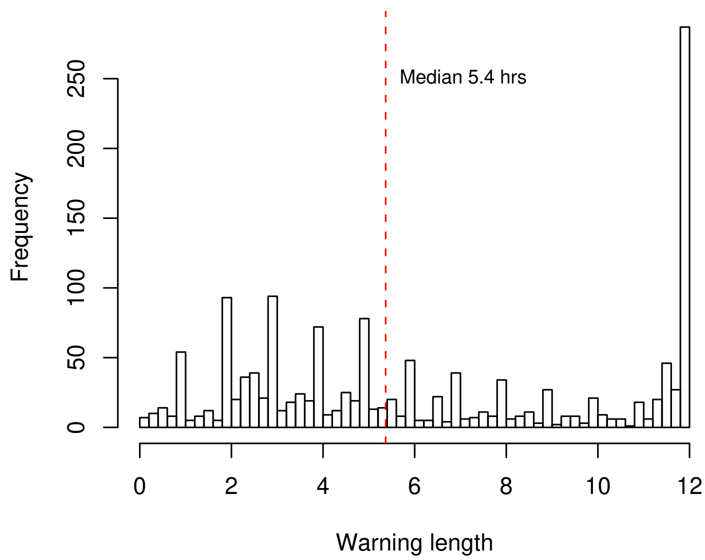


Figure 6-38: Septic shock onset warning lengths

Table 6.17: SSOM data

	Count
SIRS patients	5449
HDFR patients	2802
Included instances	237412
HDFR 12-hour warn	22482
SIRS without HDFR within 12 hours	214930

To illustrate our inclusion criteria and annotations, Figure 6-39 provides an example from `Subject_ID` 13325. The top plot in Figure 6-39 marks the episodes of SIRS *and* HDFR. The lower plot shows the automatically annotated 12-hour warnings for the top plot; periods without SIRS or with both SIRS and HDFR are indicated by the dashed blue line.

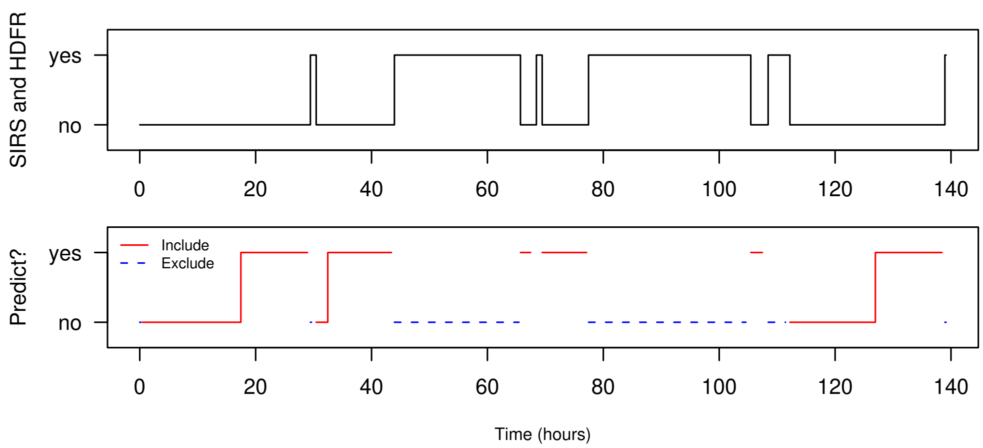


Figure 6-39: `SSOM` example annotations for `Subject_ID` 13325

6.4.3 Model Development

To develop `SSOM`, I follow the methodology described in Chapter 4. I first describe the model selection process and the resulting logistic regression model. Following the model selection description, I describe validation on the training data.

Model Selection

Candidate variables were initially ranked against the outcome variable (onset of septic shock). Variables with a p-value greater than 0.05 were excluded. Furthermore, if multiple variables were strongly correlated (Spearman's rank correlation test > 0.8) the best univariate variable was retained. After the initial screening of the variables, variable selection for the `SSOM` model was based on the best 30 variables from each of the 5 cross-validation folds (the individual cross validation plots are provided in Appendix F). When combined, the best 30 variables from the 5 folds resulted in 56 unique candidate variables. Figure 6-40 shows the AUC that results from gradually increasing the AIC backward elimination threshold and greedily dropping additional variables.

The initial model was trained using the top 32 variables obtained from cross validation. From the variables included in the initial model, the Arterial PaO_2 slope (`Art_PaO2_Slope_la`), magnesium (`Mg_lam`), and PaO_2:FiO_2 (`PaO2toFiO2`) were

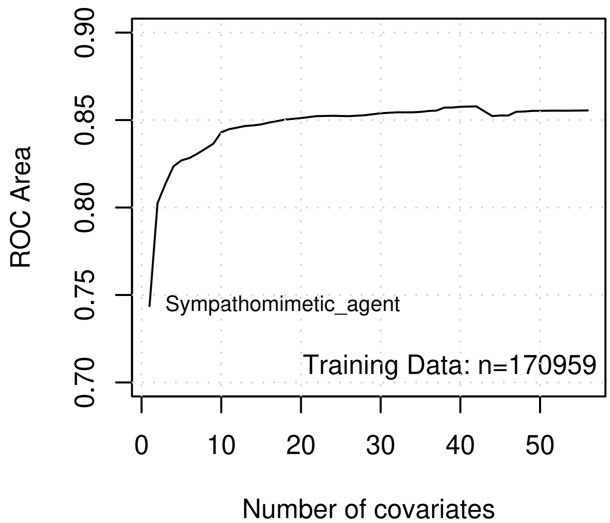


Figure 6-40: SSOM model selection (all development data)

dropped because of limited availability. Removing the PaO_2 slope, magnesium, and PaO_2:FiO_2 variables did not noticeably change the model's performance. In addition, the diastolic blood pressure (DBPm_sq) was replaced with the systolic blood pressure (SBPm_la) and the temperature (temp_am) was added with a negligible improvement in performance. The final model, with 30 inputs, is shown in Model 6.4.

Development Validation

To validate the SSOM model, I examine calibration performance and AUC performance. In addition, I also examine plots of the positive predictive value (PPV) versus sensitivity and the negative predictive value (NPV) versus specificity. Table 6.18 shows the deciles used for the Hosmer-Lemeshow H statistic and Table 6.19 shows the deciles used for the Hosmer-Lemeshow C statistic. The classification performance of SSOM on the training data is shown by the ROC curve in 6-41. In addition, plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-42.

Figure 6-43 shows the context surrounding the onset of HDFR for patients who lived (left) and patients who died (right). Similarly, Figure 6-44 shows the context surrounding the onset of HDFR for all predictions, ignoring the inclusion criteria (left) and only patients that satisfied the inclusion criteria (right).

As an illustration of predictions for an individual patient, Figure 6-45 shows the predictions for the patient used as an example of the annotation process (Figure 6-39). The patient, *Subject_ID* 13325, was admitted to the ICU with sepsis and expired after

Model 6.4 Final SSOM model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
144567	3e-07	21292.15	30	0	0.855	0.709
Gamma	Tau-a	R2	Brier			
0.713	0.116	0.302	0.067			

	Coef	S.E.	Wald Z	P
pressorTime_i	-1.177e-04	4.044e-06	-29.10	0
SBPm_la	-2.333e+00	1.491e-01	-15.65	0
alloutput_sqrt	-1.041e-02	7.644e-04	-13.62	0
Esmolol_i	-1.550e-04	1.202e-05	-12.90	0
X24hBal	-6.837e-05	5.473e-06	-12.49	0
GCSrdv	-3.930e-02	3.941e-03	-9.97	0
Morphine_Sulfate_sqrt	-4.136e-01	4.215e-02	-9.81	0
SICU	-7.848e-01	8.078e-02	-9.71	0
UrineEvnts.24h	-5.484e-03	6.042e-04	-9.08	0
Lidocaine_i	-9.965e-05	1.099e-05	-9.06	0
ventLenC_am	-7.732e-05	8.571e-06	-9.02	0
CO2_Slope_1680	-4.313e+01	5.182e+00	-8.32	0
Output_60_i	-2.560e-05	3.214e-06	-7.97	0
LOSBal	-3.390e-05	4.314e-06	-7.86	0
Vasodilating_agent	-3.038e-01	4.201e-02	-7.23	0
cvpM	1.730e-01	2.352e-02	7.35	0
temp_am	7.486e-02	9.911e-03	7.55	0
GCSrng_i	2.862e-05	3.255e-06	8.79	0
PulsePres_sqrt	1.502e-01	1.607e-02	9.35	0
Intercept	6.902e+00	7.362e-01	9.38	0
SpO2CritEvnts.24h_lam	1.351e-01	1.275e-02	10.60	0
Input_60_sqrt	1.710e-02	1.515e-03	11.29	0
allinput	4.785e-05	4.030e-06	11.88	0
totIV_sqrt	9.596e-03	7.845e-04	12.23	0
SBPmrdv	1.399e-02	1.116e-03	12.53	0
Age_la	5.291e-01	3.909e-02	13.54	0
WBC	2.100e-02	1.393e-03	15.07	0
Sympathomimetic_agent	6.394e-01	4.157e-02	15.38	0
mechVent	4.462e-01	2.636e-02	16.93	0
Neosynephrine_perKg_la	6.732e-02	3.215e-03	20.94	0
ShockIdx	1.541e+00	6.730e-02	22.91	0

Table 6.18: SSOM Hosmer-Lemeshow H risk deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[0.00119,0.0132)	0.01	45	143.1	14412	14313.9	14457
2	[0.01322,0.0179)	0.016	154	225.7	14303	14231.3	14457
3	[0.01792,0.0225)	0.02	166	292	14291	14165	14457
4	[0.02255,0.0279)	0.025	234	363.3	14222	14092.7	14456
5	[0.02785,0.0345)	0.031	338	448.7	14119	14008.3	14457
6	[0.03446,0.0441)	0.039	491	562.6	13966	13894.4	14457
7	[0.04412,0.0618)	0.052	910	749.3	13546	13706.7	14456
8	[0.06184,0.1190)	0.083	1735	1201.1	12722	13255.9	14457
9	[0.11903,0.2857)	0.2	3165	2887.9	11292	11569.1	14457
10	[0.28571,0.9724]	0.424	5765	6129.3	8691	8326.7	14456

$\chi^2 = 597.40,\ d.f. = 8;\ p = 0.000$

Table 6.19: SSOM Hosmer-Lemeshow C probability deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.033	3642	3689.2	109285	109237.8	112927
2	(0.1,0.2]	0.144	1739	1451.1	8313	8600.9	10052
3	(0.2,0.3]	0.25	2237	2076.7	6068	6228.3	8305
4	(0.3,0.4]	0.346	2234	2215.7	4175	4193.3	6409
5	(0.4,0.5]	0.445	1529	1635.1	2147	2040.9	3676
6	(0.5,0.6]	0.544	879	1021	998	856	1877
7	(0.6,0.7]	0.643	419	532.4	409	295.6	828
8	(0.7,0.8]	0.741	218	256.5	128	89.5	346
9-10	(0.8,1]	0.852	106	125.3	41	21.7	147

$\chi^2 = 249.86,\ d.f. = 7;\ p = 0.000$

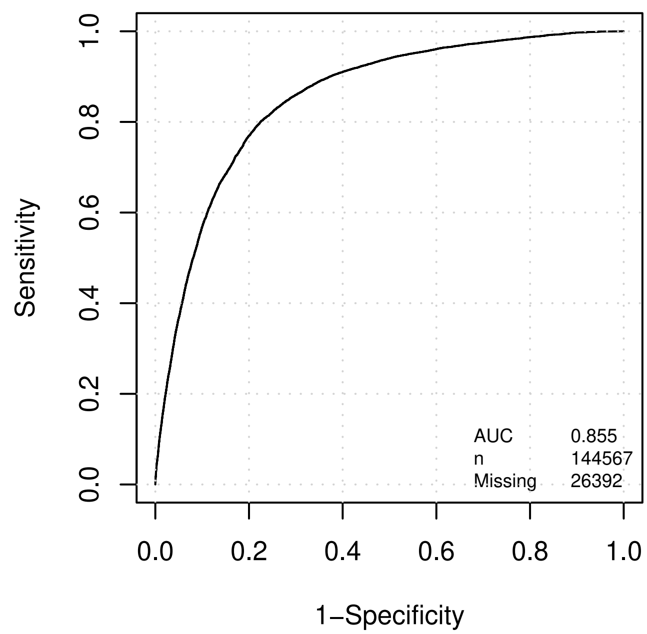


Figure 6-41: `SSOM` ROC curve (development data).

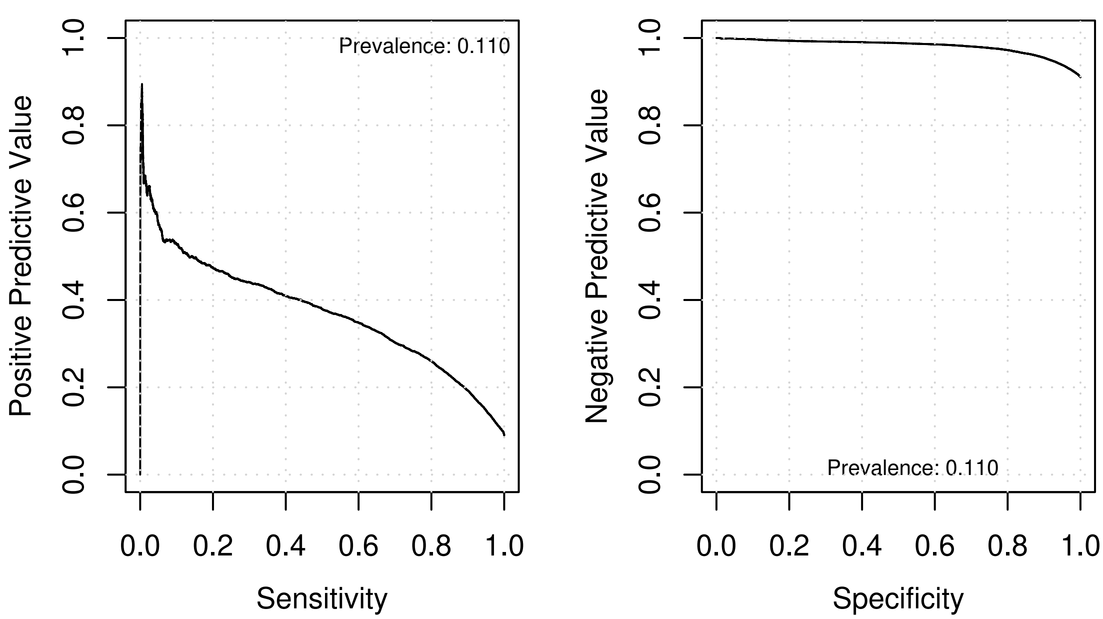


Figure 6-42: `SSOM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (development data).

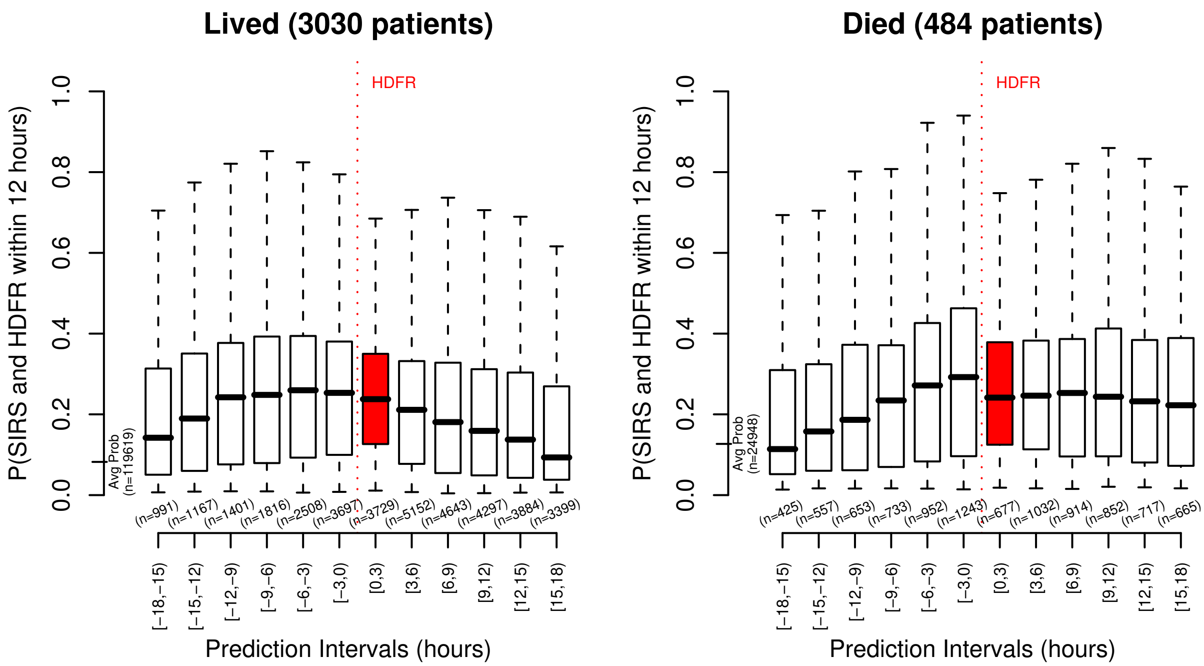


Figure 6-43: SSOM prediction context surrounding HDFR (development data). *Avg Prob*: the mean SSOM probability from all patients who lived (left) and died (right).

about 16 days. Other specific examples of SSOM predictions for individual patients (e.g., subjects 1124 and 24019) can be found in Appendix E.

6.4.4 Model Validation

As a final step, I validate the SSOM model on the separate validation data. To evaluate calibration, Table 6.20 and Table 6.21 provide the deciles used for the Hosmer-Lemeshow statistics. A plot of the calibration — actual probability versus estimated probability — is shown in Figure 6-46. The nonparametric curve on Figure 6-46 shows large deviations from the ideal calibration for large predicted probabilities, but the number of such estimates is very limited as shown by the histogram along the x-axis. Due to the limited number of estimates and the local lowess smoothing, the nonparametric curve for the large probabilities should be considered unreliable. The SSOM classification performance is summarized by the ROC curve in Figure 6-47. For comparison purposes, Figure 6-47 includes a curve generated by the RAS model developed in the previous chapter applied to the same prediction task (dotted blue).

Plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-48. The dotted blue lines show the performance obtained by using the RAS model output as a proxy to predict the same outcome as SSOM.

Finally, as done previously with the development patients, the context surrounding

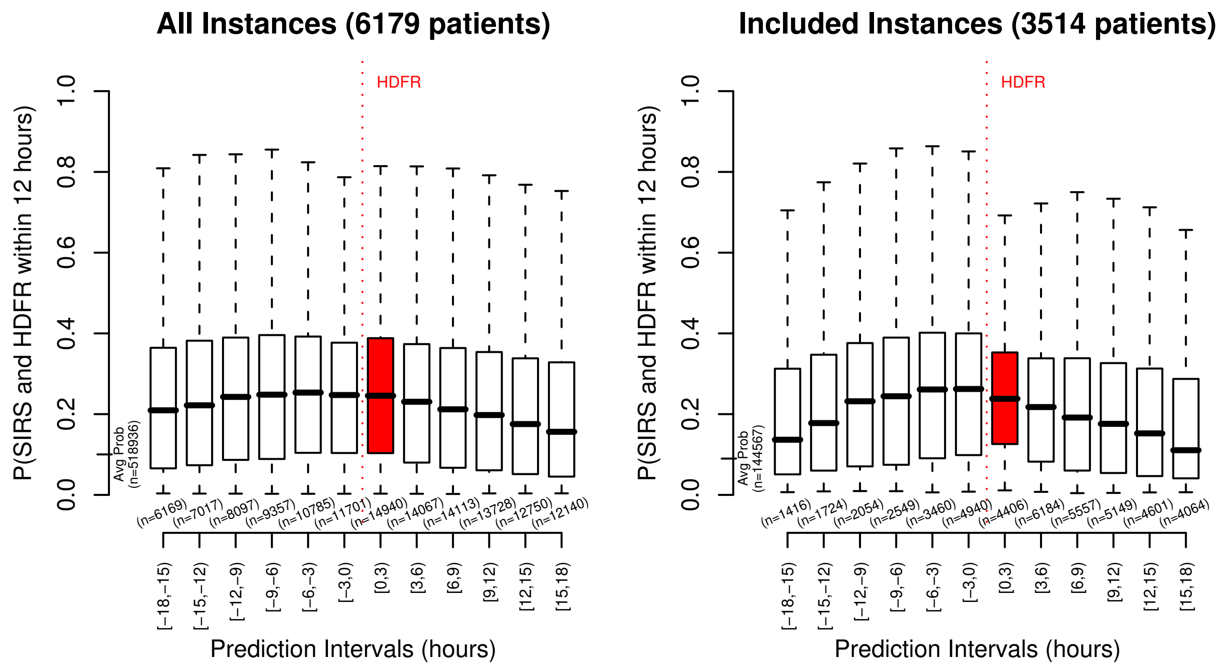


Figure 6-44: SSOM prediction context surrounding HDFR (development data). *Avg Prob*: the mean SSOM probability from all patient instances (left) and valid instances (right).

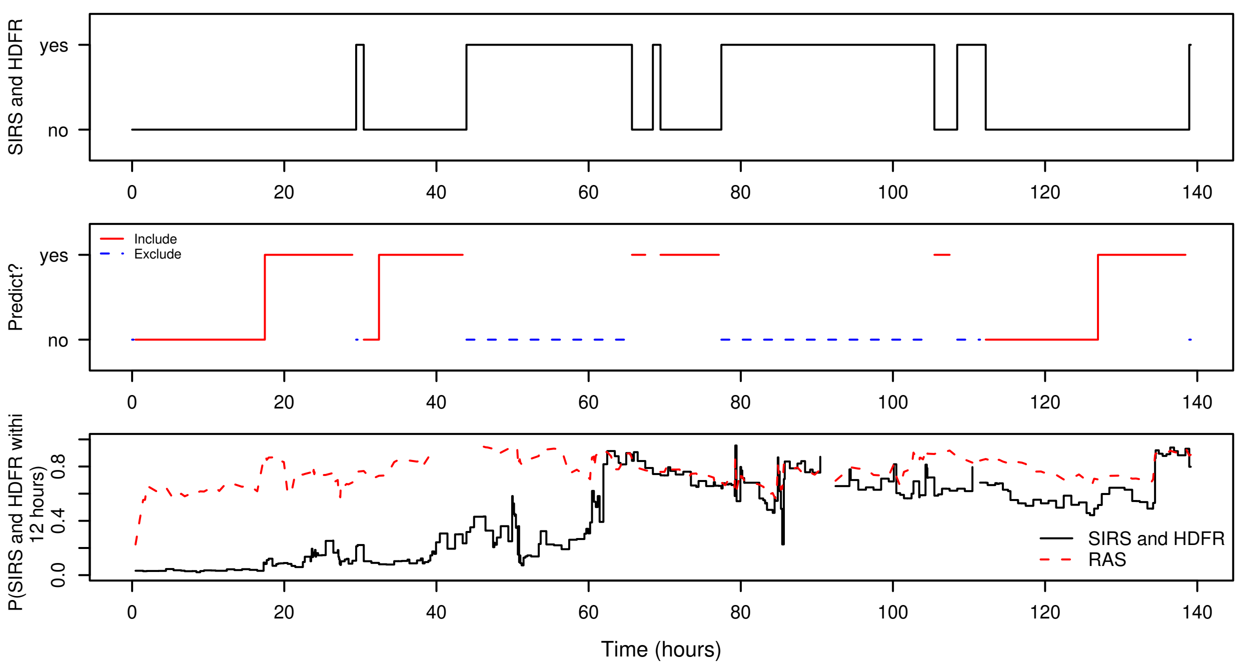


Figure 6-45: SSOM annotations for Subject_ID 13325 with SSOM and RAS predictions

Table 6.20: SSOM Hosmer-Lemeshow H risk deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[0.00237,0.0135)	0.01	30	56	5495	5469	5525
2	[0.01350,0.0184)	0.016	49	88.2	5476	5436.8	5525
3	[0.01837,0.0228)	0.021	86	113.6	5439	5411.4	5525
4	[0.02284,0.0280)	0.025	116	140	5409	5385	5525
5	[0.02797,0.0344)	0.031	167	171.6	5358	5353.4	5525
6	[0.03441,0.0440)	0.039	216	214.2	5309	5310.8	5525
7	[0.04400,0.0623)	0.052	370	287.1	5155	5237.9	5525
8	[0.06231,0.1242)	0.085	670	470.2	4855	5054.8	5525
9	[0.12419,0.3004)	0.21	1590	1159.6	3935	4365.4	5525
10	[0.30036,0.9749]	0.439	2193	2426.3	3331	3097.7	5524

$\chi^2 = 401.38,\ d.f. = 10;\ p = 0.000$

Table 6.21: SSOM Hosmer-Lemeshow C probability deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.033	1479	1401.8	41469	41546.2	42948
2	(0.1,0.2]	0.142	813	517.7	2831	3126.3	3644
3	(0.2,0.3]	0.249	997	777.4	2124	2343.6	3121
4	(0.3,0.4]	0.346	881	883.5	1675	1672.5	2556
5	(0.4,0.5]	0.447	645	717.1	959	886.9	1604
6	(0.5,0.6]	0.543	374	450	454	378	828
7	(0.6,0.7]	0.643	183	224.5	166	124.5	349
8	(0.7,0.8]	0.744	77	98.2	55	33.8	132
9-10	(0.8,1]	0.847	38	56.7	29	10.3	67

$\chi^2 = 404.33,\ d.f. = 9;\ p = 0.000$

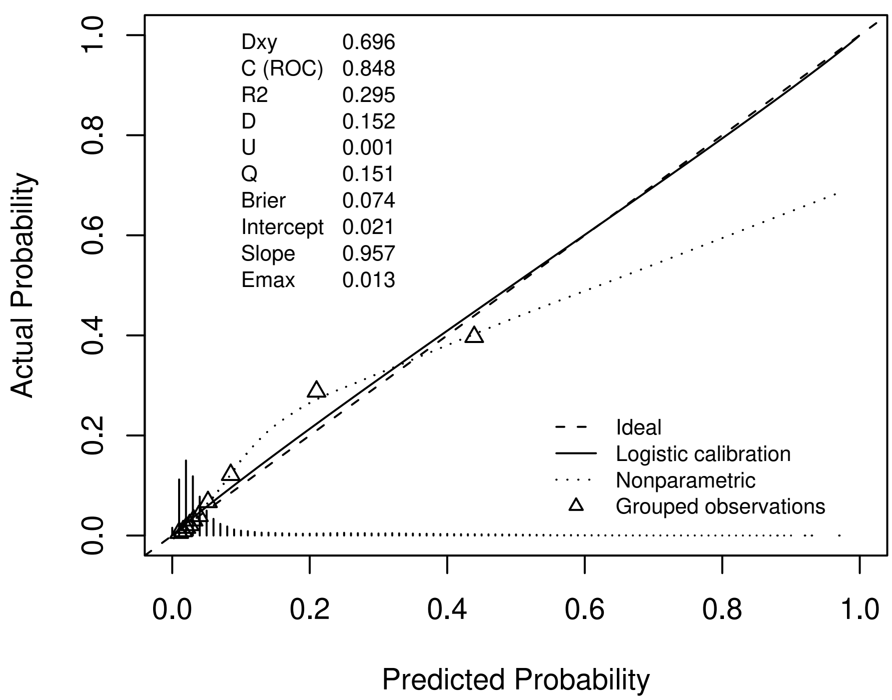


Figure 6-46: `SSOM` calibration plot

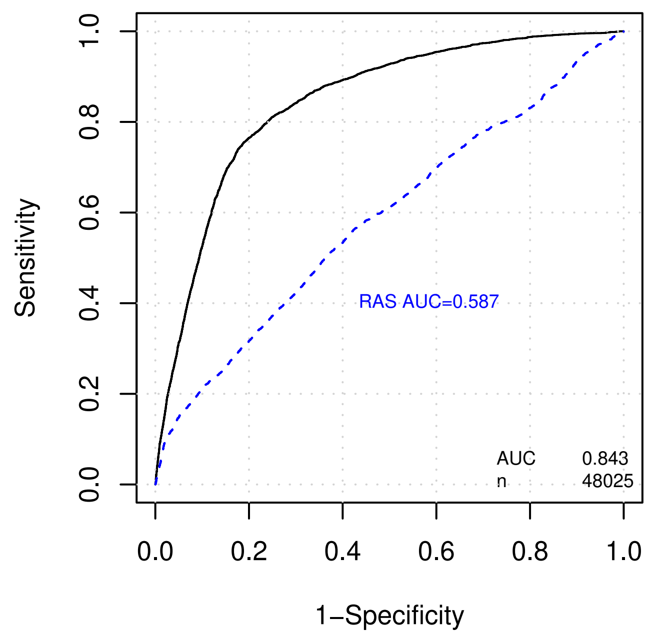


Figure 6-47: SSOM ROC curve (validation data).

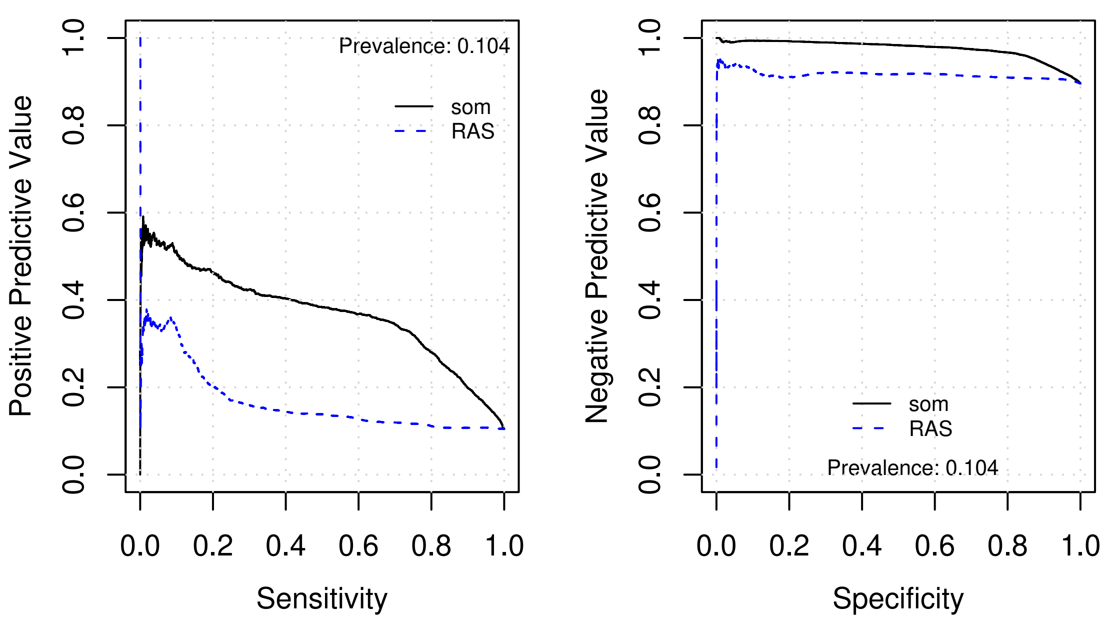


Figure 6-48: SSOM positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (validation data).

the onset of HDFR is examined for the validation data. Figure 6-49 shows the context surrounding HDFR onset for patients who survived (left) and patients who died (right). In addition to the SSOM predictions, the figure also shows *mortality* predictions from the RAS model. Figure 6-50 shows the prediction context for all predictions, including ones that did not satisfy inclusion criteria (left), and the prediction context for patients that did satisfy the inclusion criteria (right).

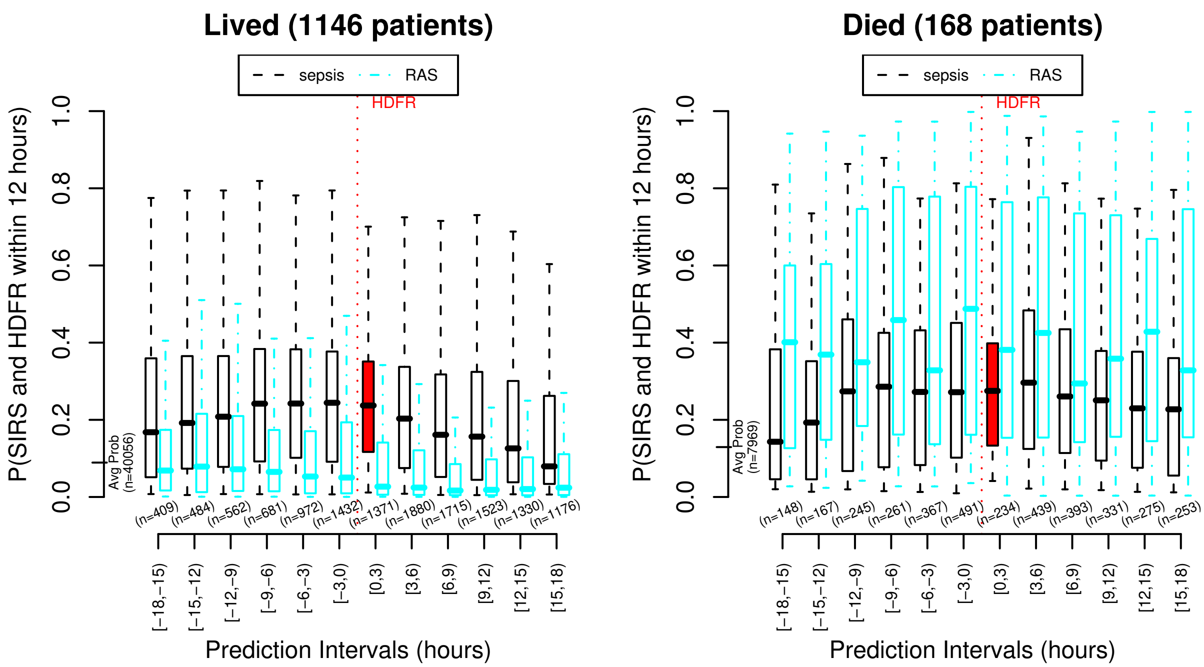


Figure 6-49: SSOM prediction context surrounding HDFR (validation data). *Avg Prob*: the mean SSOM probability from all patients who lived (left) and died (right).

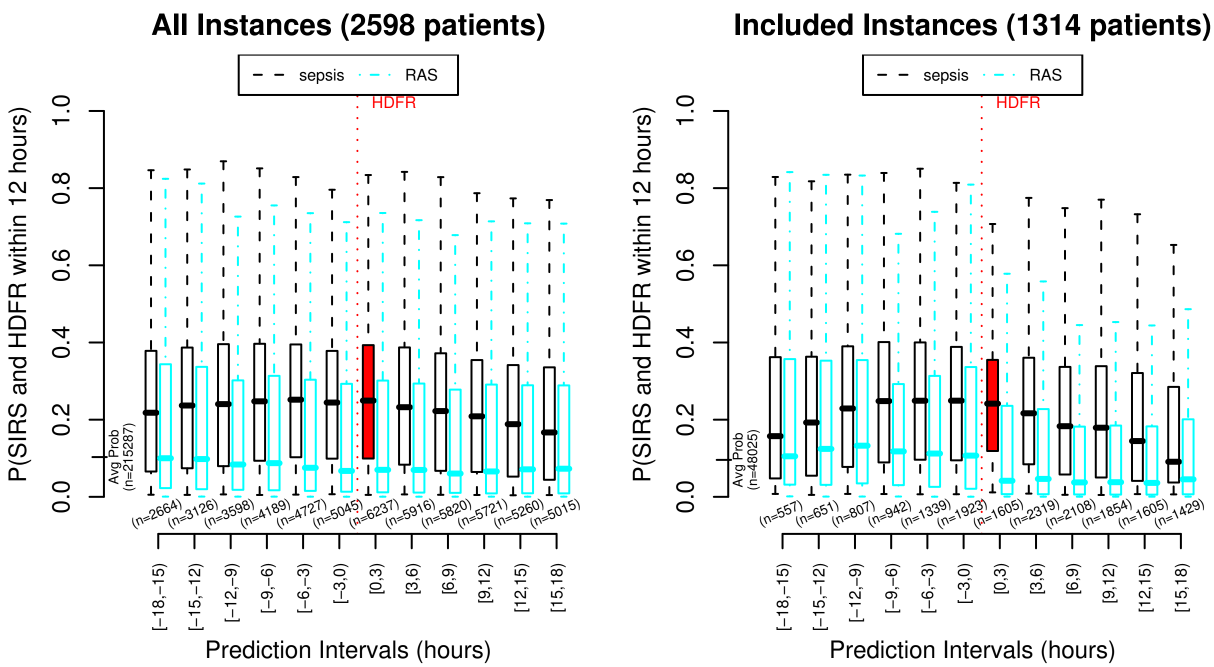


Figure 6-50: SSOM prediction context surrounding HDFR (validation data). *Avg Prob*: the mean SSOM probability from all patient instances (left) and valid instances (right).

6.4.5 Discussion

The `SSOM` model developed in this section attempts to predict a transition from SIRS *without* HDFR to SIRS *with* HDFR within 12 hours.

The `SSOM` model shown in Model 6.4 includes 30 inputs. Many of the most influential variables reflect the treatment that the patient is receiving. For example, the time that the patient has spent on pressors (`pressorTime_i`) is quite significant. Since the time spent on pressors input has units of 1/min (from the inverse transformation), it effectively serves as an indicator of the presence of pressors. If a patient is not receiving pressors, the value of $1/(0min + 0.0001) = 1000/min$ is used for the input and the estimate for the patient entering HDFR during the next 12 hours is lowered. As the time that the patient has spent on pressors increases, the contribution of the `pressorTime_i` input quickly decreases.

Other important treatment variables include (1) the amount of Neo-Synephrine the patient is receiving (`Neosynephrine_perKg_la`), which increases the probability of HDFR, (2) mechanical ventilation (`mechVent`), which increases the probability of HDFR, (3) the presence of sympathomimetic agents (`Sympathomimetic_agent`), which increases the probability of HDFR, and (4) the administration of esmolol (`Esmolol_i`), which also increases the probability of HDFR. The role of esmolol is especially interesting. Esmolol is a beta_1 receptor blocker with a very short half-life (10 min) that is used to treat acute arrhythmias. Its presence identifies a set of patients that have caregiver-induced hypotension that is unlikely to be related to severe sepsis. A variety of other variables, such as the patient being physically located in the surgery ICU (`SICU`), or receiving vasodilating agents (`Vasodilating_agent`) also decrease the risk of HDFR and likely lower the risk of septic shock.

In addition to the intervention variables, a number of physiological variables have significant influence in the model. Some of physiological variables include the shock index (HR/SBPm) (`ShockIdx`), the white blood cell count (`WBC`), the total amount of output from the patient (`alloutput_sqrt`), the patient's age (`Age_la`), the 24-hour fluid balance (`X24hBal`), and the total input received (`allinput`). In general, the physiological interpretations of these variables are expected and each play an important role in the diagnosis of HDFR and septic shock.

The calibration performance of the `SSOM` model is generally poor. The Hosmer-Lemeshow H and C statistics are quite significant for both the development and the validation data. The calibration plot (Figure 6-46) shows that the high probabilities are especially poor, but the model is dominated by low-probability predictions which, when corrected using the technique discussed in 4, translate to reasonable calibration (slope=0.957, intercept=0.021, and E_{max} of 0.013).

Given the definitions used, the classification performance for the `SSOM` model is strong. On the validation data, `SSOM` obtains an AUC of 0.843. The PPV is generally weak as a result of the low prevalence of HDFR episodes.

Severe sepsis is one of the primary causes of mortality in the ICU. Considering the increased mortality risk for septic patients, the group of patients considered by the SSOM model is expected to also represent high risk patients. The increased risk is confirmed by comparing the average RAS score between the episodes included by the SSOM model and instances not included by the SSOM model. The average RAS prediction for each patient is 0.217 when only the SIRS periods preceding HDFR (within the 12-hour limit) are considered. In contrast, the average value for SIRS-only periods without HDFR within 12 hours is 0.127. For non-SIRS periods, the average RAS prediction falls to 0.0907. The differences between each group are highly significant. An increasing trend in RAS predictions can often be observed in septic patients who continue to fight septic shock. Other patients with a better response to treatment, often show a generally decreasing trend. The weak sensitivity for RAS when predicting HDFR appears to be a result of a generally increased baseline for periods where patients are at risk for septic shock. The RAS model was not trained to limit this increase to a fixed time prior to HDFR and the RAS baseline often grows between multiple periods of HDFR for the same patient and over the course of several days.

In fact, the definitions used for the SSOM model are a major limitation. By manually reviewing a sample of patients that have the term "septic shock" present in their nursing notes, it appears that the definition needs further refinement to maximize predictive utility. Many of the patients experience an HDFR episode concurrently with SIRS and thus no predictions are available for the HDFR episode. Similarly, after an episode has been resolved, the patient often remains in SIRS. The post-HDFR SIRS helps to explain the low prevalence of warning annotations. In general it seems that the caregivers are quite mindful of SIRS. When SIRS is present without hemodynamic decompensation the patient is typically placed on a strict sepsis protocol and monitored closely for worsening sepsis. One of the more interesting cases appears to be when there is less warning and the cause of hypotension is not immediately clear to the clinicians (i.e., suspected to be sepsis or cardiogenic shock).

Furthermore, the broad definitions used by the SSOM model include a number of non-sepsis episodes. Manual review of the patients used for training and validation could help make the SSOM model more specific to septic shock and exclude patient episodes with HDFR but no sepsis (e.g., patients that are hypotensive due to beta blockers). Another way to tighten the definition of sepsis might be to better understand the septic shock treatment protocol used by the physicians in the units under consideration. The surviving sepsis campaign, for example, suggests treating septic shock with Levophed or dopamine as the initial vasopressor and epinephrine as the first alternative to septic shock refractory to Levophed or dopamine [14].

The average mortality prediction of 0.217 prior to HDFR episodes from the RAS model is reasonable. While the reported mortality rates for septic shock are typically higher (i.e., 28-50%), 22% is not far below the reported mortality range and our definitions of HDFR include a variety of sepsis severities and even non-sepsis cases. A stricter definition of septic shock would likely result in a higher average mortality prediction from `RAS`.

In conclusion, the `SSOM` model performs well at predicting the transition from SIRS without HDFR to SIRS with HDFR. Manual review of individual patients included in the `SSOM` model, however, reveals that the model is quite sensitive to the definitions used for SIRS and HDFR and the clinical utility of the model would likely benefit from additional refinement.

6.5 `AKIM`: Kidney Injury

Acute kidney injury (AKI) is a critical condition marked by a rapid loss of renal function. With decreased renal function, several waste products are not removed from the blood and the body's homeostatic balance can become compromised. The diagnosis of AKI is based on elevated creatinine or blood urea nitrogen values and/or decreased urine output. Due to varying kidney function between individuals (especially those with chronic renal failure), absolute measurements need to be compared with previous baseline measurements for a given patient.

An attempt was made in 2004 to provide a consensus definition to describe acute kidney injury. The result was the RIFLE acronym for classifying kidney function: *R*isk of renal dysfunction, *I*njury to the kidney, *F*ailure of kidney function, *L*oss of kidney function and *E*nd-stage kidney disease [3]. Previously more than 30 definitions existed.

The RIFLE classification scheme relies on relative changes in creatinine (as a proxy for the glomerular filtration rate or GFR), and urine output. The thresholds used to classify each stage are provided in Table 6.22. A patient's RIFLE classification is based on the most deranged value for GFR or urine output. The serum creatinine (SCreat) thresholds in Table 6.22 refer to change from baseline. Accurate baseline measurements are unavailable for the MIMIC II patients. Instead, the abbreviated "modification of diet in renal disease" (MDRD) study equation — based on serum creatinine, age, gender, and race — can be used to to estimate the baseline serum creatinine levels:

$$\text{eGFR} = 186 \times \text{SCreat}^{-1.154} \times \text{Age}^{-0.203} \times [1.210 \text{ if Black}] \times [0.742 \text{ if Female}].$$

Without a baseline creatinine value, an estimated GFR (eGFR) of 75 mL/min per 1.73 m^2, which is at the lower end of normal range, has been suggested for use with the MDRD equation to estimate serum creatinine [19, 3].

For the purposes of this section, I used the MDFR SCreat baseline estimate for RIFLE classification. In the MIMIC II data, patient race is unknown, so no separate (higher) baselines were available for African-Americans. The baseline estimates from the MDFR study equation are given in Table 6.23. For comparison, my entire dataset had a median serum creatinine value of 0.90 mg/dL, a mean value of 1.29 mg/dL, and a standard deviation of 1.19 mg/dL.

In this section, I develop a model to predict acute kidney injury (AKI). I use the term "kidney injury" to include both acute kidney injury and acute kidney failure. The model developed in this section will be referred to as the acute kidney injury model (`AKIM`).

Table 6.22: RIFLE Classification Scheme [3]. A patient is classified as the worst stage resulting from GFR *or* urine output.

Stage	GFR	Urine Output
Risk	Increased SCreat x1.5 or GFR decrease > 25%	< 0.5 ml/kg/h x 6 hr
Injury	Increased SCreat x2 or GFR decrease > 50%	< 0.5 ml/kg/h x 12 hr
Failure	Increased SCreat x3 or GFR decrease 75% or SCreat $\geq$ 4 mg/dL	< 0.3 ml/kg/h x 24 hrs or Anuria x 12 hrs

Table 6.23: Estimated baseline creatinine using the abbreviated MDRD equation and an estimated GFR = 75 ml/min per 1.73 m^2

Age (years)	Males (mg/dL)	Females (mg/dL)
20-24	1.3	1.0
25-29	1.2	1.0
30-39	1.2	0.9
40-54	1.1	0.9
55-65	1.1	0.8
>65	1.0	0.8

6.5.1 Data and Patient Inclusion Criteria

As described in Chapter 3, I attempted to remove patients with chronic renal failure by dropping patients that had an ICD-9 code of 585. The 225 patients that were dropped based on an ICD-9 code of 585 were likely only a subset of the chronic renal failure cases. In addition to the general inclusion criteria, I required that patient episodes used to develop and validate the `AKIM` model were classified by RIFLE under the "Risk" category. Patient episodes with a more severe RIFLE classification are not included in order to emphasize the *onset* of kidney injury. The Risk category requirement also omitted patient episodes with no valid creatinine measurement and urine measurement.

While many patients only received one daily creatinine measurement, it was common for measurements to be made more frequently. Figure 6-51 shows that 6-hour and 12-hour measurement intervals were also quite common.

After annotating the final dataset, a number of instances were excluded that did not contain episodes annotated as kidney risk. Table 6.24 provides a summary of the included data.

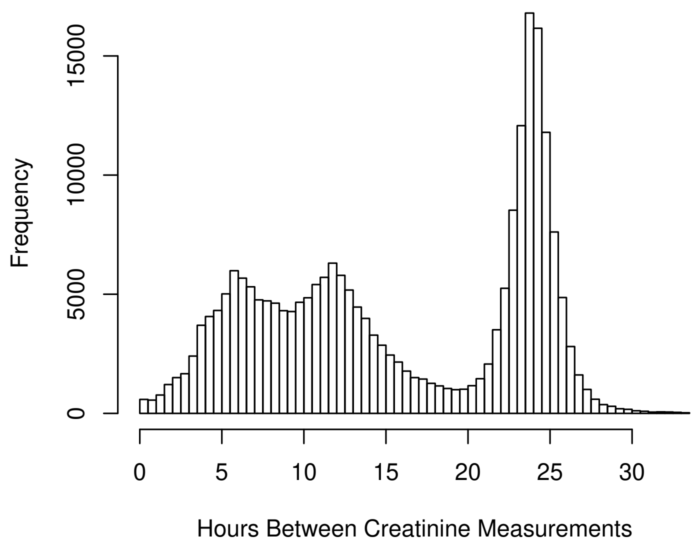


Figure 6-51: Creatinine Measurement Intervals

Table 6.24: `AKIM` data

	Count
Kidney risk patients	4591
Kidney injury/failure patients	3249
Included instances	103543
AKI 18-hour warn instances	31289
Risk but no AKI within 18 hours	72254

6.5.2 Outcome

The outcome for the `AKIM` model is the transition from kidney risk to acute kidney injury or failure, as specified by the RIFLE scheme, within 18 hours. The 18-hour window, in addition to the 6 hours for the RIFLE oliguria kidney risk classification, allows the development of injury and failure as determined by urine output, specifically the kidney failure criterion that requires 24 hours of oliguria.

The entire prediction window of 18 hours is often not fully utilized. The classification of kidney risk, on average, only occurs 6 hours prior to the classification of kidney injury. Figure 6-52 shows the distribution of lengths for the warning window. As the figure shows, most kidney risk episodes limit the warning window to within 6 hours and only a small number of predictions are limited by the 18-hour window.

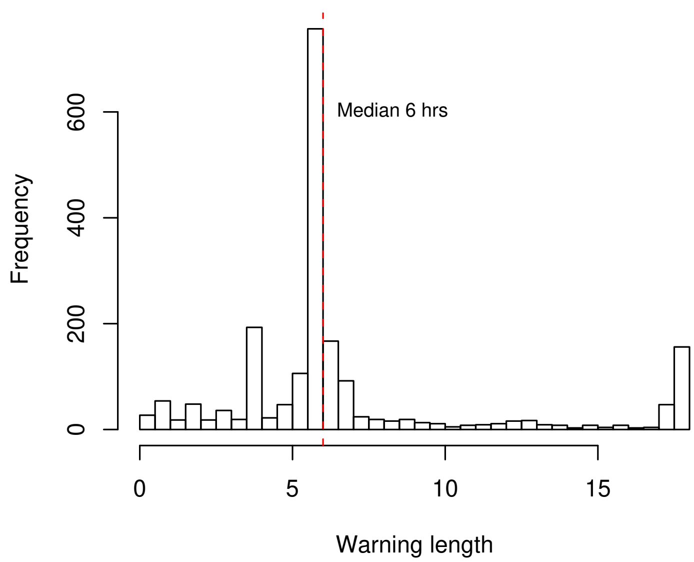


Figure 6-52: Acute kidney injury onset warning lengths

To illustrate the annotation process, Figure 6-53 shows how an example patient (`Subject_ID` 2539) was annotated. The figure in the top plot shows the kidney injury indicator. The bottom plot in the figure shows the 18 hour warning annotations; episodes where no annotations are made (i.e., did not satisfy the kidney risk criteria) are marked with the dashed blue line.

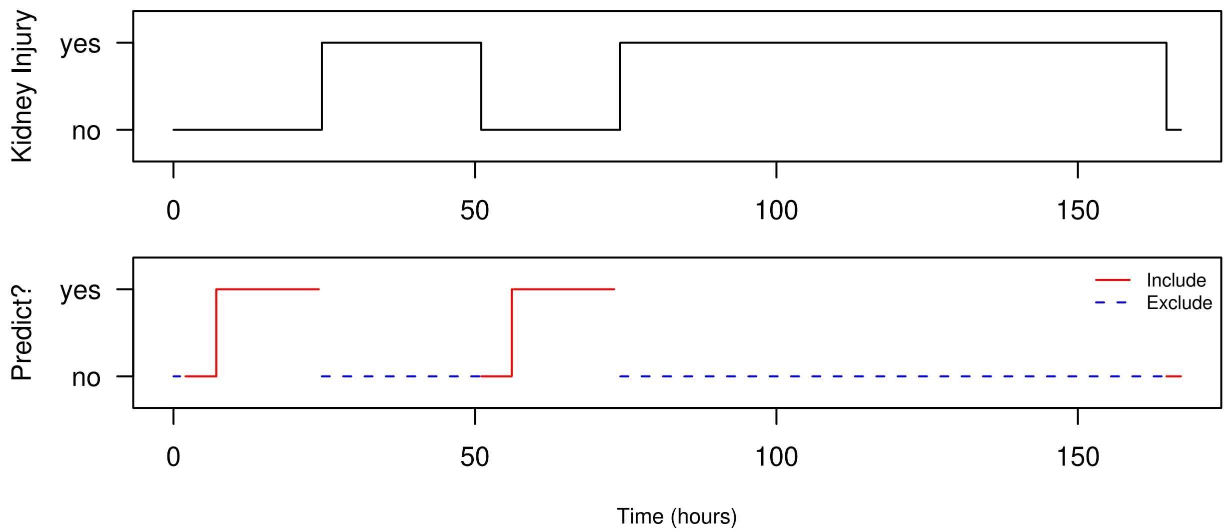


Figure 6-53: `AKIM` example annotations for `Subject_ID` 2539

6.5.3 Model Development

The `AKIM` model was developed using the same methodology as the other secondary outcome models discussed in this chapter (described in Chapter 4). In this section I describe the `AKIM` model selection process and the resulting logistic regression model. I also provide a description of the model's performance on the training data.

Model Selection

Candidate variables were initially ranked against the outcome variable (kidney injury). Variables with a p-value greater than 0.05 were excluded. Furthermore, if multiple variables were strongly correlated (Spearman's rank correlation test > 0.8) the best univariate variable was retained. After the initial screening of the variables, variable selection for the `AKIM` model was based on the best 20 variables from each of the 5 cross-validation folds (the individual cross validation plots are provided in Appendix F). For each cross-validation fold, the validation performance decreased after 20 variables. When combined, the best 20 variables from the 5 folds resulted in 40 unique candidate variables. Figure 6-54 shows the AUC that resulted from gradually increasing the AIC backward elimination threshold and greedily dropping additional variables.

The final `AKIM` model was trained using the top 26 variables. This model is shown in Model 6.5.

Model 6.5 Final `AKIM` model

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
56952	9e-07	10581.15	26	0	0.768	0.537
Gamma	Tau-a	R2	Brier			
0.538	0.217	0.244	0.165			

	Coef	S.E.	Wald Z	P
X24hUrOut_sqrt	-2.701e-02	1.642e-03	-16.45	0
MBPm_la	-1.745e+00	1.070e-01	-16.31	0
Nitroglycerine_i	-5.674e-05	3.706e-06	-15.31	0
UrineOut_sqrt	-5.105e-02	3.634e-03	-14.05	0
alloutput_sqrt	-1.288e-02	9.633e-04	-13.37	0
BUN_la	-3.480e-01	2.794e-02	-12.46	0
CO2_Slope_1680	-7.199e+01	5.986e+00	-12.03	0
Creatinine_sqrt	-7.317e-01	7.656e-02	-9.56	0
Output_60_sqrt	-2.208e-02	2.621e-03	-8.43	0
COtdM	-2.343e-01	2.800e-02	-8.37	0
CO2_am	2.815e-02	3.721e-03	7.56	0
Platelets_Slope_1680_i	2.811e-05	3.292e-06	8.54	0
CV_HR_Slope_240_i	4.156e-05	4.762e-06	8.73	0
Midazolam_sqrt	1.674e-01	1.866e-02	8.97	0
temp_am	1.064e-01	1.143e-02	9.32	0
hrmPaced	3.358e-01	3.233e-02	10.39	0
Levophed_i	4.052e-05	3.894e-06	10.41	0
pressD24	7.416e-01	6.971e-02	10.64	0
totOut_sqrt	1.923e-02	1.712e-03	11.24	0
index_sqrt	1.370e-02	1.215e-03	11.27	0
DBPm_la	9.632e-01	8.374e-02	11.50	0
HCT_sq	4.451e-04	3.622e-05	12.29	0
Intercept	4.058e+00	2.877e-01	14.10	0
Sympathomimetic_agent	3.985e-01	2.792e-02	14.27	0
totIn_am	8.290e-05	5.074e-06	16.34	0
urineByHr.oor120.t_sqrt	6.669e-02	2.711e-03	24.60	0
admitWt_sq	6.204e-05	2.258e-06	27.48	0

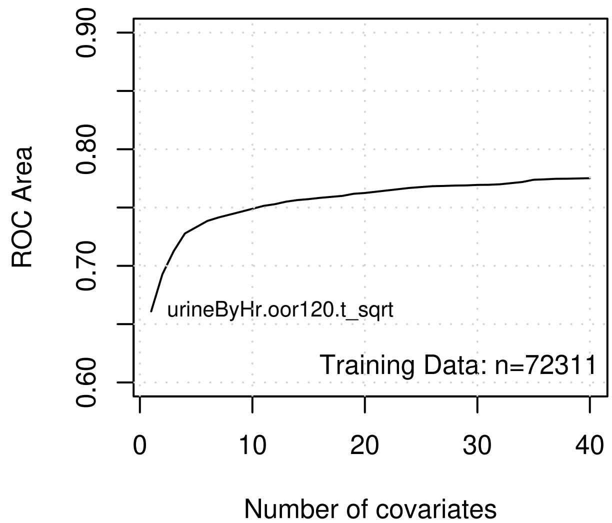


Figure 6-54: AKIM model selection (all development data)

Development Validation

To validate the AKIM model, I examine calibration performance and AUC performance. In addition, I plot the PPV versus sensitivity and the NPV versus specificity. Table 6.25 shows the deciles used for the Hosmer-Lemeshow H statistic and Table 6.26 shows the deciles used for the Hosmer-Lemeshow C statistic. The classification performance of AKIM on the training data is shown by the ROC curve in 6-55. The PPV versus sensitivity and the NPV versus specificity are plotted in Figure 6-56

Figure 6-57 shows the context surrounding transitions from kidney risk to kidney injury for patients who lived (left) and patients who died (right). Similarly, Figure 6-58 shows the context surrounding transitions from kidney risk to kidney injury, ignoring the inclusion criteria (i.e., not requiring kidney *risk*), on the left side and only the patients that satisfied the inclusion criteria on the right.

As an illustration of predictions for an individual patient, Figure 6-59 shows the predictions for the patient used to demonstrate the annotation process (Figure 6-53).

6.5.4 Model Validation

As a final step, I validate the AKIM model on the separate validation data. To evaluate calibration, Table 6.27 and Table 6.28 provide the deciles used for the Hosmer-Lemeshow statistics. A plot of the calibration — actual probability versus estimated probability — is shown in Figure 6-60. The AKIM classification performance is sum-

Table 6.25: AKIM Hosmer-Lemeshow H risk deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[0.00387,0.0774)	0.055	352	313.3	5344	5382.7	5696
2	[0.07739,0.1108)	0.095	440	539.2	5255	5155.8	5695
3	[0.11081,0.1437)	0.127	566	724	5129	4971	5695
4	[0.14365,0.1810)	0.162	828	921.6	4867	4773.4	5695
5	[0.18096,0.2275)	0.204	1137	1160	4558	4535	5695
6	[0.22754,0.2847)	0.255	1518	1452.5	4178	4243.5	5696
7	[0.28466,0.3566)	0.319	2056	1819.2	3639	3875.8	5695
8	[0.35662,0.4474)	0.4	2452	2278.1	3243	3416.9	5695
9	[0.44744,0.5704)	0.505	2940	2876.5	2755	2818.5	5695
10	[0.57038,0.9873]	0.685	3697	3901.7	1998	1793.3	5695

$\chi^2 = 184.93,\ d.f. = 8;\ p = 0.000$

Table 6.26: AKIM Hosmer-Lemeshow C probability deciles (development data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.068	625	638.6	8739	8725.4	9364
2	(0.1,0.2]	0.146	1990	2311.4	13802	13480.6	15792
3	(0.2,0.3]	0.246	2672	2551.9	7683	7803.1	10355
4	(0.3,0.4]	0.348	2824	2524.9	4434	4733.1	7258
5	(0.4,0.5]	0.447	2586	2460.9	2915	3040.1	5501
6	(0.5,0.6]	0.547	2123	2179.7	1861	1804.3	3984
7	(0.6,0.7]	0.646	1698	1717	959	940	2657
8	(0.7,0.8]	0.743	907	999	438	346	1345
9	(0.8,0.9]	0.841	415	445.7	115	84.3	530
10	(0.9,1]	0.946	146	157	20	9	166

$\chi^2 = 190.39,\ d.f. = 8;\ p = 0.000$

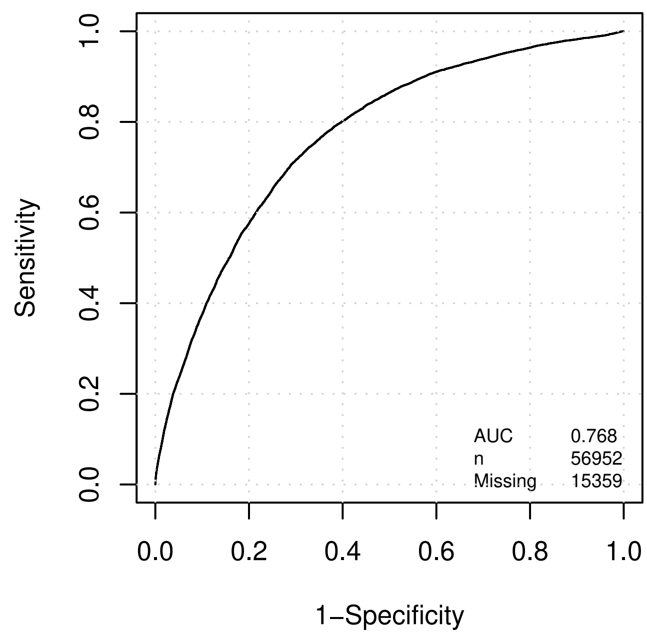


Figure 6-55: AKIM ROC curve (development data).

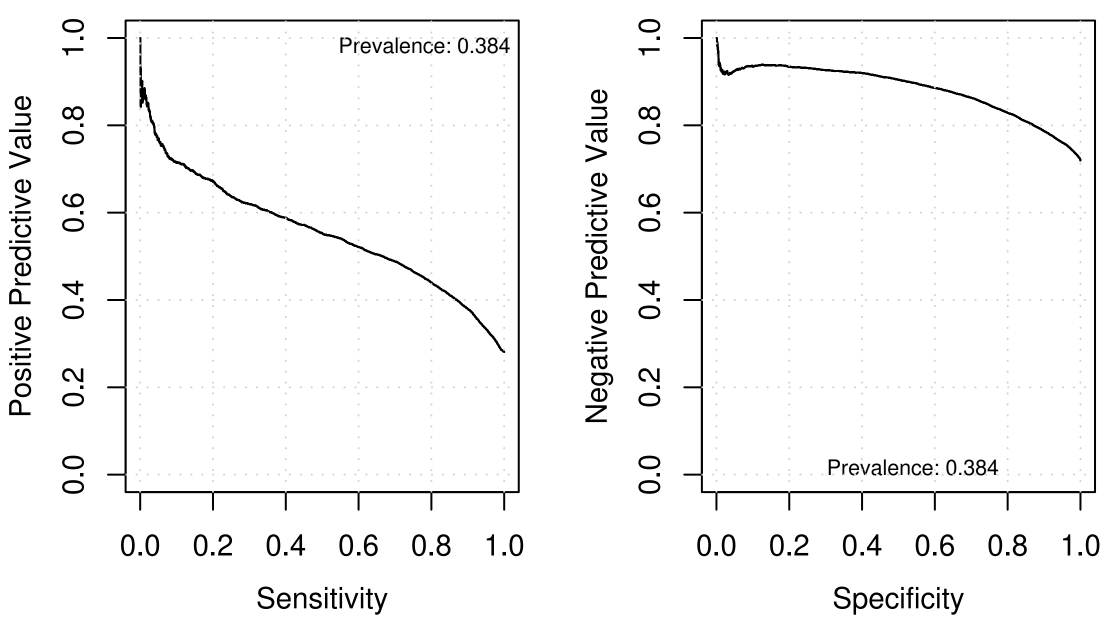


Figure 6-56: AKIM positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (development data).

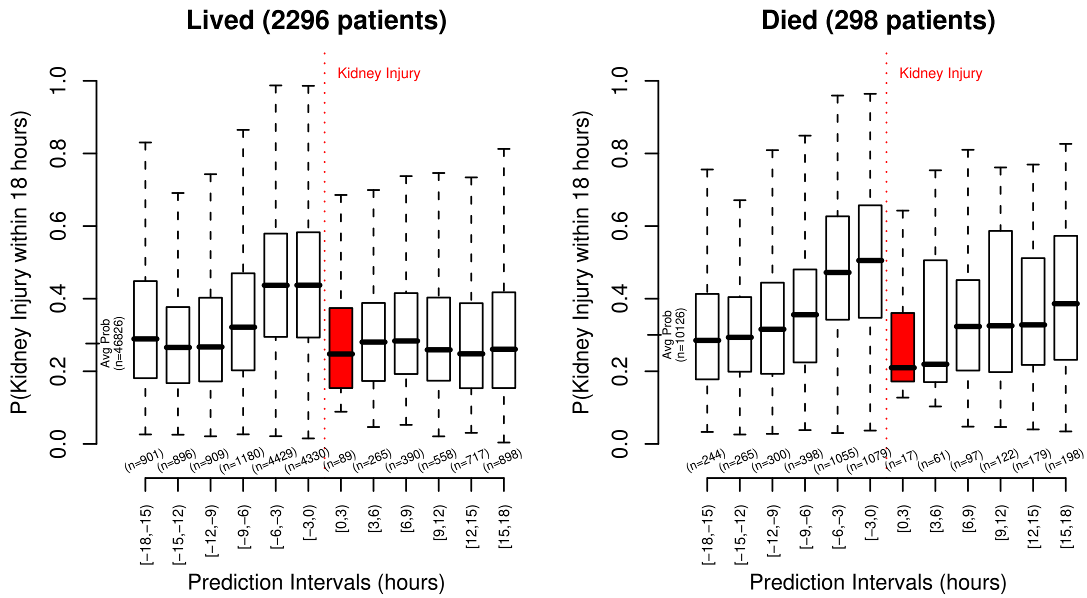


Figure 6-57: `AKIM` prediction context surrounding kidney injury (development data). *Avg Prob*: the mean `AKIM` probability from all patients who lived (left) and died (right).

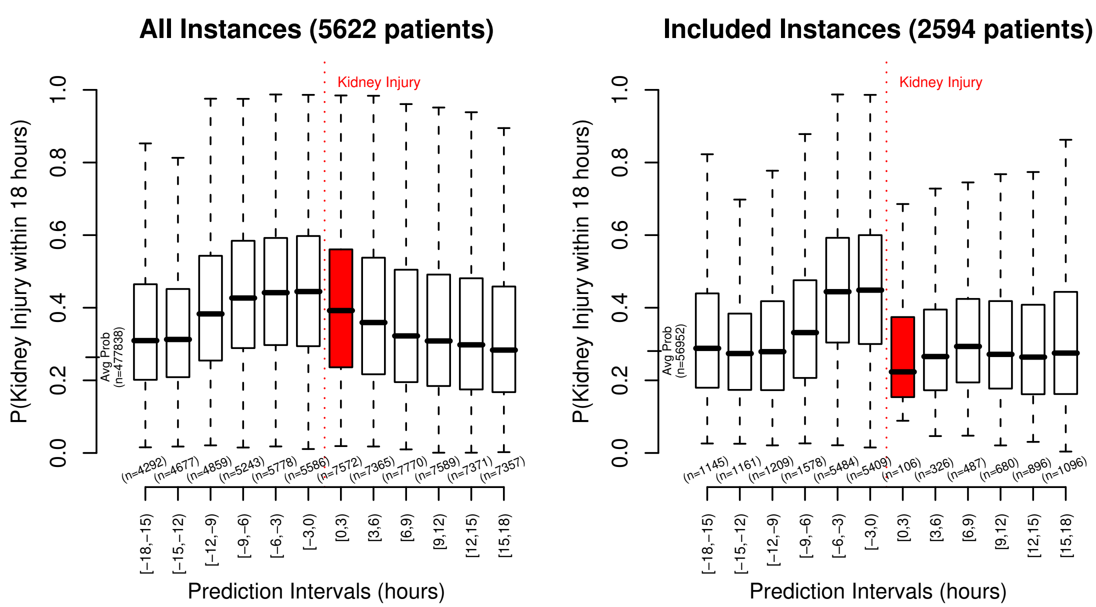


Figure 6-58: `AKIM` prediction context surrounding kidney injury (development data). *Avg Prob*: the mean `AKIM` probability from all patient instances (left) and valid instances (right).

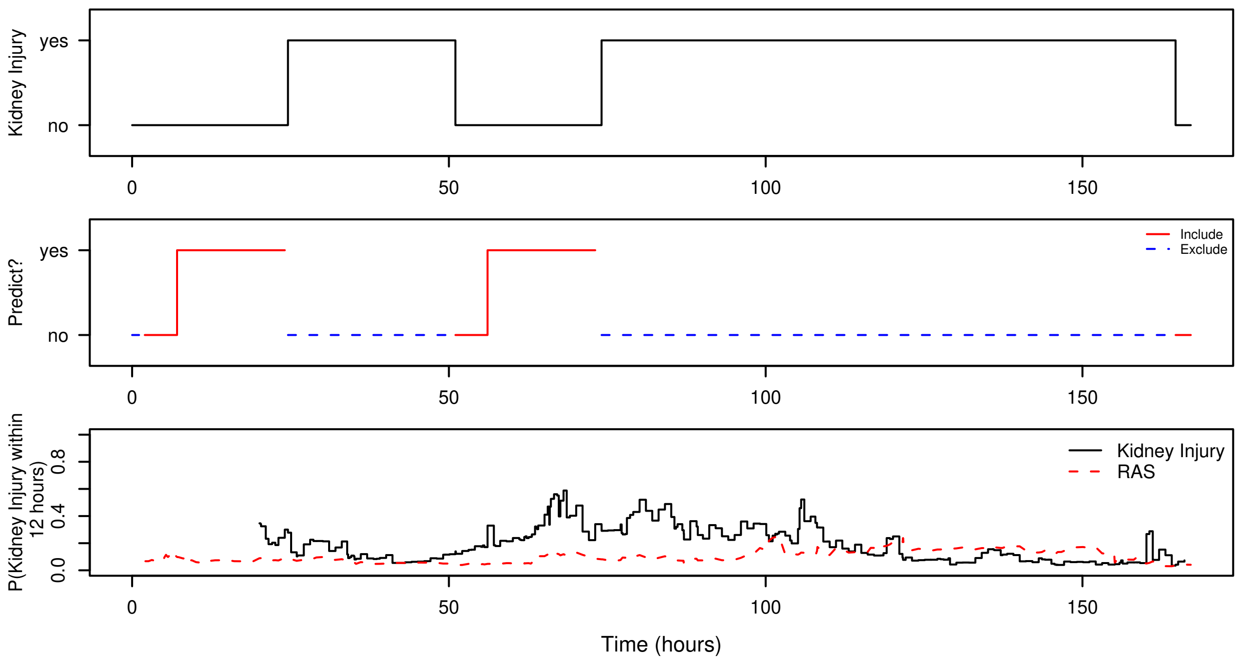


Figure 6-59: AKIM annotations for Subject_ID 2539 with AKIM and RAS predictions

marized by the ROC curve in Figure 6-61. For comparison purposes, Figure 6-61 includes a curve generated by the RAS model from the previous chapter applied to the kidney injury prediction task (dotted blue).

Plots showing the PPV versus sensitivity and the NPV versus specificity are provided in Figure 6-62. The dotted blue lines show the performance obtained by using the RAS model output as a proxy to predict the same outcome as AKIM.

Finally, as done previously with the development patients, the context surrounding the onset of kidney injury is examined for the validation data. Figure 6-63 shows the context surrounding the onset of kidney injury for patients who survived (left) and patients who died (right). In addition to the AKIM predictions, the figure also shows *mortality* predictions from the RAS model. Figure 6-64 shows the prediction context surrounding the onset of kidney injury for all predictions, including ones that did not satisfy inclusion criteria (left), and the prediction context for all patients that did satisfy the inclusion criteria (right).

Table 6.27: AKIM Hosmer-Lemeshow H risk deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	[5.05e-05,0.077)	0.051	277	125.1	2163	2314.9	2440
2	[7.70e-02,0.111)	0.094	208	230	2231	2209	2439
3	[1.11e-01,0.146)	0.129	247	314.5	2193	2125.5	2440
4	[1.46e-01,0.181)	0.163	338	397.9	2101	2041.1	2439
5	[1.81e-01,0.226)	0.203	449	495.5	1990	1943.5	2439
6	[2.26e-01,0.283)	0.253	666	618	1774	1822	2440
7	[2.83e-01,0.353)	0.317	859	773.5	1580	1665.5	2439
8	[3.53e-01,0.439)	0.394	1086	961.4	1354	1478.6	2440
9	[4.39e-01,0.549)	0.49	1235	1195.8	1204	1243.2	2439
10	[5.49e-01,0.953]	0.652	1498	1589.1	941	849.9	2439

$\chi^2 = 292.73,\ d.f. = 10;\ p = 0.000$

Table 6.28: AKIM Hosmer-Lemeshow C probability deciles (validation data)

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1	(0,0.1]	0.066	417	271	3666	3812	4083
2	(0.1,0.2]	0.148	820	1000.6	5925	5744.4	6745
3	(0.2,0.3]	0.246	1151	1092.9	3288	3346.1	4439
4	(0.3,0.4]	0.348	1269	1115.5	1939	2092.5	3208
5	(0.4,0.5]	0.447	1178	1116.1	1318	1379.9	2496
6	(0.5,0.6]	0.546	967	972.2	814	808.8	1781
7	(0.6,0.7]	0.644	621	662.9	408	366.1	1029
8	(0.7,0.8]	0.743	339	352.9	136	122.1	475
9-10	(0.8,1]	0.847	101	116.8	37	21.2	138

$\chi^2 = 188.79,\ d.f. = 9;\ p = 0.000$

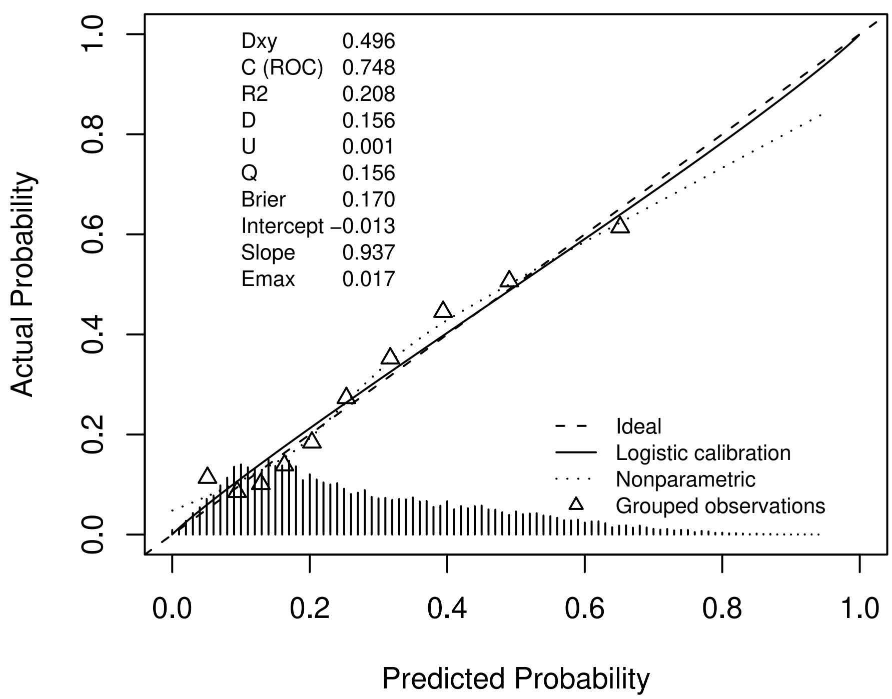


Figure 6-60: AKIM calibration plot (validation data)

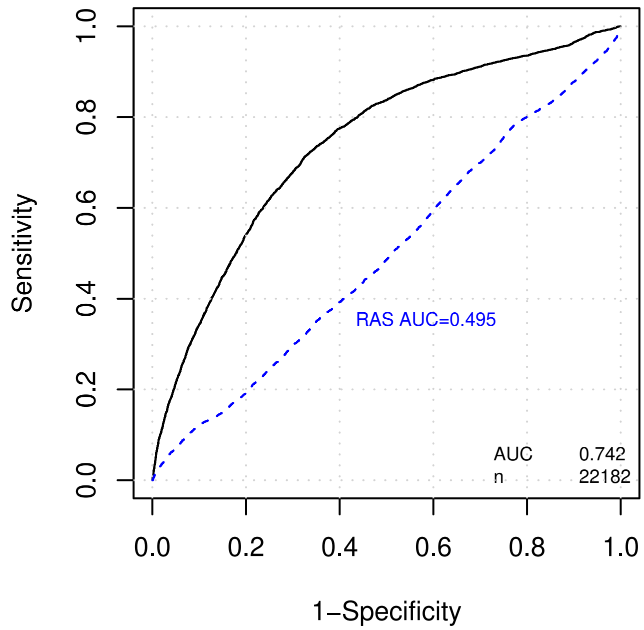


Figure 6-61: `AKIM` ROC curve (validation data).

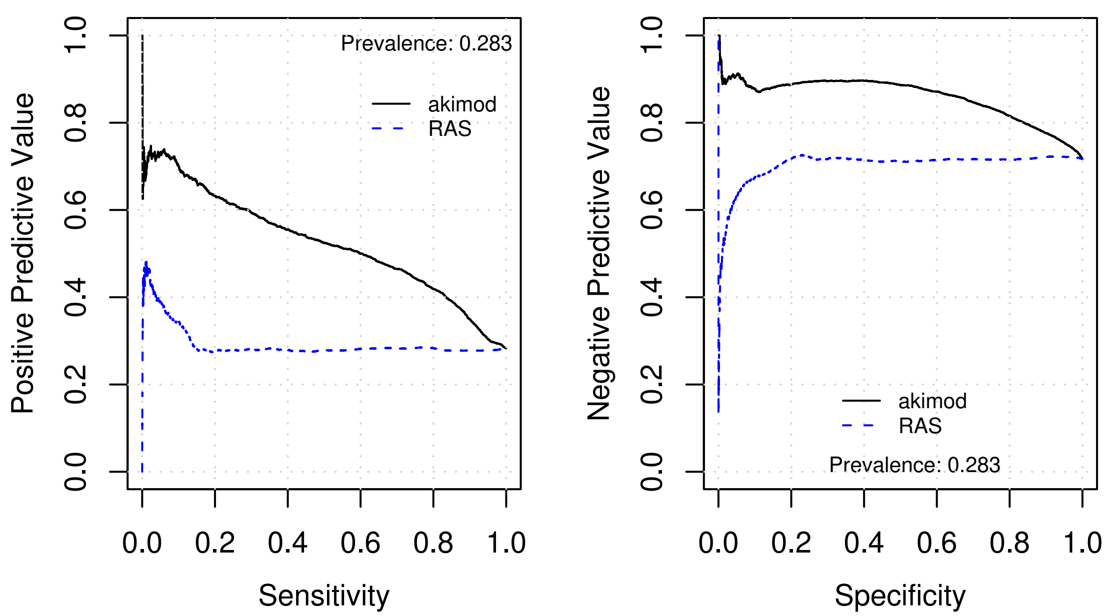


Figure 6-62: `AKIM` positive predictive value (PPV) versus sensitivity (left) and negative predictive value (NPV) versus specificity (right) (validation data).

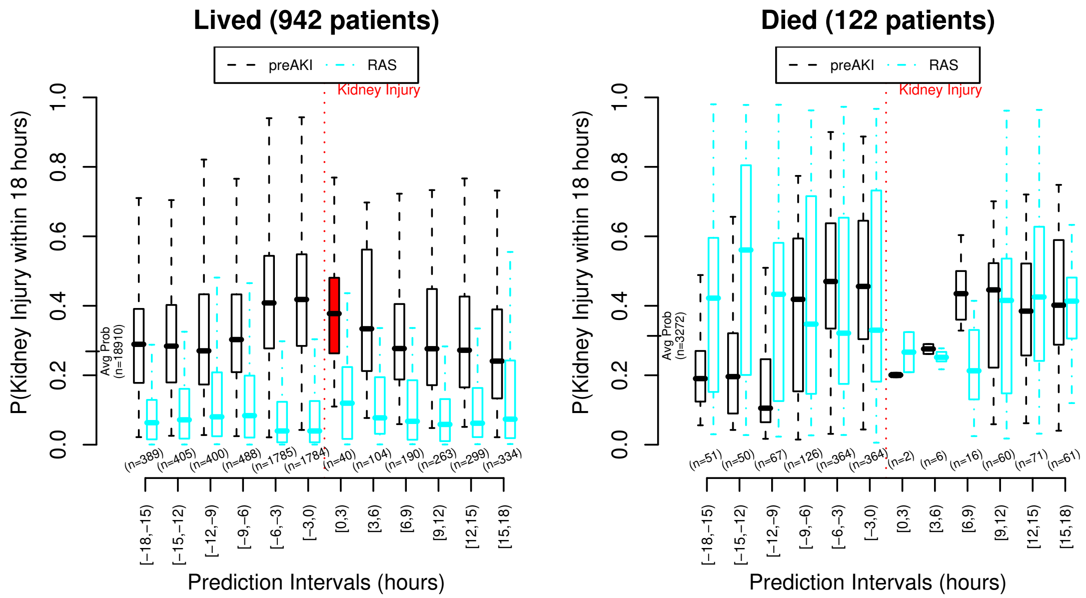


Figure 6-63: AKIM prediction context surrounding kidney injury (validation data). *Avg Prob*: the mean AKIM probability from all patients who lived (left) and died (right).

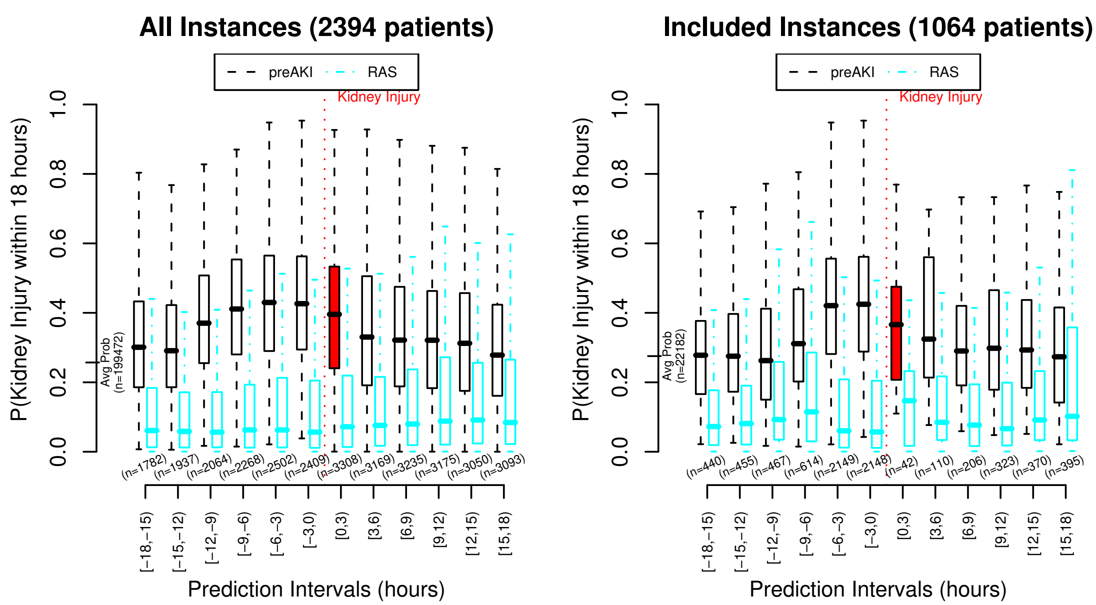


Figure 6-64: AKIM prediction context surrounding kidney injury (validation data). *Avg Prob*: the mean AKIM probability from all patient instances (left) and valid instances (right).

6.5.5 Discussion

The `AKIM` model that I presented in this section attempts to predict decreasing renal function in the form of a progression from *risk* of kidney injury to kidney injury or kidney failure. It uses the change in creatinine from an estimated baseline and the hourly urine output specified by the RIFLE classification scheme to annotate these kidney changes.

Of the models presented in this chapter, the `AKIM` model has the worst prediction performance. The relatively poor performance is not surprising given the necessary estimation of the creatinine baseline values and the rather coarse RIFLE definitions used to build the model. Despite the predictive difficulty of the `AKIM` model, it contains a number of interesting physiological inputs.

Some of the more predictive physiological inputs include the patient weight from admission (`admitWt_sq`), the number of out-of-range urine events during the past 2 hours (`urineByHr.oor120.t_sqrt`), the 24-hour urine output (`X24hUrOut_sqrt`), and the mean blood pressure (`MBPm_la`). The weights associated with these variables appear to be consistent with what one would expect in their reflection of decreased cardiac output and the symptoms of early kidney failure.

The `AKIM` model includes three trend variables. The three trend variables are the CO_2 slope over 28 hours (`CO2_Slope_1680`), the platelets slope over 28 hours (`Platelets_Slope_1680_i`), and the heart rate slope over 4 hours (`CV_HR_Slope_1680_i`). The CO_2 slope, for example, indicates that the patient is trending away from renal acidosis that often results from an accumulation of urea and creatinine in the blood.

Unlike many of the other secondary outcome models developed in this chapter, interventions appear to be less influential in the `AKIM` model. The most important drug inputs include nitroglycerine (`Nitroglycerine_i`), administration of a sympathomimetic agent (`Sympathomimetic_agent`), Levophed (`Levophed_i`), and midazolam (`Midazolam_sqrt`). Lasix was notably absent from this list.

The `AKIM` model demonstrates moderate calibration performance. The `AKIM` model has significant values for the Hosmer-Lemeshow C and H statistics, but is calibrated better than the other secondary models in this chapter. The calibration plot shown in Figure 6-60 confirms this finding with a slope of 0.937, an intercept of -0.013, and an E_{max} of 0.017 for the logistic correction.

The `AKIM` model offers only moderate performance at predicting kidney injury. The AUC value for the `AKIM` model on the validation data is 0.742. By looking at the context plots for the `AKIM` model predictions in Figure 6-57, it appears that for both the training and validation patients a consistent increase in the model's prediction value is not observed until 6 hours prior to the classification of kidney injury. The behavior displayed within 6 hours of kidney injury indicates that the prediction window of 18 hours might adversely affect the performance of the model, and a more realistic predictive window might be nine hours prior to the event. If the n values along the bottom of the context plots are examined, it is clear that the vast majority of positive training instances already occur within 6 hours of the kidney injury event. There are also notable prediction differences between patients that live (left) and patients that die (right). For patients that ultimately die, there is a significant increase in estimates between the interval [-12,-9) hours before the event and [-9,-6) hours before the event ($p < 0.00001$). When making conclusions from the context plots, however, it is important to consider the precipitous drop in the number of predictions available for intervals that precede the kidney injury event by more than 6 hours.

It is also interesting to note that, in general, the `RAS` model predictions are unhelpful in identifying increased renal dysfunction. The predictions, however, may not be entirely useless. In Figure 6-57, the `RAS` predictions, on average, reach their maximum between 12 and 9 hours prior to the kidney injury event and subsequently trend in the wrong direction. This pattern is especially noticeable in a patients who died, where the median `RAS` prediction is nearly 0.6 between 15 and 12 hours prior to the kidney injury event. A peak in mortality risk several hours prior to kidney injury is consistent with the understanding that kidney injury typically follows an event that causes reduced cardiac output [65]. When the inclusion criteria are ignored (i.e., predictions are not limited to episodes that fall within the kidney risk classification) the observed `RAS` pattern is no longer evident. If the structure of the prediction problem were changed to focus on events that ultimately caused the renal dysfunction (e.g., hypovolemia), the `RAS` model would likely perform better. Many of the events that led to kidney dysfunction are likely acute life-threatening events with interesting risk profiles. Without any changes the average prediction from `RAS` for all kidney risk periods up to 18 hours prior to a kidney injury event is 0.10 for patients who lived and 0.39 for patients who died.

A number of limitations and areas for improvements exist for the `AKIM` model. First, the model would benefit from better baseline estimates for creatinine. The RIFLE criteria have been criticized for using the MDRD equation to estimate baseline creatinine for ICU patients. The MDRD equation has not been validated for use in an ICU setting and likely misclassifies a number of patients such as those with low muscle mass [13]. Furthermore, the baseline creatinine estimate that I use does not include information regarding the patient's race. A better baseline creatinine estimate would allow the model to include a larger range of kidney function in its relative classification. Currently, with the coarse creatinine baseline estimates used, patients with an underestimated baseline are directly classified in the injury or failure group and are not represented in the model because no kidney risk episode precedes the kidney injury classification. Similarly, patients with an overestimated creatinine baseline may only satisfy the RIFLE risk category while experiencing renal failure. Providing a better baseline creatinine estimate might add cases that provide additional creatinine-based injury classifications and adjust the model's current emphasis on urine output. In other cases, a patient's renal function rapidly deteriorated and the patient transitioned directly from no kidney risk to kidney injury. Such transitions were observed a number of times, and indicate that my inclusion criteria would likely benefit from a more inclusive definition of kidney risk than that specified by the RIFLE criteria. My model is currently not evaluated against acute kidney injury episodes that are not preceded by kidney risk as defined by RIFLE.

In conclusion, the prediction model developed in this section does a moderate job of predicting acute kidney injury. By examining the predictions from our model, however, it appears that warnings more than 6 hours in advance are difficult. In contrast, the `RAS` model tends to peak between 12 and 6 hours before the kidney injury classification is made, but drops for the 6 hours where the `AKIM` predictions tend to rise.

6.6 Other Outcomes

6.6.1 Weaning of Mechanical Ventilator

Many of the ICU patients in the data receive mechanical ventilation support. In general, when patients are mechanically ventilated, CareVue has fields that indicate the ventilator type and the ventilator mode that are updated every 4 hours (see Figure 6-65). It is not uncommon for the charting interval to be 5 hours and sometimes 6 hours. Long delays between ventilator status entries make the task of finding the hour that the patient was weaned difficult. As noted in Chapter 3, in my data preparations I used a 5 hour hold window to retain ventilator information between intermittent updates in my dataset. Consequently, the exact point of weaning was difficult to discern. Further work is necessary to better isolate when exactly a patient was weaned from the mechanical ventilator. If an accurate timestamp for ventilator removal could be found, predicting when the ventilator weaning happens would provide an interesting regression task.

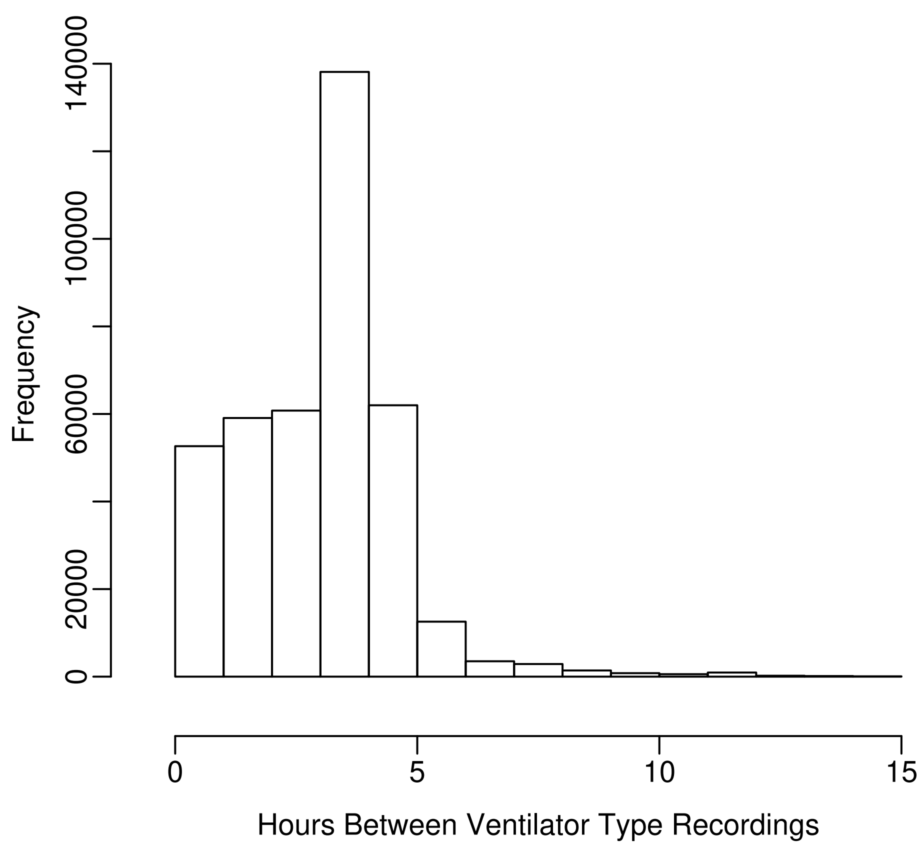


Figure 6-65: Ventilator type recording frequency

6.6.2 Tracheotomy Insertion

One relatively common procedure in the ICU is the surgical insertion of an opening that provides direct access to a patient's trachea. The tracheotomy procedure is performed for a variety of reasons, e.g., to assist in ventilator weaning, and provides a number of benefits over endotracheal intubation (especially when needed for longer periods of time).

In the dataset that we consider in this report (limited to 7 days), a relatively small number of patients receive a tracheotomy. Several patients have tracheotomies upon admission but admission cases are not helpful for training predictive models. If the data are examined without the 7-day limitation, most of the tracheotomy insertions occurred for patients who have long stays. But using my dataset, which included only 90 tracheotomy insertions in the development data and 23 insertions in the validation data, it was difficult to train a model to predict tracheotomy insertions. Without the 7-day limitation, predicting tracheotomy insertion may provide an interesting modeling problem.

6.6.3 Pressor Dependence

Possibly of more clinical interest than whether or not a patient is entirely weaned from pressors within a fixed time period, is how a patient responds when pressors are first reduced. A model that could predict the nature of the first response to a weaning attempt could be useful for determining which patients are ready to be weaned. Given a patient on a high dose of pressors, can I predict if he or she will tolerate a 25% reduction in pressor dosage?

As an initial step in exploring pressor dependence, Figure 6-66 shows a discretized joint distribution between the relative change in pressors and the time from the first high pressor dosage recorded for each patient. Only the first high dose pressor episode is used for each patient; if a patient has pressors removed for at least 4 hours, the remaining data for that patient are ignored. If pressors are removed for less than 4 hours, the drop is ignored and additional pressor doses are still represented relative to the first high dose (i.e., 0 on the y-axis). Similarly, Figure 6-67 shows the same joint distribution for a much lower dose-inclusion threshold.

Figures 6-66 and 6-67 are discretized into fixed 3-hour periods along the x-axis and 10% or 20% change increments along the y-axis. the median time until a 50% reduction in pressors and the median time until a 90% reduction in pressors are provided on each plot. By looking at a fixed value along either the x-axis or the y-axis, one can gain insight into the distribution that follows from conditioning on the fixed value.

In Figure 6-66, for example, the distribution for time conditioned on a 45-55% drop in pressor infusions has a mode of about 5 hours and a median of about 21 hours. Similarly, one might examine the distribution of relative pressor changes after 37.5 to 40.5 hours. For this case, the conditional distribution is skewed towards decreased pressor dosages with a mode of about -80% but it also has a long tail with a number of cases where the pressors rise by more than 50%. The most common path for relative pressor doses, starting at the origin (relative pressor change between -5% and 5%, time < 1.5 hours), is to decrease along the ridge of the contours shown in the plot, quickly falling to about -60% within 6 hours. At this point, the ridge is not as clear, but the pressors continue to decline albeit at a generally reduced rate. The final reduction from -75% to -100% indicates diffuse weaning trajectories at these lower pressor infusion rates.

The differences in the joint distributions can also be stratified by the final outcome of the patient. Figures 6-68 and 6-69 show the pressor change versus time for patients who died and lived, respectively. As expected, patients who live have a clear downward trajectory in pressor dosages. The downward trend is less evident for patients who do not survive.

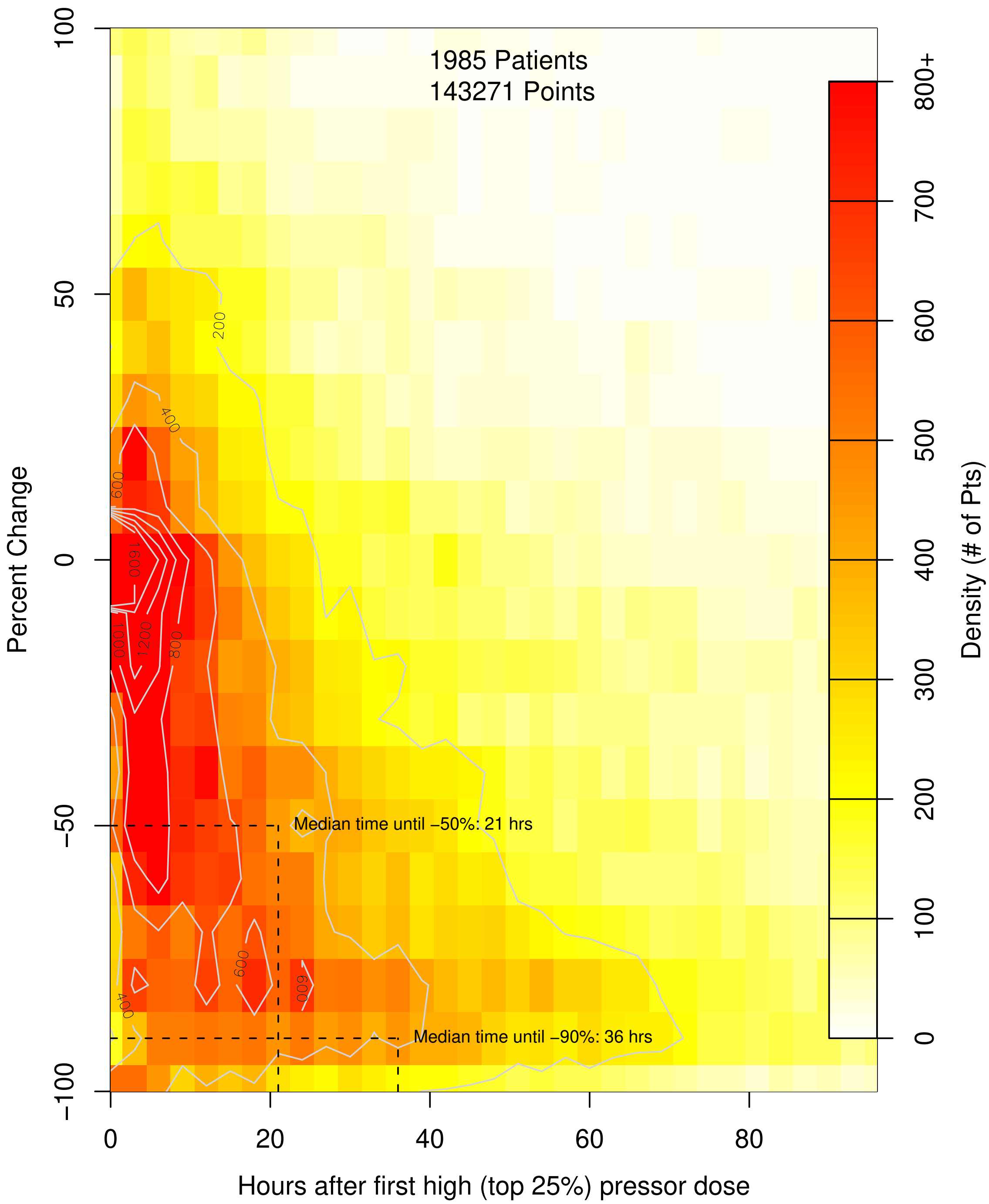


Figure 6-66: Joint density showing relationship between relative changes in pressors and time following first large (top 25%) pressor infusion. Each cell represents a 10% change over a period of 3 hours. Pressor changes following four hours with no pressors are ignored.

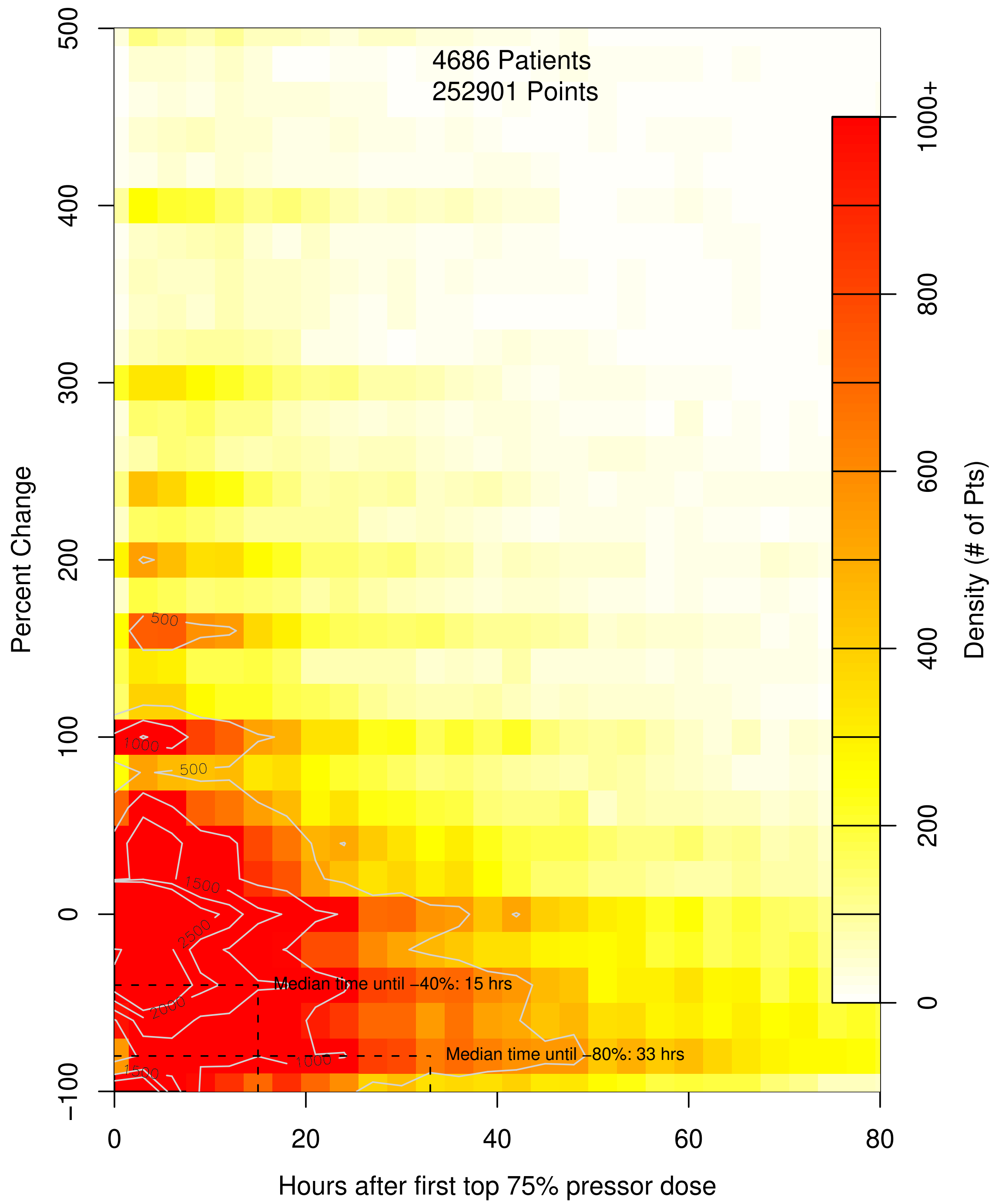


Figure 6-67: Joint density showing relationship between relative changes in pressors and time following first moderate-high (top 75%) pressor infusion. Each cell represents a 20% change over a period of 3 hours. Pressor changes following four hours with no pressors are ignored.

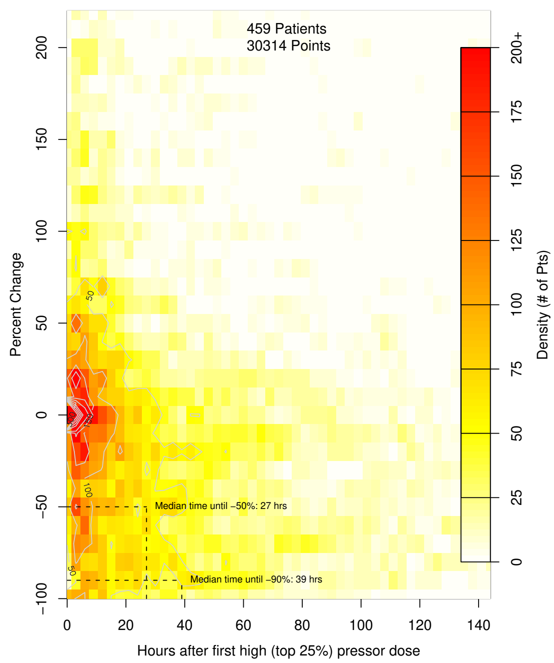


Figure 6-68: Non-survivors. Joint density for pressor changes and time following first large pressor infusions. Pressor changes following four hours without pressors are ignored.

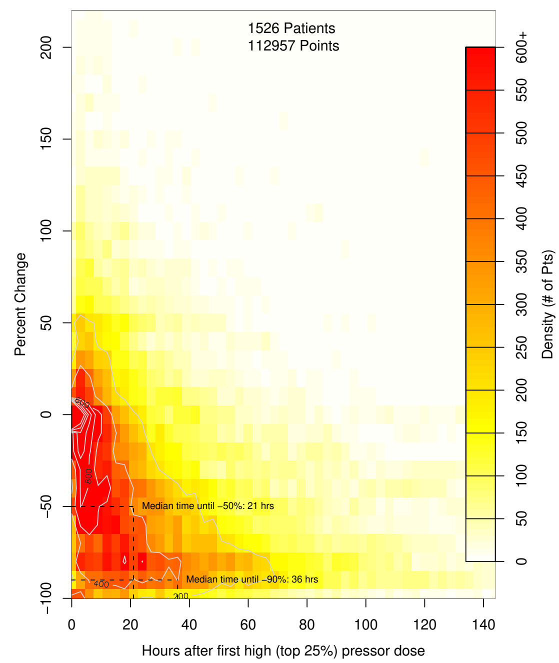


Figure 6-69: Survivors. Joint density for pressor changes and time following first large pressor infusions. Pressor changes following four hours without pressors are ignored.

6.7 Conclusion

The secondary outcome models developed in this chapter illustrate a variety of the prediction tasks that are possible with the framework that we established in Chapter 4. On the separate validation patients, the models generally performed well with ROC areas of 0.809 (pressor wean model), 0.825 (pressor wean *and* live model), 0.816 (IABP wean model), 0.843 (septic shock onset model), and 0.742 (acute kidney injury model). In terms of calibration, the probabilities produced by the models were generally less reliable with highly significant values for all of the Hosmer-Lemeshow tests considered. The calibration plots, however, generally reflected acceptable calibration. These calibration observations were expected given similar calibration behavior displayed by the real-time model (`RAS`) developed in Chapter 5.

The more specific populations of patients that each secondary outcome model examined resulted in a variety of interesting inputs previously not utilized in the mortality models considered in Chapter 5. The `BPWM` model, for example, included six 28-hour trend variables as it attempted to understand the patient's response to weaning of his or her intraaortic balloon pump.

The usefulness of the `RAS` model varied between the secondary outcomes considered. In terms of AUC for secondary outcomes, `RAS` predictions resulted in an area of 0.679 for the `PWM` outcome, 0.727 for the `PWLM` outcome, 0.679 for the `BPWM` outcome, 0.587 for the `SSOM` outcome, and 0.495 for the `AKIM` outcome. The pressor wean *and* live model outcome is the only outcome that had an AUC > 0.7 using `RAS` predictions. The `RAS` performance is reasonably strong if one considers the wide range of baseline risks associated with the secondary outcomes and the diversity of patients included by each model. With the `AKIM` model, for example, there is a pattern of higher `RAS` predictions about 12 hours before classification of kidney injury.

Based on the findings in this chapter, a number of conclusions can be drawn. First, with enough data it is possible to develop models that perform well at predicting intermediate or secondary ICU outcomes. These models are better at predicting their respective outcomes than a general model that is trained to predict mortality. For some of the secondary outcomes, however, general mortality model predictions appear to be strongly correlated with the predictions of a specialized model. Second, the general acuity score synthesizes a variety of mortality risk factors, so it is often much less sensitive to the fixed prediction windows used to train and validate models that predict secondary outcomes. Instead, the general acuity score focuses on the severity of the patient's condition (which often increases in the context of a severe secondary event). While a septic shock warning might look similar between a generally stable septic patient and a septic patient with chronic heart failure, the baseline risk profiles between the patients is likely to be quite different. In fact, for the acute kidney injury prediction task the general acuity model provided essentially no early warning assistance for the secondary event. Similarly, despite a significant increase in the `RAS` mortality probabilities during septic episodes, the `RAS` predictions provided only marginal utility for the septic shock onset prediction as defined.

For decision support it may be beneficial to look at the output from a specialized model in conjunction with a general acuity model. Such an approach might help caregivers decide on the likely cause of an increase in mortality risk and how to proceed. The positive predictive value requirements for a secondary warning model may decrease with a patient's mortality risk. For example, as a patient's mortality risk increases, the acceptable threshold for acting on a septic shock warning may decrease.

One important observation is that it is difficult to make clinically meaningful definitions for use in training the models. The models that rely on a fixed outcome that can be directly inferred from the data (e.g. successful pressor wean or IABP wean) have a significant advantage. Other definitions, such as those used for the `SSOM` model, are much more sophisticated and include a variety of parameters that can be adjusted. In addition, careful consideration is needed in developing the inclusion criteria used by the secondary outcome definitions to remove patients that might weaken the model's clinical utility (such as patients weaned from pressors following a surgical procedure). With careful consideration, however, models that rely on more complex outcome definitions (e.g., classifying pressor-dependent patients) will likely prove to be the most beneficial to clinical practice.

Chapter 7

Conclusion

7.1 Summary of Contributions

Large bodies of rich ICU data have only recently been available for data mining and analysis. With ongoing advances in computing technology, digital charting, disk storage capacities, and electronic medical records, ICU data archives continue to grow in size, quality, and scope.

In this thesis, I utilized one such resource — namely, the MIMIC II database — to develop predictive models for ICU patients. Two categories of models were explored. First, I developed general models that predicted patient mortality. Second, I developed models that predicted a number of intermediate patient outcomes.

7.1.1 Mortality Models

For the mortality prediction task, I compared four types of models at predicting patient mortality. First, I trained a model that used aggregate daily data to predict mortality for any ICU day (`SDAS`). Next, I developed five daily models that were specialized for ICU day n (`DAS`n for $n \in \{1, 2, 3, 4, 5\}$). Third, I developed a real-time acuity model that utilized any unique observation to make a prediction of mortality (`RAS`). Finally, I created a customized SAPS II algorithm for comparison ($\texttt{SAPSII}_a$).

I found that for each day, AUC validation performance was significantly greater ($p < 0.05$) for my mortality models developed on MIMIC II patient data than it was for the MIMIC II-customized $\texttt{SAPSII}_a$ model. Furthermore, no significant performance difference was observed between the `RAS` model and the other models that were based on aggregate daily data. The performance of `RAS` was somewhat surprising as the task of predicting mortality based on daily data is presumably easier than predicting mortality based on intra-day moment-to-moment observations. The greater temporal granularity available from the `RAS` model allows one to track intra-day changes during periods of clinical interest. The `RAS` model might, for example, indicate a strong decreasing trend over the course of a day (e.g., due to the resolution of cardiogenic shock). In contrast, a daily score is forced to summarize intra-day trends with a single number and thereby offers less clinical insight into the patient's changing risk.

These findings indicate that real-time risk assessment for ICU patients is feasible. While others have augmented existing severity of illness scores with a few frequently updated inputs (e.g., [82, 67]), as far as I know, no one else has yet explored models that use rich patient data to provide estimates that are updated more frequently than once per day.

My real-time mortality model also includes a variety of complex inputs that are customized to individual patients. Two examples of complex inputs include: (1) trend information such as the platelet slope over the past 28 hours and (2) the number of threshold events that occurred for a particular observation in recent history such as the number of times the SpO_2 fell below 90% in the past 2 hours. Several of these complex inputs provide important summaries of observations and make a significant contribution to the model's understanding of patient risk. Most existing severity scores, with their emphasis on simplicity, do not take advantage of such computationally-complex inputs.

The inclusion of therapeutic interventions provides another important distinguishing characteristic for my real-time acuity model. Most existing severity of illness metrics ignore interventions and instead focus solely on physiologic observations. One notable exception is the SOFA score which includes vasopressor administration, along with the mean arterial pressure, to assess a patient's cardiovascular system [92, 91]. The other model that uses some intervention input (albeit in a limited manner) is the MPM model. The MPM_{24} and MPM_{48} models each include the number of hours with mechanical ventilation, and the 24-hour version includes the "number of lines" while the 48-hour version includes hours of vasoactive IV drugs. Apart from SOFA the later MPM models, other common scores explicitly exclude interventions from consideration as score inputs. While including interventions made by caregivers may worry someone who wants to only observe physiologic indicators, the changes brought by interventions often influence how one interprets a patient's observations. In the `RAS` model that I developed, the most influential intervention input was the quantity of pressors that the patient was receiving. The importance of this input, however, was lower than might be expected; six physiologic variables were more influential in the model than the pressor level. Nonetheless, the pressor administration was helpful in interpreting the patient's risk profile. Knowledge of pressor dose is often necessary to properly interpret physiologic observations during an acute event that requires aggressive treatment. Furthermore, the differences between inputs for the daily acuity model for ICU day 1 (`DAS1`) and subsequent daily models indicate that intervention considerations are more important for later ICU days. From these observations I can conclude that therapeutic interventions were especially important for providing meaningful intra-day predictions and predictions following the initial 24 hours of a patient's ICU stay.

7.1.2 Secondary Outcome Models

Chapter 6 discussed models that were trained to predict five separate secondary outcomes. These secondary outcomes represent clinically significant events that might occur during a patient's ICU stay. The predictive models included (1) the successful weaning of pressors (`PWM`), (2) the successful weaning of pressors and ICU survival (`PWLM`), (3) the removal of an intraaortic balloon pump (IABP) (`BPWM`), (4) the onset of septic shock (`SSOM`), and (5) acute kidney injury (`AKIM`):

1. The `PWM` model predicted if a patient would be successfully weaned from pressors within 12 hours. On the separate validation data, the `PWM` model performed well with an AUC of 0.809. To predict weaning, the `PWM` model relied heavily on the level of pressors being administered, how long the patient had received pressors, and several other intervention inputs. A number of physiologic variables were also important such as the Glasgow Coma Scale and the creatinine level.

2. The `PWLM` used an augmented version of the `PWM` outcome to include patient survival in addition to weaning of pressors within 12 hours. On this slightly different prediction task, the `PWLM` model obtained an AUC of 0.825 on the separate validation data. The prognostic utility of this model is not clear, but it allows a contrast to be drawn with the `PWM` model in that it does not penalize for patient weans that in reality may not have truly been "successful". The set of inputs selected for the outcome of *pressor wean within 12 hours and survival* included many of the inputs selected for the `PWM` model.

3. The `BPWM` model predicted the successful removal of an IABP pump within 12 hours (`BPWM`). The `BPWM` predictions on the separate validation data yielded an AUC of 0.816. To make predictions, the `BPWM` model relied on a number of therapies that were indicative of IABP weaning but also a number of physiologic trends that indicated how the patient's heart was responding to increased afterloads.

4. The `SSOM` model predicted the onset of septic shock within 12 hours. Septic shock was defined as hypotension despite fluid resuscitation (HDFR) in addition to the systemic inflammatory response syndrome (SIRS). The model attempted to identify transitions from SIRS only to SIRS and HDFR. The performance for the `SSOM` model was quite strong with an AUC of 0.843 on separate validation patients. To make the predictions, the `SSOM` model relied heavily on medication interventions (e.g., the time that the patient had spent on pressors) and obvious physiologic inputs such as the shock index (heart rate divided by systolic blood pressure).

5. The `AKIM` model used the RIFLE kidney classification scheme to predict the transition from kidney risk to kidney injury within 12 hours. The performance

of the `AKIM` model was weaker than the other models, but it still managed to obtain an AUC of 0.742 on the separate validation patients. To predict kidney injury, the model had a strong focus on urine output, fluid balance and the mean arterial blood pressure.

For each of the secondary models described above, I examined the performance of the `RAS` model in the context of the secondary event. Specifically, I compared the performance of using the `RAS` mortality estimate as an indicator of the secondary event. This resulted in the following AUC numbers: (1) AUC of 0.679 for pressor wean in 12 hours, (2) AUC of 0.727 for pressor wean in 12 hours and survival, (3) AUC of 0.679 for IABP removal within 12 hours, (4) AUC of 0.587 for the onset of septic shock within 12 hours, and (5) AUC of 0.495 for kidney injury within 12 hours. The first three of these performance numbers indicate the `RAS` risk estimates are reasonably correlated with the estimates produced by the specialized models. For the last two outcomes of septic shock and kidney injury the `RAS` prediction was less useful. In general, it is expected that models trained on specific outcomes would be more sensitive to those outcomes. Consequently, it may be most useful to use secondary outcome predictions in conjunction with `RAS` estimates for early warnings of specific pathologies in the context of a patient's general risk assessment.

Appendix E provides 40 randomly selected patients (20 who expired and 20 who survived) to illustrate the application of each of my models on individual patients.

7.2 Limitations and Future Work

A number of limitations exist with the present work. In addition to these limitations, over the course of this work several ideas for further exploration were identified. For discussion, I divide these limitations and areas for future work into three categories: mortality models, secondary models, and general methodology.

Mortality models While the real-time mortality model presented in this report generally performed well, it is important to be clear about its many limitations. First, the model did not demonstrate perfect calibration. The high probability estimates were particularly unreliable. In fact, for the `RAS` model on the validation data, the highest probability estimates often needed significant correction to align with the actual mortality probabilities. Many high estimates come from short-lived "peaks" in an otherwise generally low risk profile. Abrupt changes should be accompanied by additional uncertainty especially if the cause of the change is known and expected to be ephemeral.[1] Likewise, it is important to understand that low risk estimates have

[1] If the `RAS` predictions are averaged by taking the mean prediction for each patient, this problem is greatly mitigated. A variety of smoothing techniques could also be applied to decrease local variance.

better, but still imperfect, calibration. This is especially important in the intensive care setting where every patient has an increased mortality risk and could deteriorate quickly. A patient's current observations might reflect stability, but they may have other developing or ongoing conditions that concern caregivers. A patient with a failing heart, for example, may receive an IABP and stabilize well. The same patient's ICU survival, however, may be contingent on a successful high-risk heart surgery. Important context information, often limited to unstructured or semi-structured free text, is generally lacking from the MIMIC II data that my model is built from. For these reasons, the `RAS` score should be interpreted along the lines of "the patient's current clinical profile is similar to other patients who had a mortality risk of x", and the individual considerations for the patient should continue to guide therapy.

With a wide range of patient conditions, it is essential to remember that a model will simply not account for some patients. The exact composition of my patient population, in relation to mortality risk estimation, merits further investigation. That is, are there specific categories of patients where the model consistently underestimates or overestimates patient risk? One such category of "difficult" patients that I identified early in this work was severe head injury patients. Often the most important observations for such patients are the head scans that reflect the amount of intracranial pressure. Image data were unavailable in the MIMIC II data that I considered. Without such information, many of high-risk head trauma patients appeared to present normal physiologic observations. Identifying other such groups by thoroughly examining the population used for building the `RAS` model could reduce the number of patients that have an inherently different risk profile requiring special observations and clinical considerations.

The definition of mortality used in this thesis represents an interesting area for further exploration. For example, the advantage of using a 30-day mortality window over a 10-day mortality window is unclear. A perfect prediction of mortality (despite being impossible) would not necessarily be useful for patient tracking. There appears to be a trade-off between acute risk and baseline mortality risk. If one considers the `RAS` model as a similarity metric, by increasing the mortality window, the score should reflect more of a baseline mortality risk (taken to the extreme, everyone eventually dies). On the other hand, if the mortality window is short, the score is weighted more toward acute risk. With too short of a window, however, many individuals — such as elderly individuals with dire chronic illnesses who are successfully stabilized for a short while — are perhaps considered less severely ill because they are not rapidly deteriorating.

Secondary Outcomes The secondary outcome models presented in this report also have a variety of limitations and areas for future work. Further refinement of the secondary outcome models could help maximize their clinical utility. A number of adjustments to the secondary model prediction tasks are possible: (1) the length window used for early warning, (2) the inclusion criteria for episodes that can have warnings issued (e.g., SIRS), and (3) the definition of the events of interest.

With the `SSOM` model I observed that the most difficult episodes were often also the most interesting. The first onset of septic shock, for example, was often not preceded by SIRS. In other cases that I manually reviewed, it seemed that there was little doubt that the shock episode was cardiogenic shock and not septic shock. This is supported by the presence of beta blockers as an input in the `SSOM` model. Further refinement, such as loosening the SIRS requirement or further restricting the HDFR definition, might allow the `SSOM` model to target a more specific epidemiology.

General Methodology A number of methodological limitations also exist in this work. Many of these represent areas for further exploration. A selection of important limitations include:

- **External validation** The results of this work are confined to a single hospital's ICU population. Before fully generalizing my results it is necessary to validate them on external data.

- **Independence assumptions** While a number of variables were included to try to summarize the temporal dynamics of a patient, a strong assumption is made to consider subsequent patient observations as independent of each other. One might consider an alternative modeling technique, such as a hidden Markov model, that has "memory" of recent observations. This could potentially lead to better prediction performance and a smoother model with less local variation.

- **Therapeutic interventions** It is important that clinical application of models such as the ones developed in this thesis proceed with an understanding of the limitations that come by including patient interventions. While it is assumed that caregivers generally make the correct decisions, it is possible that a model which relies on caregiver interventions could propagate suboptimal treatment. This is felt to be somewhat mitigated by the large sample sizes and variance between individual caregivers in the data, but it is still a concern. To better understand the role that therapies take in my models, therapy-free models could be built to contrast the exact benefits of including caregiver therapies in predictive models.

- **Standardized pressor measurements** The method I used for combining pressor medications into an overall pressor level indicator (pressorSum.std) warrants further exploration. Instead of combining pressor medications based on their general dosage patterns, one might consider a more principled approach by analyzing dose response curves and the interactive effects of combining multiple pressors.

- **Other machine learning techniques** In addition to logistic regression, a number of other advanced machine learning algorithms exist. As the MIMIC II data is refined, such algorithms may present attractive alternatives. For most of our models, with the large quantity of data available (about 10000 patients), overfitting was not a significant problem. More powerful techniques might be used to create better performing predictive models. From my experience, however, *transparent* models are nearly essential as the data continue to be refined and better understood. The ability of a Bayesian network to model nonlinear relationships and handle missing data, for example, might prove to be a useful modeling technique for future work.

- **Missing observations** The sparsity of the data matrix used for my modeling could be reduced by applying sophisticated imputation techniques. This would be especially helpful for predicting secondary outcomes that occur rarely and therefore have severely limited data. Imputation on other predictive inputs, such as the bilirubin level — which my current methodology often excludes based on infrequent availability — might allow additional variables to be included in my modeling process and further enhance my predictive performance. More complete data would also allow other techniques such as principal component analysis that do not tolerate missing observations.

7.3 Conclusion

Real-time mortality prediction is a feasible way to provide continuous risk assessment for ICU patients. RAS offers similar discrimination ability when compared to models computed once per day, based on aggregate data over that day. Moreover, RAS mortality predictions are better at discrimination than a customized SAPS II score (Day 3 AUC=0.878 vs AUC=0.849, $p < 0.05$). The secondary outcome models also provide interesting insights into patient responses to care and patient risk profiles. While models trained for specifically recognizing secondary outcomes consistently outperform the RAS model at their specific tasks, RAS provides useful baseline risk estimates throughout these events and in some cases offers a notable level of predictive utility. By providing a similarity measure between a patient and other patients who ultimately died, the `RAS` model offers a succinct summary of a patient's acuity. The availability of a real-time acuity summary may affect the use of other models that predict specific pathologies which develop over a diverse population with a wide range of risks and confounders. Additional work remains to be done in order to better understand the future clinical utility of such real-time acuity models.

While much work remains to be done to add to and improve the work described in this report, my results contribute to the following broad ideas: (1) relatively simple modeling frameworks can produce highly predictive models using a large volume multi-resolution temporal ICU data; (2) real-time risk assessment is feasible in the ICU; and (3) generic patient tracking is a reasonable goal that can be advanced through a variety of predictive models and real-time risk models may play an important role in this advancement.

Bibliography

[1] M. Antonelli, R. Moreno, J. L. Vincent, C. L. Sprung, A. Mendoa, M. Passariello, L. Riccioni, and J. Osborn. Application of SOFA score to trauma patients. Sequential organ failure assessment. *Intensive Care Med*, 25(4):389–394, Apr 1999.

[2] T. K. Bavin and M. A. Self. Weaning from intra-aortic balloon pump support. *Am J Nurs*, 91(10):54–59, Oct 1991.

[3] Rinaldo Bellomo, Claudio Ronco, John A Kellum, Ravindra L Mehta, Paul Palevsky, and Acute Dialysis Quality Initiative workgroup. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the second international consensus conference of the acute dialysis quality initiative (adqi) group. *Crit Care*, 8(4):R204–R212, Aug 2004.

[4] H Bolooki. *Clinicial Application of the Intra-Aortic Balloon Pump*. Futura Publishing Co., Mount Kisco, NY, 2nd edition, 1984.

[5] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald. 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. *Chest*, 101(6):1644–1655, Jun 1992.

[6] Daliana Peres Bota, Christian Melot, Flavio Lopes Ferreira, Vinh Nguyen Ba, and Jean-Louis Vincent. The multiple organ dysfunction score (MODS) versus the sequential organ failure assessment (SOFA) score in outcome prediction. *Intensive Care Med*, 28(11):1619–1624, Nov 2002.

[7] Thomas A Buckley, Charles D Gomersall, and Sarah J Ramsay. Validation of the multiple organ dysfunction (MOD) score in critically ill medical and surgical patients. *Intensive Care Med*, 29(12):2216–2222, Dec 2003.

[8] Hanqing Cao, Larry Eshelman, Nicolas Chbat, Larry Nielsen, Brian Gross, and Mohammed Saeed. Predicting icu hemodynamic instability using continuous multiparameter trends. *Conf Proc IEEE Eng Med Biol Soc*, 2008:3803–3806, 2008.

[9] R. W. Chang. Individual outcome prediction models for intensive care units. *Lancet*, 2(8655):143–146, Jul 1989.

[10] Children's Hospital Informatics Program. Indivo. Available at: http://www.indivohealth.org. Accessed on August 28, 2008.

[11] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3-4):562–565, 1958.

[12] Carol J DeFrances, Karen A Cullen, and Lola Jean Kozak. National hospital discharge survey: 2005 annual summary with detailed diagnosis and procedure data. *Vital Health Stat 13*, (165):1–209, Dec 2007.

[13] Pierre Delanaye, Jean-Marie Krzesinski, Etienne Cavalier, and Bernard Lambermont. The rifle criteria: are the foundations robust? *Crit Care Med*, 35(11):2669; author reply 2669–2669; author reply 2670, Nov 2007.

[14] R. Phillip Dellinger, Mitchell M Levy, Jean M Carlet, Julian Bion, Margaret M Parker, Roman Jaeschke, Konrad Reinhart, Derek C Angus, Christian Brun-Buisson, Richard Beale, Thierry Calandra, Jean-Francois Dhainaut, Herwig Gerlach, Maurene Harvey, John J Marini, John Marshall, Marco Ranieri, Graham Ramsay, Jonathan Sevransky, B. Taylor Thompson, Sean Townsend, Jeffrey S Vender, Janice L Zimmerman, Jean-Louis Vincent, International Surviving Sepsis Campaign Guidelines Committee, American Association of Critical-Care Nurses, American College of Chest Physicians, American College of Emergency Physicians, Canadian Critical Care Society, European Society of Clinical Microbiology, Infectious Diseases, European Society of Intensive Care Medicine, European Respiratory Society, International Sepsis Forum, Japanese Association for Acute Medicine, Japanese Society of Intensive Care Medicine, Society of Critical Care Medicine, Society of Hospital Medicine, Surgical Infection Society, World Federation of Societies of Intensive, and Critical Care Medicine. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Crit Care Med*, 36(1):296–327, Jan 2008.

[15] Viktor Y Dombrovskiy, Andrew A Martin, Jagadeeshan Sunderram, and Harold L Paz. Rapid increase in hospitalization and mortality rates for severe sepsis in the united states: a trend analysis from 1993 to 2003. *Crit Care Med*, 35(5):1244–1250, May 2007.

[16] Colleen M Ennett, K. P. Lee, Larry J Eshelman, Brian Gross, Larry Nielsen, Joseph J Frassica, and Mohammed Saeed. Predicting respiratory instability in the icu. *Conf Proc IEEE Eng Med Biol Soc*, 2008:2848–2851, 2008.

[17] Larry J Eshelman, K. P. Lee, Joseph J Frassica, Wei Zong, Larry Nielsen, and Mohammed Saeed. Development and evaluation of predictive alerts for hemodynamic instability in icu patients. *AMIA Annu Symp Proc*, pages 379–383, 2008.

[18] F. L. Ferreira, D. P. Bota, A. Bross, C. Mlot, and J. L. Vincent. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA*, 286(14):1754–1758, Oct 2001.

[19] National Kidney Foundation. K/doqi clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis*, 39(2 Suppl 1):S1–266, Feb 2002.

[20] Google Corp. Google Health. Available at: https://www.google.com/health. Accessed on August 28, 2008.

[21] R. J. Goris, T. P. te Boekhorst, J. K. Nuytinck, and J. S. Gimbrre. Multiple-organ failure. generalized autodestructive inflammation? *Arch Surg*, 120(10):1109–1115, Oct 1985.

[22] John D Halamka, Kenneth D Mandl, and Paul C Tang. Early experiences with personal health records. *J Am Med Inform Assoc*, 15(1):1–7, 2008.

[23] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modelling strategies for improved prognostic prediction. *Stat Med*, 3(2):143–152, 1984.

[24] Frank E. Harrell. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis.* Springer series in statistics. Springer, New York, 2001.

[25] Frank E Harrell, Jr. *Design: Design Package*, 2005. R package version 2.0-12.

[26] Khosro Hekmat, Axel Kroener, Hartmut Stuetzer, Robert H G Schwinger, Sandra Kampe, Gerardus B W E Bennink, and Uwe Mehlhorn. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. *Ann Thorac Surg*, 79(5):1555–1562, May 2005.

[27] Thomas Higgins. Daily versus admission mortality estimates: Is admission severity yesterday's news? *Crit Care Med*, 29(1):202–219, Jan 2001.

[28] Kalon Ho. Re: Predictive models for iabp weaning. email correspondence, April 2009.

[29] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* New York: Wiley, 2nd edition, 2000.

[30] HSM Group, Ltd. Acute care hospital survey of RN vacancy and turnover rates in 2000. *J Nurs Adm*, 32(9):437–439, Sep 2002.

[31] Caleb Hug. Predicting the risk and trajectory of intensive care patients using survival models. Master's thesis, Massachusetts Institute of Technology, 2006.

[32] Caleb Hug and Gari D Clifford. An analysis of the errors in recorded heart rate and blood pressure in the ICU using a complex set of signal quality metrics. In *Computers in Cardiology 2007*, volume 34, pages 641–644, 2007.

[33] Caleb Hug, Gari D Clifford, and Andrew T Reisner. Clinician blood pressure documentation of stable ICU patients: an intelligent archiving agent has a higher association with future hypotension. *JAMIA*, 2009. In submission.

[34] A Kantrowitz, R R Cardona, and P S Freed. *Comprehensive Intraaortic Balloon Counterpulsation*. Mosby, Sydney, 2nd edition, 1993.

[35] M. Kayaalp, G. F. Cooper, and G. Clermont. Predicting ICU mortality: a comparison of stationary and nonstationary temporal models. *Proc AMIA Symp*, pages 418–422, 2000.

[36] Mark A Kelley, Derek Angus, Donald B Chalfin, Edward D Crandall, David Ingbar, Wanda Johanson, Justine Medina, Curtis N Sessler, and Jeffery S Vender. The critical care crisis in the United States: a report from the profession. *Chest*, 125(4):1514–1517, Apr 2004.

[37] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: a severity of disease classification system. *Crit Care Med*, 13(10):818–829, Oct 1985.

[38] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100(6):1619–1636, Dec 1991.

[39] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*, 9(8):591–597, Aug 1981.

[40] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson. *To Err is Human: Building a Safer Health System*. Institute of Medicine. Washington, DC: National Academy Press, 1999.

[41] Kenneth Krell. Critical care workforce. *Crit Care Med*, 36(4):1350–1353, Apr 2008.

[42] Thomas A Lasko, Jui G Bhagwat, Kelly H Zou, and Lucila Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*, 38(5):404–415, Oct 2005.

[43] J. R. Le Gall, J. Klar, S. Lemeshow, F. Saulnier, C. Alberti, A. Artigas, and D. Teres. The logistic organ dysfunction system. a new way to assess organ dysfunction in the intensive care unit. ICU scoring group. *JAMA*, 276(10):802–810, Sep 1996.

[44] J. R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963, 1993.

[45] J. R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for ICU patients. *Crit Care Med*, 12(11):975–977, Nov 1984.

[46] Jean Roger Le Gall, Anke Neumann, Franois Hemery, Jean Pierre Bleriot, Jean Pierre Fulgencio, Bernard Garrigues, Christian Gouzes, Eric Lepage, Pierre Moine, and Daniel Villers. Mortality prediction using SAPS II: an update for french intensive care units. *Crit Care*, 9(6):R645–R652, 2005.

[47] R. Lefering, B. Wolfrum, H. Wauer, and E. A M Neugebauer. Limitations of score-based daily outcome predictions in the individual intensive care patient. an example of the RIAHDH algorithm. *Inflamm Res*, 53 Suppl 2:S169–S174, Aug 2004.

[48] Rolf Lefering, R. Jan A Goris, Ernst J van Nieuwenhoven, and Edmund Neugebauer. Revision of the multiple organ failure score. *Langenbecks Arch Surg*, 387(1):14–20, Apr 2002.

[49] S. Lemeshow and D. W. Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*, 115(1):92–106, Jan 1982.

[50] S. Lemeshow, J. Klar, D. Teres, J. S. Avrunin, S. H. Gehlbach, J. Rapoport, and M. Rué. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. *Crit Care Med*, 22(9):1351–1358, Sep 1994.

[51] S. Lemeshow and J. R. Le Gall. Modeling the severity of illness of ICU patients. a systems update. *JAMA*, 272(13):1049–1055, Oct 1994.

[52] S. Lemeshow, D. Teres, J. S. Avrunin, and R. W. Gage. Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Crit Care Med*, 16(5):470–477, May 1988.

[53] Bruno Levy, Benjamin Dusang, Djillali Annane, Sebastien Gibot, Pierre-Edouard Bollaert, and College Interregional des Réanimateurs du Nord-Est. Cardiovascular response to dopamine and early prediction of outcome in septic shock: a prospective multiple-center study. *Crit Care Med*, 33(10):2172–2177, Oct 2005.

[54] P. A. Lewis and M. Courtney. Weaning intraaortic balloon counterpulsation: the evidence. *British Journal of Cardiac Nursing*, 1:385–389, 2006.

[55] Walter Linde-Zwirble and Derek Angus. Severe sepsis epidemiology: sampling, selection, and society. *Critical Care*, 8(4):222–226, 2004.

[56] J. C. Marshall, D. J. Cook, N. V. Christou, G. R. Bernard, C. L. Sprung, and W. J. Sibbald. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. *Crit Care Med*, 23(10):1638–1652, Oct 1995.

[57] Neil McIntosh. Intensive care monitoring: past, present and future. *Clin Med*, 2(4):349–355, 2002.

[58] C. Meredith and J. Edworthy. Are there too many alarms in the intensive care unit? An overview of the problems. *J Adv Nurs*, 21(1):15–20, Jan 1995.

[59] Philipp G H Metnitz, Fabienne Fieux, Barbara Jordan, Thomas Lang, Rui Moreno, and Jean-Roger Le Gall. Critically ill patients readmitted to intensive care units–lessons to learn? *Intensive Care Med*, 29(2):241–248, Feb 2003.

[60] Microsoft Corp. Microsoft HealthVault. Available at: http://www.healthvault.com. Accessed on August 28, 2008.

[61] M. E. Miller, S. L. Hui, and W. M. Tierney. Validation techniques for logistic regression models. *Stat Med*, 10(8):1213–1226, Aug 1991.

[62] S. Oda, H. Hirasawa, T. Sugai, H. Shiga, K. Nakanishi, N. Kitamura, T. Sadahiro, and T. Hirano. Comparison of sepsis-related organ failure assessment (SOFA) score and CIS (cellular injury score) for scoring of severity for patients with multiple organ dysfunction syndrome (MODS). *Intensive Care Med*, 26(12):1786–1793, Dec 2000.

[63] Lucila Ohno-Machado, Frederic S Resnic, and Michael E Matheny. Prognosis in critical care. *Annu Rev Biomed Eng*, 8:567–599, 2006.

[64] Gustavo Ospina-Tascón, Ricardo Cordioli, and Jean-Louis Vincent. What type of monitoring has been shown to improve outcomes in acutely ill patients? *Intensive Care Med*, Jan 2008.

[65] M. Palazzo. *Anaesthesia, Pain, Intensive Care and Emergency Medicine A. P. I. C. E.* Springer, 2007.

[66] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[67] Ricardo Rivera-Fernndez, Raoul Nap, Guillermo Vzquez-Mata, and Dinis Reis Miranda. Analysis of physiologic alterations in intensive care unit patients and their relationship with mortality. *J Crit Care*, 22(2):120–128, Jun 2007.

[68] Graeme Rocker, Deborah Cook, Peter Sjokvist, Bruce Weaver, Simon Finfer, Ellen McDonald, John Marshall, Anne Kirby, Mitchell Levy, Peter Dodek, Daren Heyland, Gordon Guyatt, and Level of Care Study Investigators and Canadian Critical Care Trials Group. Clinician predictions of intensive care unit mortality. *Crit Care Med*, 32(5):1149–1154, May 2004.

[69] J. Rogers and H. D. Fuller. Use of daily acute physiology and chronic health evaluation (APACHE) II scores to predict individual patient survival rate. *Crit Care Med*, 22(9):1402–1405, Sep 1994.

[70] A. L. Rosenberg, T. P. Hofer, R. A. Hayward, C. Strachan, and C. M. Watts. Who bounces back? physiologic and other predictors of intensive care unit readmission. *Crit Care Med*, 29(3):511–518, Mar 2001.

[71] Jeffrey M Rothschild, Christopher P Landrigan, John W Cronin, Rainu Kaushal, Steven W Lockley, Elisabeth Burdick, Peter H Stone, Craig M Lilly, Joel T Katz, Charles A Czeisler, and David W Bates. The critical care safety study: The incidence and nature of adverse events and serious medical errors in intensive care. *Crit Care Med*, 33(8):1694–1700, Aug 2005.

[72] M. Rué, S. Quintana, M. Alvarez, and A. Artigas. Daily assessment of severity of illness and mortality prediction for individual patients. *Crit Care Med*, 29(1):45–50, Jan 2001.

[73] U. E. Ruttimann. Statistical approaches to development and validation of predictive instruments. *Crit Care Clin*, 10(1):19–35, Jan 1994.

[74] Erica E Ryherd, Kerstin Persson Waye, and Linda Ljungkvist. Characterizing noise and perceived work environment in a neurological intensive care unit. *J Acoust Soc Am*, 123(2):747–756, Feb 2008.

[75] M. Saeed, C. Lieu, G. Raber, and R. G. Mark. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29:641–644, 2002.

[76] J. H. Schäfer, A. Maurer, F. Jochimsen, C. Emde, K. Wegscheider, H. R. Arntz, J. Heitz, B. Krell-Schroeder, and A. Distler. Outcome prediction models on

admission in a medical intensive care unit: do they predict individual outcome? *Crit Care Med*, 18(10):1111–1118, Oct 1990.

[77] David Schoenfeld. Survival methods, including those using competing risk analysis, are not appropriate for intensive care unit outcome studies. *Crit Care*, 10(1):103, Feb 2006.

[78] N. Scholz, K. Bsler, P. Saur, H. Burchardi, and S. Felder. Outcome prediction in critical care: physicians' prognoses vs. scoring systems. *Eur J Anaesthesiol*, 21(8):606–611, Aug 2004.

[79] Wouter Jan Schuiling, Al W de Weerd, Paul J W Dennesen, Ale Algra, and Gabril J E Rinkel. The simplified acute physiology score to predict outcome in patients with subarachnoid hemorrhage. *Neurosurgery*, 57(2):230–6; discussion 230–6, Aug 2005.

[80] H. P. Schuster, F. P. Schuster, P. Ritschel, S. Wilts, and K. F. Bodmann. The ability of the simplified acute physiology score (SAPS II) to predict outcome in coronary care patients. *Intensive Care Med*, 23(10):1056–1061, Oct 1997.

[81] Dewang Shavdia. Septic shock: Providing early warnings through multivariate logistic regression models. Master's thesis, Massachusetts Institute of Technology, 2007.

[82] Alvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, and Jose Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial Intelligence in Medicine*, 36:3, 2005.

[83] Tasnim Sinuff, Neill K J Adhikari, Deborah J Cook, Holger J Schünemann, Lauren E Griffith, Graeme Rocker, and Stephen D Walter. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. *Crit Care Med*, 34(3):878–885, Mar 2006.

[84] J. F. Timsit, J. P. Fosse, G. Troch, A. De Lassence, C. Alberti, M. Garrouste-Orgeas, E. Azoulay, S. Chevret, P. Moine, and Y. Cohen. Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. *Intensive Care Med*, 27(6):1012–1021, Jun 2001.

[85] Tudor Toma, Ameen Abu-Hanna, and Robert Bosman. Predicting mortality in the intensive care using episodes. In *IWINAC (1)*, pages 447–458, 2005.

[86] Tudor Toma, Ameen Abu-Hanna, and Robert-Jan Bosman. Discovery and inclusion of SOFA score episodes in mortality prediction. *J Biomed Inform*, 40(6):649–660, Dec 2007.

[87] Atnthony J. Trevor, Bertram G. Katzung, and Susan B. Masters. *Pharmacology*. McGraw Hill, 2005.

[88] Jeffrey C Trost and L. David Hillis. Intra-aortic balloon counterpulsation. *Am J Cardiol*, 97(9):1391–1398, May 2006.

[89] C. L. Tsien and J. C. Fackler. Poor prognosis for existing monitors in the intensive care unit. *Crit Care Med*, 25(4):614–619, Apr 1997.

[90] Christine L. Tsien. *TrendFinder: Automated Detection of Alarmable Trends*. PhD thesis, Massachusetts Institute of Technology, 2000.

[91] J. L. Vincent, A. de Mendona, F. Cantraine, R. Moreno, J. Takala, P. M. Suter, C. L. Sprung, F. Colardyn, and S. Blecher. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Crit Care Med*, 26(11):1793–1800, Nov 1998.

[92] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*, 22(7):707–710, Jul 1996.

[93] Jean-Louis Vincent, Edward Abraham, Djillali Annane, Gordon Bernard, Emanuel Rivers, and Greet Van den Berghe. Reducing mortality in sepsis: new directions. *Crit Care*, 6 Suppl 3:S1–18, Dec 2002.

[94] D. P. Wagner, W. A. Knaus, F. E. Harrell, J. E. Zimmerman, and C. Watts. Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. *Crit Care Med*, 22(9):1359–1372, Sep 1994.

[95] W. Zong, G. B. Moody, and R. G. Mark. Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure. *Med Biol Eng Comput*, 42(5):698–706, Sep 2004.

[96] David A Zygun, Kevin B Laupland, Gordon H Fick, James Dean Sandham, and Christopher J Doig. Limited ability of SOFA and MOD scores to discriminate outcome: a prospective evaluation in 1,436 patients. *Can J Anaesth*, 52(3):302–308, Mar 2005.

Appendix A

Summary of Statistical Measures

A variety of statistical measures are referenced in this thesis. A summary of these statistics are provided in Table A.1.

Table A.1: Statistical Measures

Statistic	Description
D_{xy}	The Somers' D_{xy} rank correlation between the predicted outcome and the actual outcome
R^2	Nagelkerke-Cox-Snell-Maddala-Magee R^2 index
γ_0	*Intercept* for the fitted logistic calibration curve
γ_1	*Slope* for the fitted logistic calibration curve
$E_{\max}$	The maximum absolute difference in predicted and calibrated probabilities $E_{\max}$
D	The discrimination index D [(model L.R. $\chi^2 - 1$)/n]
U	The unreliability index (difference in -2 log likelihood between uncalibrated $X\beta$ terms and the $X\beta$ terms with overall intercept and slope calibrated to the test sample)$/n$
Q	the overall quality index ($Q = D - U$)
B	The Brier score (average squared difference between predicted outcome probability and actual outcome)
C	Hosmer-Lemeshow calibration statistic using fixed probability deciles
H	Hosmer-Lemeshow calibration statistic using deciles of risk
AUC or `C`	Area under the ROC curve
Z	Wald Z score, $Z = \hat{\beta}/S\hat{E}(\hat{\beta})$

Appendix B

Summary of Final Dataset

Each variable included in my final dataset, along with the number of instances (n), the number of missing instances (missing), the mean, the median, and the standard deviation (std dev) are listed in this appendix. Chapter 2 describes the preparation of this dataset and the meaning of the variable naming notations.

Variable	num	missing	mean	median	std.dev
Index	1044982	0	2.86e+03	2.04e+03	2.59e+03
AdmitWt	1044982	51538	8.18e+01	7.94e+01	2.23e+01
Age	1044982	6441	6.52e+01	6.70e+01	1.55e+01
Aggrastat	1044982	0	2.38e-02	0.00e+00	4.82e-01
Aggrastat_perKg	1044982	0	2.74e-04	0.00e+00	5.56e-03
AIDS	1044982	0	1.07e-02	0.00e+00	1.03e-01
Albumin	1044982	863268	2.81e+00	2.80e+00	6.32e-01
Albumin_Slope_1680	1044982	784195	-2.42e-06	0.00e+00	1.04e-04
Allinput	1044982	26925	8.82e+03	7.04e+03	7.68e+03
Alloutput	1044982	26925	5.50e+03	3.90e+03	5.49e+03
ALT	1044982	828439	8.07e+01	3.60e+01	1.30e+02
ALT_Slope_1680	1044982	755633	-1.79e-04	0.00e+00	3.40e-02
Amicar	1044982	0	4.61e-04	0.00e+00	3.15e-02
Amicar_perKg	1044982	0	5.86e-06	0.00e+00	3.77e-04
Aminophylline	1044982	0	3.32e-03	0.00e+00	2.45e-01
Aminophylline_perKg	1044982	0	5.45e-05	0.00e+00	4.02e-03
Amiodarone	1044982	26	4.47e-02	0.00e+00	1.80e-01
Amiodarone_perKg	1044982	26	5.37e-04	0.00e+00	2.50e-03
Amrinone	1044982	0	5.78e-03	0.00e+00	1.30e+00
Amrinone_perKg	1044982	0	9.52e-05	0.00e+00	2.20e-02
Antiarrhythmic_agent	1044982	0	7.19e-02	0.00e+00	2.58e-01
Anticoagulant	1044982	0	1.20e-01	0.00e+00	3.25e-01
Antiplatelet_agent	1044982	0	2.13e-02	0.00e+00	1.44e-01
Argatroban	1044982	0	2.47e-01	0.00e+00	6.61e+00
Argatroban_perKg	1044982	0	3.47e-03	0.00e+00	1.01e-01
Art_BE	1044982	333506	-1.93e-01	0.00e+00	4.31e+00
Art_BE_Slope_1680	1044982	342098	6.34e-05	0.00e+00	2.97e-03
Art_CO2	1044982	296679	2.55e+01	2.50e+01	5.09e+00
Art_CO2_Slope_1680	1044982	236662	1.04e-04	0.00e+00	2.08e-03
Art_PaCO2	1044982	296899	4.06e+01	4.00e+01	8.56e+00
Art_PaCO2_Slope_1680	1044982	236870	-1.66e-04	0.00e+00	4.63e-03
Art_PaO2	1044982	297402	1.30e+02	1.10e+02	7.05e+01
Art_PaO2_Slope_1680	1044982	237530	-2.90e-02	-5.00e-03	6.46e-02
Art_pH	1044982	279769	7.39e+00	7.40e+00	6.57e-02
Art_pH_Slope_1680	1044982	229813	3.68e-09	0.00e+00	1.92e-06
Art_pH.basedev	1044982	279769	1.18e-02	7.53e-03	4.59e-02
Art_pH.range	1044982	196547	1.35e-01	1.20e-01	9.80e-02

Variable	num	missing	mean	median	std.dev
AST	1044982	819367	1.05e+02	5.10e+01	1.47e+02
AST_Slope_1680	1044982	742903	-2.50e-03	0.00e+00	4.74e-02
Ativan	1044982	102	1.26e-01	0.00e+00	9.64e-01
Ativan_perKg	1044982	102	1.56e-03	0.00e+00	1.24e-02
Atracurium	1044982	0	2.61e-03	0.00e+00	2.35e-01
Atracurium_perKg	1044982	0	3.44e-05	0.00e+00	3.08e-03
Barbiturate	1044982	0	3.89e-04	0.00e+00	1.97e-02
Benzodiazepine	1044982	0	1.17e-01	0.00e+00	3.22e-01
Beta.Blocking_agent	1044982	0	1.45e-02	0.00e+00	1.20e-01
Bivalirudin	1044982	0	1.14e-04	0.00e+00	1.07e-02
Bivalirudin_perKg	1044982	0	1.25e-05	0.00e+00	1.24e-03
Bpcor	1044982	536208	1.00e+00	1.00e+00	0.00e+00
BUN	1044982	129322	2.71e+01	2.00e+01	2.08e+01
BUN_Slope_1680	1044982	134662	3.41e-04	0.00e+00	3.97e-03
BUN.basedev	1044982	129322	1.03e+00	0.00e+00	6.82e+00
BUN.range	1044982	124717	7.59e+00	4.00e+00	1.07e+01
BUNtoCr	1044982	129852	2.32e+01	2.07e+01	1.20e+01
Ca	1044982	407850	8.19e+00	8.20e+00	7.39e-01
Ca_Slope_1680	1044982	362136	1.08e-05	0.00e+00	3.51e-04
Calcium_channel_blocking_agent	1044982	0	1.01e-02	0.00e+00	9.98e-02
Calprevflg	1044982	27087	9.99e-01	1.00e+00	2.40e-02
Calprevflg_Slope_1680	1044982	36893	0.00e+00	0.00e+00	0.00e+00
CCU	1044982	0	2.00e-01	0.00e+00	4.00e-01
Cisatracurium	1044982	0	1.00e-01	0.00e+00	2.48e+00
Cisatracurium_perKg	1044982	0	1.08e-03	0.00e+00	2.67e-02
Cl	1044982	146557	1.06e+02	1.06e+02	5.68e+00
Cl_Slope_1680	1044982	148449	-1.68e-04	0.00e+00	2.22e-03
CO2	1044982	145831	2.41e+01	2.40e+01	4.62e+00
CO2_Slope_1680	1044982	148093	2.78e-04	0.00e+00	1.72e-03
COfick	1044982	910451	5.92e+00	5.50e+00	2.12e+00
COfick_Slope_1680	1044982	838925	-2.93e-05	0.00e+00	1.36e-03
COfick_Slope_240	1044982	896519	-1.04e-04	0.00e+00	4.82e-03
ComfortMeas	1044982	0	0.00e+00	0.00e+00	0.00e+00
COtd	1044982	798107	5.29e+00	5.10e+00	1.59e+00
COtd_Slope_1680	1044982	725412	1.32e-04	0.00e+00	8.82e-04
COtd_Slope_240	1044982	786908	2.43e-04	0.00e+00	4.18e-03
COtdM	1044982	0	2.36e-01	0.00e+00	4.25e-01
CrdIndx	1044982	744396	2.77e+00	2.65e+00	7.35e-01
CrdIndx_Slope_1680	1044982	675140	6.75e-05	0.00e+00	5.52e-04
CrdIndx_Slope_240	1044982	733203	1.36e-04	0.00e+00	2.66e-03
CrdIndxM	1044982	0	2.88e-01	0.00e+00	4.53e-01
Creatinine	1044982	128631	1.29e+00	9.00e-01	1.19e+00
Creatinine_Slope_1680	1044982	134053	7.44e-06	0.00e+00	2.05e-04
Creatinine.basedev	1044982	128631	1.60e-02	0.00e+00	3.09e-01
Creatinine.range	1044982	124086	3.22e-01	2.00e-01	5.32e-01
CSRU	1044982	0	4.86e-01	0.00e+00	5.00e-01
CumPressorTime	1044982	0	7.80e+02	2.50e+01	1.44e+03
HR	1044982	28131	8.66e+01	8.50e+01	1.74e+01
HR.oor120.c	1044982	0	8.39e+00	0.00e+00	2.70e+01
HR.oor120.t	1044982	0	8.63e+00	0.00e+00	2.75e+01
HR.oor30.c	1044982	0	1.67e+00	0.00e+00	6.40e+00
HR.oor30.t	1044982	0	1.68e+00	0.00e+00	6.44e+00
HR_Slope_1680	1044982	34948	-4.48e-04	0.00e+00	9.92e-03
HR_Slope_240	1044982	36021	6.14e-05	0.00e+00	5.39e-02
HR.basedev	1044982	28131	-8.08e-01	-6.96e-01	1.19e+01
HR.range	1044982	25108	4.01e+01	3.60e+01	2.61e+01
CVP	1044982	574885	1.12e+01	1.10e+01	5.82e+00
CVP_Min_1440	1044982	474652	5.74e+00	5.00e+00	4.71e+00
CVP_Slope_1680	1044982	485061	-6.93e-05	0.00e+00	6.09e-03
CVP_Slope_240	1044982	568430	6.13e-04	0.00e+00	2.21e-02
CvpM	1044982	0	4.50e-01	0.00e+00	4.97e-01
DBili	1044982	1021961	3.82e+00	1.50e+00	5.89e+00
DBili_Slope_1680	1044982	1011019	1.39e-05	0.00e+00	8.86e-04

Variable	num	missing	mean	median	std.dev
DBP	1044982	354854	5.91e+01	5.80e+01	1.23e+01
DBP_Slope_1680	1044982	284266	-2.78e-04	0.00e+00	9.23e-03
DBP_Slope_240	1044982	342191	-4.72e-04	0.00e+00	4.82e-02
DBPm	1044982	31914	5.82e+01	5.70e+01	1.32e+01
DBPm.basedev	1044982	31914	-2.89e-01	-6.25e-01	9.82e+00
DBPm.range	1044982	25495	4.17e+01	4.00e+01	2.03e+01
Dilaudid	1044982	0	5.38e-03	0.00e+00	1.67e-01
Dilaudid_perKg	1044982	0	5.92e-05	0.00e+00	2.04e-03
Diltiazem	1044982	0	9.41e-02	0.00e+00	1.12e+00
Diltiazem_perKg	1044982	0	1.13e-03	0.00e+00	1.37e-02
Diuretic	1044982	0	1.26e-02	0.00e+00	1.12e-01
DNI	1044982	0	4.78e-06	0.00e+00	2.19e-03
DNR	1044982	0	3.35e-05	0.00e+00	5.79e-03
Dobutamine	1044982	0	7.51e+00	0.00e+00	6.81e+01
Dobutamine_perKg	1044982	0	8.93e-02	0.00e+00	7.76e-01
Dopamine	1044982	39	2.00e+01	0.00e+00	1.29e+02
Dopamine_perKg	1044982	39	2.45e-01	0.00e+00	1.54e+00
DopLg	1044982	39	5.46e-03	0.00e+00	7.37e-02
DopMd	1044982	39	2.77e-02	0.00e+00	1.64e-01
DopSm	1044982	39	6.41e-03	0.00e+00	7.98e-02
Doxacurium	1044982	0	1.41e-03	0.00e+00	5.02e-02
Doxacurium_perKg	1044982	0	1.83e-05	0.00e+00	6.58e-04
ECO	1044982	36119	3.33e+01	3.21e+01	1.08e+01
ECOSlope	1044982	39154	-5.90e-04	4.63e-04	1.04e-01
EctFreq	1044982	0	4.24e-01	0.00e+00	8.35e-01
Epinephrine	1044982	135	8.91e-02	0.00e+00	9.14e-01
Epinephrine_perKg	1044982	135	1.00e-03	0.00e+00	9.35e-03
Esmolol	1044982	0	4.62e+01	0.00e+00	7.85e+02
Esmolol_perKg	1044982	0	5.57e-01	0.00e+00	9.18e+00
FallRisk	1044982	0	5.19e-01	1.00e+00	5.00e-01
Fentanyl	1044982	0	1.08e+01	0.00e+00	4.59e+01
Fentanyl_Conc	1044982	2	5.77e-01	0.00e+00	1.01e+01
Fentanyl_Conc_perKg	1044982	2	7.30e-03	0.00e+00	1.36e-01
Fentanyl_perKg	1044982	0	1.31e-01	0.00e+00	5.75e-01
FiO2Set	1044982	393806	5.32e-01	5.00e-01	1.84e-01
FiO2Set_Slope_1680	1044982	328620	-1.10e-05	0.00e+00	1.47e-04
FlushSkin	1044982	0	8.40e-03	0.00e+00	9.13e-02
FullCode	1044982	0	8.86e-01	1.00e+00	3.18e-01
GCS	1044982	51831	1.13e+01	1.40e+01	4.23e+00
GCS_Slope_1680	1044982	61291	8.07e-04	0.00e+00	2.76e-03
GCS.basedev	1044982	51831	1.05e+00	1.67e-01	2.78e+00
GCS.range	1044982	51508	5.51e+00	5.00e+00	4.70e+00
General_anesthetic	1044982	0	1.97e-01	0.00e+00	3.98e-01
Glucose	1044982	70035	1.36e+02	1.26e+02	4.84e+01
Glucose_Slope_1680	1044982	75671	-5.63e-03	0.00e+00	3.75e-02
HCT	1044982	69067	3.10e+01	3.05e+01	4.66e+00
HCT_Slope_1680	1044982	75778	-1.17e-04	0.00e+00	2.69e-03
HCTM	1044982	0	9.34e-01	1.00e+00	2.48e-01
HCT.basedev	1044982	69067	-3.34e-01	-1.07e-14	2.70e+00
HCT.range	1044982	65885	6.80e+00	6.00e+00	5.37e+00
HemMalig	1044982	0	3.04e-02	0.00e+00	1.72e-01
Hemostatic_agent	1044982	0	2.87e-02	0.00e+00	1.67e-01
Heparin	1044982	3	1.19e+02	0.00e+00	3.54e+02
Heparin_perKg	1044982	3	1.39e+00	0.00e+00	4.06e+00
Hgb	1044982	150841	1.05e+01	1.04e+01	1.63e+00
Hgb_Slope_1680	1044982	153327	-4.30e-05	0.00e+00	8.33e-04
Hgb.basedev	1044982	150841	-1.26e-01	-1.78e-15	8.90e-01
Hgb.range	1044982	143443	1.64e+00	1.30e+00	1.58e+00
HospTime	1044982	0	2.72e+03	1.15e+03	5.13e+03
HRCritEvnts.24h	1044982	0	6.67e+00	0.00e+00	1.65e+01
HRCritEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
HREvnts.24h	1044982	0	7.97e+00	0.00e+00	1.74e+01
HREvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00

Variable	num	missing	mean	median	std.dev
HrmHB	1044982	0	2.94e-02	0.00e+00	2.40e-01
HrmPaced	1044982	0	1.26e-01	0.00e+00	3.31e-01
HrmSA	1044982	0	6.87e-01	0.00e+00	1.91e+00
HrmVA	1044982	0	1.49e-02	0.00e+00	1.77e-01
HRThreshCnt	1044982	0	6.89e+00	0.00e+00	1.60e+01
HRThreshCntF	1044982	25108	8.62e-02	0.00e+00	1.81e-01
HRThreshCntN	1044982	0	2.86e+00	0.00e+00	6.80e+00
IABP	1044982	0	4.98e-02	0.00e+00	2.18e-01
ImpairedSkin	1044982	0	2.25e-01	0.00e+00	4.18e-01
Inotropic_agent	1044982	0	4.92e-02	0.00e+00	2.16e-01
Input_60	1044982	26925	2.70e+02	8.99e+01	7.53e+02
Input_60.basedev	1044982	26925	-2.52e+02	-1.39e+02	6.48e+02
Input_60.range	1044982	24840	2.51e+03	1.42e+03	2.70e+03
InputB	1044982	429799	1.53e+02	6.70e+01	4.13e+02
InputB_60	1044982	429799	2.36e+02	8.25e+01	6.75e+02
InputOtherBlood	1044982	0	1.74e+01	0.00e+00	1.00e+02
InputOtherBloodB	1044982	0	1.60e+00	0.00e+00	2.73e+01
InputRBCs	1044982	0	4.72e+01	0.00e+00	1.78e+02
InputRBCsB	1044982	0	3.68e+00	0.00e+00	4.75e+01
INR	1044982	0	1.35e+00	1.20e+00	5.49e-01
INR_Slope_1680	1044982	234157	-2.60e-05	0.00e+00	3.18e-04
INR.basedev	1044982	0	-9.12e-03	0.00e+00	3.61e-01
INR.range	1044982	0	5.51e-01	3.00e-01	8.96e-01
Insulin	1044982	0	1.73e-01	0.00e+00	3.78e-01
Insulin_perKg	1044982	0	8.62e-03	0.00e+00	4.02e-02
Integrelin	1044982	27	2.73e+00	0.00e+00	2.14e+01
Integrelin_perKg	1044982	27	3.28e-02	0.00e+00	2.48e-01
IonCa	1044982	474733	1.13e+00	1.14e+00	1.11e-01
IonCa_Slope_1680	1044982	416770	-1.03e-07	0.00e+00	2.70e-05
JaundiceSkin	1044982	0	8.34e-03	0.00e+00	9.10e-02
K	1044982	61530	4.11e+00	4.10e+00	5.59e-01
K_Slope_1680	1044982	68977	-3.01e-05	0.00e+00	4.99e-04
Ketamine	1044982	0	8.23e-05	0.00e+00	9.07e-03
Ketamine_perKg	1044982	0	9.14e-04	0.00e+00	1.23e-01
Labetolol	1044982	0	1.53e-02	0.00e+00	2.05e-01
Labetolol_perKg	1044982	0	1.82e-04	0.00e+00	2.45e-03
Lactate	1044982	756439	2.45e+00	1.80e+00	2.24e+00
Lactate_Slope_1680	1044982	666906	-1.24e-04	0.00e+00	9.67e-04
LactateM	1044982	0	2.76e-01	0.00e+00	4.47e-01
Lactate.basedev	1044982	756439	-2.55e-01	-4.44e-16	1.12e+00
Lactate.range	1044982	570744	1.29e+00	5.00e-01	2.13e+00
Lasix	1044982	0	1.07e-01	0.00e+00	1.16e+00
Lasix_perKg	1044982	0	1.25e-03	0.00e+00	1.38e-02
Lepirudin	1044982	0	1.11e-03	0.00e+00	1.15e-01
Lepirudin_perKg	1044982	0	1.39e-05	0.00e+00	1.42e-03
Levophed	1044982	67	8.70e-01	0.00e+00	4.58e+00
Levophed_perKg	1044982	67	1.01e-02	0.00e+00	5.20e-02
Lidocaine	1044982	5	1.19e-02	0.00e+00	1.49e-01
Lidocaine_perKg	1044982	5	1.36e-04	0.00e+00	1.74e-03
LOSBal	1044982	235013	3.11e+03	2.10e+03	5.10e+03
LOSBal.basedev	1044982	235013	4.64e+02	0.00e+00	2.27e+03
LOSBal.range	1044982	232778	2.37e+03	1.00e+03	3.55e+03
MAP	1044982	358738	7.88e+01	7.70e+01	1.49e+01
MAP_Slope_1680	1044982	285006	-2.91e-04	0.00e+00	1.14e-02
MAP_Slope_240	1044982	343167	-9.13e-04	0.00e+00	6.38e-02
MBPm	1044982	35683	7.86e+01	7.67e+01	1.49e+01
MBPm.pr	1044982	0	9.82e-01	1.00e+00	8.58e-02
MBPm.basedev	1044982	35683	9.52e-02	-4.02e-01	1.16e+01
MBPm.range	1044982	25854	4.80e+01	4.60e+01	2.26e+01
MeanObsIntv	1044982	10066	3.34e+01	3.29e+01	1.52e+01
MechVent	1044982	0	4.85e-01	0.00e+00	5.00e-01
MetCarcinoma	1044982	0	1.86e-02	0.00e+00	1.35e-01
Mg	1044982	199888	2.07e+00	2.00e+00	3.86e-01

Variable	num	missing	mean	median	std.dev
Mg_Slope_1680	1044982	189486	2.03e-05	0.00e+00	2.53e-04
MICU	1044982	0	2.01e-01	0.00e+00	4.01e-01
Midazolam	1044982	0	3.14e-01	0.00e+00	1.55e+00
Midazolam_perKg	1044982	0	3.81e-03	0.00e+00	1.94e-02
Milrinone	1044982	14	1.43e+00	0.00e+00	7.17e+00
Milrinone_perKg	1044982	14	1.64e-02	0.00e+00	7.95e-02
Morphine_Sulfate	1044982	0	4.98e-02	0.00e+00	5.35e-01
Morphine_Sulfate_perKg	1044982	0	6.30e-04	0.00e+00	7.21e-03
MSICU	1044982	0	8.38e-02	0.00e+00	2.77e-01
Na	1044982	86686	1.38e+02	1.38e+02	4.44e+00
Na_Slope_1680	1044982	91126	1.56e-04	0.00e+00	1.99e-03
Narcan	1044982	0	2.10e-04	0.00e+00	1.45e-02
Narcan_perKg	1044982	0	1.74e-05	0.00e+00	2.34e-03
Natrecor	1044982	11	1.31e-02	0.00e+00	1.41e-01
Natrecor_perKg	1044982	11	1.49e-04	0.00e+00	1.58e-03
NBPDias	1044982	593020	5.69e+01	5.50e+01	1.49e+01
NBPDias_Slope_1680	1044982	439728	-6.51e-04	0.00e+00	1.30e-02
NBPDias_Slope_240	1044982	567504	-1.52e-03	0.00e+00	6.12e-02
NBPMean	1044982	594471	7.66e+01	7.50e+01	1.51e+01
NBPMean_Slope_1680	1044982	440194	-6.71e-04	0.00e+00	1.37e-02
NBPMean_Slope_240	1044982	568402	-1.78e-03	0.00e+00	5.93e-02
NBPSys	1044982	592489	1.17e+02	1.15e+02	2.22e+01
NBPSys_Slope_1680	1044982	439347	-7.29e-04	0.00e+00	1.92e-02
NBPSys_Slope_240	1044982	567327	-2.34e-03	0.00e+00	8.03e-02
Neosynephrine	1044982	72	1.24e+01	0.00e+00	4.63e+01
Neosynephrine_perKg	1044982	72	1.43e-01	0.00e+00	5.21e-01
Nicardipine	1044982	0	1.76e-01	0.00e+00	5.64e+00
Nicardipine_perKg	1044982	0	2.00e-03	0.00e+00	6.22e-02
Nitroglycerine	1044982	75	1.00e+01	0.00e+00	4.42e+01
Nitroglycerine_perKg	1044982	75	1.22e-01	0.00e+00	5.42e-01
Nitroprusside	1044982	9	2.72e+00	0.00e+00	2.31e+01
Nitroprusside_perKg	1044982	9	3.39e-02	0.00e+00	2.79e-01
NoCPR	1044982	0	1.91e-06	0.00e+00	1.38e-03
Nondepolarizing_agent	1044982	0	1.01e-02	0.00e+00	9.97e-02
ObsFreq2hr	1044982	0	2.07e+00	1.50e+00	1.37e+00
Opiate	1044982	0	1.27e-01	0.00e+00	3.33e-01
Orientation	1044982	0	6.41e-01	0.00e+00	8.10e-01
OrientUnableAs	1044982	0	2.15e-01	0.00e+00	4.11e-01
OtherCode	1044982	0	0.00e+00	0.00e+00	0.00e+00
Output_60	1044982	26925	1.60e+02	8.00e+01	2.98e+02
OutputB_60	1044982	429799	1.49e+02	8.00e+01	2.74e+02
PAC	1044982	0	7.86e-02	0.00e+00	2.69e-01
Pacemkr	1044982	0	3.52e-01	0.00e+00	4.78e-01
PaleSkin	1044982	0	1.89e-02	0.00e+00	1.36e-01
Pancuronium	1044982	0	2.77e-04	0.00e+00	3.84e-02
Pancuronium_perKg	1044982	0	3.71e-06	0.00e+00	5.21e-04
PaO2toFiO2	1044982	67328	4.64e-01	0.00e+00	7.90e-01
PAPmean	1044982	940747	3.03e+01	2.90e+01	9.22e+00
PAPmean_Slope_1680	1044982	916138	-1.59e-04	0.00e+00	6.70e-03
PAPmean_Slope_240	1044982	936868	1.51e-04	0.00e+00	3.03e-02
PAPmeanM	1044982	0	9.97e-02	0.00e+00	3.00e-01
PAPsd	1044982	734789	3.81e+01	3.60e+01	1.20e+01
PAPsd_Slope_1680	1044982	659671	3.26e-04	0.00e+00	7.51e-03
PAPsd_Slope_240	1044982	723053	1.30e-03	0.00e+00	3.27e-02
PCWP	1044982	998797	1.80e+01	1.70e+01	6.66e+00
PCWP_Slope_1680	1044982	967522	-7.26e-05	0.00e+00	3.45e-03
PCWP_Slope_240	1044982	992876	-9.90e-05	0.00e+00	1.28e-02
PCWPM	1044982	0	4.42e-02	0.00e+00	2.06e-01
PEEPSet	1044982	381198	5.74e+00	5.00e+00	3.44e+00
PEEPSet_Slope_1680	1044982	350684	2.26e-05	0.00e+00	1.56e-03
Pentobarbitol	1044982	0	1.10e-01	0.00e+00	5.83e+00
Pentobarbitol_perKg	1044982	0	9.96e-04	0.00e+00	5.21e-02
PIP	1044982	473547	2.72e+01	2.70e+01	6.98e+00

Variable	num	missing	mean	median	std.dev
PIP_Slope_1680	1044982	377420	-3.24e-04	0.00e+00	3.24e-03
PlateauPres	1044982	510143	2.17e+01	2.10e+01	5.59e+00
PlateauPres_Slope_1680	1044982	410636	-2.45e-05	0.00e+00	2.37e-03
Platelets	1044982	106401	1.88e+02	1.70e+02	1.01e+02
Platelets_Slope_1680	1044982	109682	-1.73e-03	0.00e+00	2.82e-02
PNC	1044982	0	7.77e-04	0.00e+00	2.79e-02
Precedex	1044982	3	3.11e-01	0.00e+00	4.17e+00
Precedex_perKg	1044982	2	3.33e-03	0.00e+00	4.36e-02
PressD01	1044982	0	4.73e-01	0.00e+00	4.99e-01
PressD12	1044982	0	3.29e-02	0.00e+00	1.78e-01
PressD24	1044982	0	1.75e-02	0.00e+00	1.31e-01
PressD4	1044982	0	4.60e-03	0.00e+00	6.76e-02
PressorCnt	1044982	342	3.57e-01	0.00e+00	6.81e-01
PressorM	1044982	342	2.64e-01	0.00e+00	4.41e-01
PressorSum.std	1044982	342	3.56e-01	0.00e+00	9.98e-01
PressorTime	1044982	0	3.38e+02	0.00e+00	1.02e+03
Procainamide	1044982	6	1.96e-03	0.00e+00	7.30e-02
Procainamide_perKg	1044982	6	2.45e-05	0.00e+00	9.08e-04
Propofol	1044982	58	5.87e+02	0.00e+00	1.46e+03
Propofol_perKg	1044982	58	6.66e+00	0.00e+00	1.59e+01
PT	1044982	325961	1.48e+01	1.41e+01	2.72e+00
PT_Slope_1680	1044982	235875	-2.08e-04	0.00e+00	1.15e-03
PT.basedev	1044982	325961	-3.31e-01	-1.07e-14	1.38e+00
PT.range	1044982	175440	1.87e+00	9.00e-01	2.68e+00
PTT	1044982	319543	4.02e+01	3.30e+01	2.16e+01
PTT_Slope_1680	1044982	231764	-1.13e-03	0.00e+00	1.37e-02
PulsePres	1044982	33686	6.03e+01	5.80e+01	1.84e+01
PVC	1044982	0	1.85e-01	0.00e+00	3.88e-01
PVR	1044982	1028665	2.33e+02	2.00e+02	1.51e+02
PVR_Slope_1680	1044982	1018344	-1.97e-03	0.00e+00	6.82e-02
PVR_Slope_240	1044982	1026750	-3.61e-03	0.00e+00	2.83e-01
RBC	1044982	161758	3.49e+00	3.44e+00	5.66e-01
RBC_Slope_1680	1044982	163450	3.13e-06	0.00e+00	2.65e-04
Reopro	1044982	0	7.17e-02	0.00e+00	7.50e+00
Reopro_perKg	1044982	0	8.82e-04	0.00e+00	9.21e-02
RESP	1044982	47616	1.92e+01	1.90e+01	6.04e+00
RESP_Slope_1680	1044982	40344	5.48e-04	0.00e+00	3.76e-03
RESP_Slope_240	1044982	50458	1.21e-03	0.00e+00	2.57e-02
Respiratory_smooth_muscle_relaxant	1044982	0	2.09e-04	0.00e+00	1.44e-02
RespSet	1044982	474001	1.41e+01	1.20e+01	4.94e+00
RespSet_Slope_1680	1044982	377164	8.59e-05	0.00e+00	1.87e-03
RespSpon	1044982	717716	6.87e+00	5.00e+00	6.12e+00
RespSpon_Slope_1680	1044982	635585	1.70e-04	0.00e+00	3.21e-03
RespTot	1044982	398799	1.83e+01	1.80e+01	6.31e+00
RespTot_Slope_1680	1044982	327635	1.04e-03	0.00e+00	3.83e-03
RikerSAS	1044982	0	3.73e+00	4.00e+00	7.44e-01
Sandostatin	1044982	0	6.67e-01	0.00e+00	5.80e+00
Sandostatin_perKg	1044982	0	7.77e-03	0.00e+00	6.94e-02
SaO2	1044982	595009	9.69e+01	9.70e+01	2.10e+00
SaO2_Slope_1680	1044982	522394	-7.76e-05	0.00e+00	1.30e-03
SBP	1044982	355024	1.18e+02	1.15e+02	2.28e+01
SBP_Slope_1680	1044982	284252	2.03e-04	0.00e+00	1.59e-02
SBP_Slope_240	1044982	342184	-7.04e-04	0.00e+00	9.09e-02
SBPm	1044982	31710	1.19e+02	1.16e+02	2.22e+01
SBPm.oor120.c	1044982	0	7.28e+00	0.00e+00	2.14e+01
SBPm.oor120.t	1044982	0	7.76e+00	0.00e+00	2.24e+01
SBPm.oor30.c	1044982	0	1.79e+00	0.00e+00	6.33e+00
SBPm.oor30.t	1044982	0	1.82e+00	0.00e+00	6.41e+00
SBPm.pr	1044982	0	9.78e-01	1.00e+00	8.70e-02
SBPmCritEvnts.24h	1044982	0	5.69e+00	0.00e+00	1.22e+01
SBPmCritEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
SBPmEvnts.24h	1044982	0	9.69e+00	4.00e+00	1.51e+01
SBPmEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00

Variable	num	missing	mean	median	std.dev
SBPm.basedev	1044982	31710	1.49e+00	3.86e-01	1.67e+01
SBPm.range	1044982	25462	6.75e+01	6.50e+01	3.22e+01
SBPmThreshCnt	1044982	0	7.03e+00	3.00e+00	1.20e+01
SBPmThreshCntF	1044982	25462	8.77e-02	4.30e-02	1.31e-01
SBPmThreshCntN	1044982	0	3.13e+00	1.00e+00	4.86e+00
Sedatives	1044982	0	3.19e-01	0.00e+00	4.66e-01
Sex	1044982	4897	5.83e-01	1.00e+00	4.93e-01
ShockIdx	1044982	32009	7.57e-01	7.31e-01	2.17e-01
SICU	1044982	0	2.81e-02	0.00e+00	1.65e-01
Sid	1044982	0	1.32e+04	1.32e+04	7.67e+03
Somatostatin_preparation	1044982	0	1.36e-02	0.00e+00	1.16e-01
SpO2	1044982	42290	9.71e+01	9.80e+01	2.94e+00
SpO2.oor120.c	1044982	0	1.77e+00	0.00e+00	1.11e+01
SpO2.oor120.t	1044982	0	1.82e+00	0.00e+00	1.14e+01
SpO2.oor30.c	1044982	0	4.13e-01	0.00e+00	3.17e+00
SpO2.oor30.t	1044982	0	4.16e-01	0.00e+00	3.19e+00
SpO2_Slope_1680	1044982	37253	-2.71e-04	0.00e+00	1.97e-03
SpO2_Slope_240	1044982	43240	-5.25e-04	0.00e+00	1.32e-02
SpO2CritEvnts.24h	1044982	0	1.31e+00	0.00e+00	5.49e+00
SpO2CritEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
SpO2Evnts.24h	1044982	0	2.16e+00	0.00e+00	6.52e+00
SpO2Evnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
SpO2LowCnt	1044982	0	1.57e+00	0.00e+00	5.40e+00
SpO2LowCntF	1044982	27091	1.89e-02	0.00e+00	6.78e-02
SpO2LowCntN	1044982	0	7.40e-01	0.00e+00	1.92e+00
SvCCU	1044982	0	1.72e-01	0.00e+00	3.77e-01
SvCSICU	1044982	0	0.00e+00	0.00e+00	0.00e+00
SvCSRU	1044982	0	3.51e-01	0.00e+00	4.77e-01
SvMICU	1044982	0	2.21e-01	0.00e+00	4.15e-01
SvMSICU	1044982	0	9.40e-02	0.00e+00	2.92e-01
SvNSICU	1044982	0	0.00e+00	0.00e+00	0.00e+00
SvOther	1044982	0	0.00e+00	0.00e+00	0.00e+00
SVR	1044982	756061	1.01e+03	9.57e+02	3.61e+02
SVR_Slope_1680	1044982	690076	-5.08e-02	-1.50e-02	2.44e-01
SVR_Slope_240	1044982	745295	-1.13e-01	0.00e+00	1.31e+00
Sympathomimetic_agent	1044982	0	2.68e-01	0.00e+00	4.43e-01
TBili	1044982	830068	2.72e+00	8.00e-01	5.69e+00
TBili_Slope_1680	1044982	762419	2.38e-05	0.00e+00	6.05e-04
Temp	1044982	37576	9.86e+01	9.86e+01	1.47e+00
Temp_Slope_1680	1044982	47164	2.65e-04	0.00e+00	1.15e-03
Thrombolytic_agent	1044982	0	1.90e-04	0.00e+00	1.38e-02
TidVolObs	1044982	467579	6.03e+02	6.00e+02	1.31e+02
TidVolObs_Slope_1680	1044982	373740	-7.36e-03	0.00e+00	5.19e-02
TidVolSet	1044982	483249	6.06e+02	6.00e+02	1.21e+02
TidVolSet_Slope_1680	1044982	383901	-2.28e-03	0.00e+00	2.72e-02
TidVolSpon	1044982	709376	4.90e+02	4.78e+02	1.58e+02
TidVolSpon_Slope_1680	1044982	626579	3.27e-03	0.00e+00	7.84e-02
TotIn24	1044982	238077	3.33e+03	2.49e+03	2.88e+03
TotIV	1044982	247376	1.99e+03	1.53e+03	1.76e+03
TotOut24	1044982	240813	2.25e+03	2.01e+03	1.59e+03
TPA	1044982	0	4.01e-04	0.00e+00	1.10e-01
TPA_perKg	1044982	0	4.23e-06	0.00e+00	1.04e-03
TProtein	1044982	1034847	5.37e+00	5.40e+00	1.06e+00
TProtein_Slope_1680	1044982	1028940	-1.25e-06	0.00e+00	3.53e-05
Trach	1044982	0	7.70e-03	0.00e+00	8.74e-02
Troponin	1044982	1017024	8.71e+00	3.50e+00	1.16e+01
Troponin_Slope_1680	1044982	1005734	-6.31e-05	0.00e+00	3.47e-03
UrHiCntN	1044982	0	0.00e+00	0.00e+00	0.00e+00
UrineByHr	1044982	57454	1.10e+02	6.00e+01	1.66e+02
UrineByHr.oor120.c	1044982	0	1.85e+01	0.00e+00	3.90e+01
UrineByHr.oor120.t	1044982	0	1.86e+01	0.00e+00	3.92e+01
UrineByHr.oor60.c	1044982	0	9.55e+00	0.00e+00	2.17e+01
UrineByHr.oor60.t	1044982	0	9.55e+00	0.00e+00	2.17e+01

Variable	num	missing	mean	median	std.dev
UrineCritEvnts	1044982	0	7.01e-02	0.00e+00	2.55e-01
UrineCritEvnts.24h	1044982	0	5.77e+00	0.00e+00	1.40e+01
UrineCritEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
UrineEvnts.24h	1044982	0	8.36e+00	2.00e+00	1.62e+01
UrineEvnts.cum	1044982	0	0.00e+00	0.00e+00	0.00e+00
UrineOut	1044982	85448	1.38e+02	8.00e+01	1.73e+02
UrineOutB	1044982	573884	1.25e+02	8.00e+01	1.48e+02
Vasodilating_agent	1044982	0	1.20e-01	0.00e+00	3.25e-01
Vasopressin	1044982	46	2.28e-02	0.00e+00	2.40e-01
Vasopressin_perKg	1044982	46	2.59e-04	0.00e+00	2.73e-03
VasopressorCnt	1044982	328	2.79e-01	0.00e+00	5.50e-01
VasopressorSum.std	1044982	328	2.87e-01	0.00e+00	8.89e-01
Vecuronium	1044982	3	1.04e-02	0.00e+00	2.53e-01
Vecuronium_perKg	1044982	3	1.18e-04	0.00e+00	2.76e-03
Vent	1044982	0	4.97e-01	0.00e+00	5.00e-01
VentLen	1044982	0	8.45e+02	0.00e+00	1.69e+03
VentLenC	1044982	0	1.54e+03	5.84e+02	2.16e+03
VentMode	1044982	0	1.91e+00	0.00e+00	2.80e+00
WBC	1044982	158454	1.27e+01	1.16e+01	6.28e+00
WBC_Slope_1680	1044982	161553	-1.28e-04	0.00e+00	2.49e-03
Weight	1044982	441981	8.60e+01	8.31e+01	2.41e+01
Weight_Slope_1680	1044982	412468	1.94e-04	0.00e+00	3.14e-03
Weight.basedev	1044982	441981	5.26e-01	0.00e+00	4.50e+00
Weight.range	1044982	373323	2.38e+00	0.00e+00	6.91e+00
Bal24	1044982	235489	1.10e+03	5.02e+02	2.76e+03
UrOut24	1044982	245526	1.81e+03	1.56e+03	1.37e+03

Appendix C

DASn Model Selection

The `DAS`n model selection was done in two steps. The first step looked for models by performing model selection from scratch and entirely based on the specific day, n, for the model. This was difficult, especially for the later days that had more limited data as the number of candidate covariates increased in proportion to the number of observations. To complement this selection, the covariates from the final `SDAS` model were used in conjunction with the ones selected from day 1, 2, and 3. The figures in this appendix describe both stages of the `DAS`n model selection.

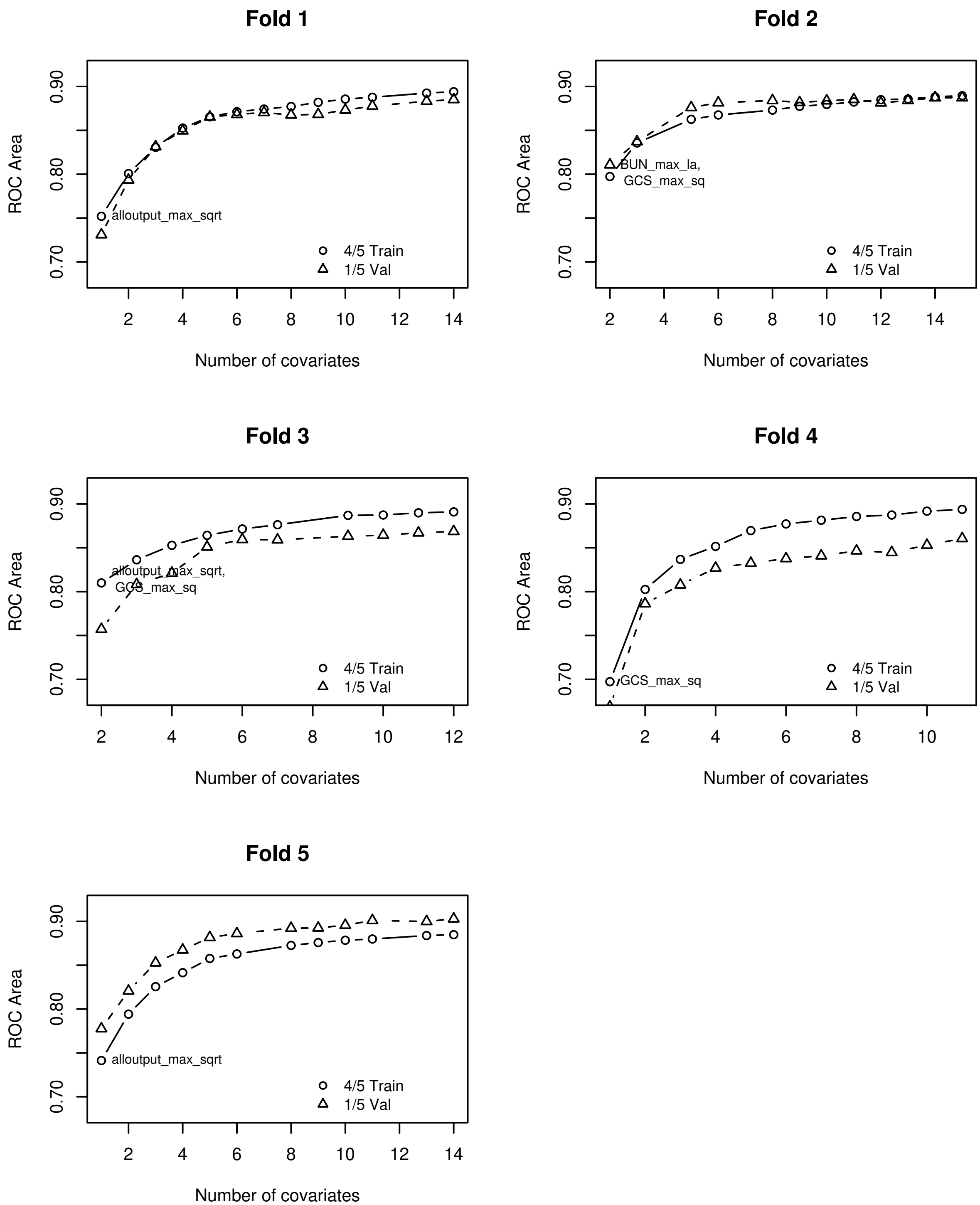


Figure C-1: `DAS1` Model Selection Stage 1: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 1)

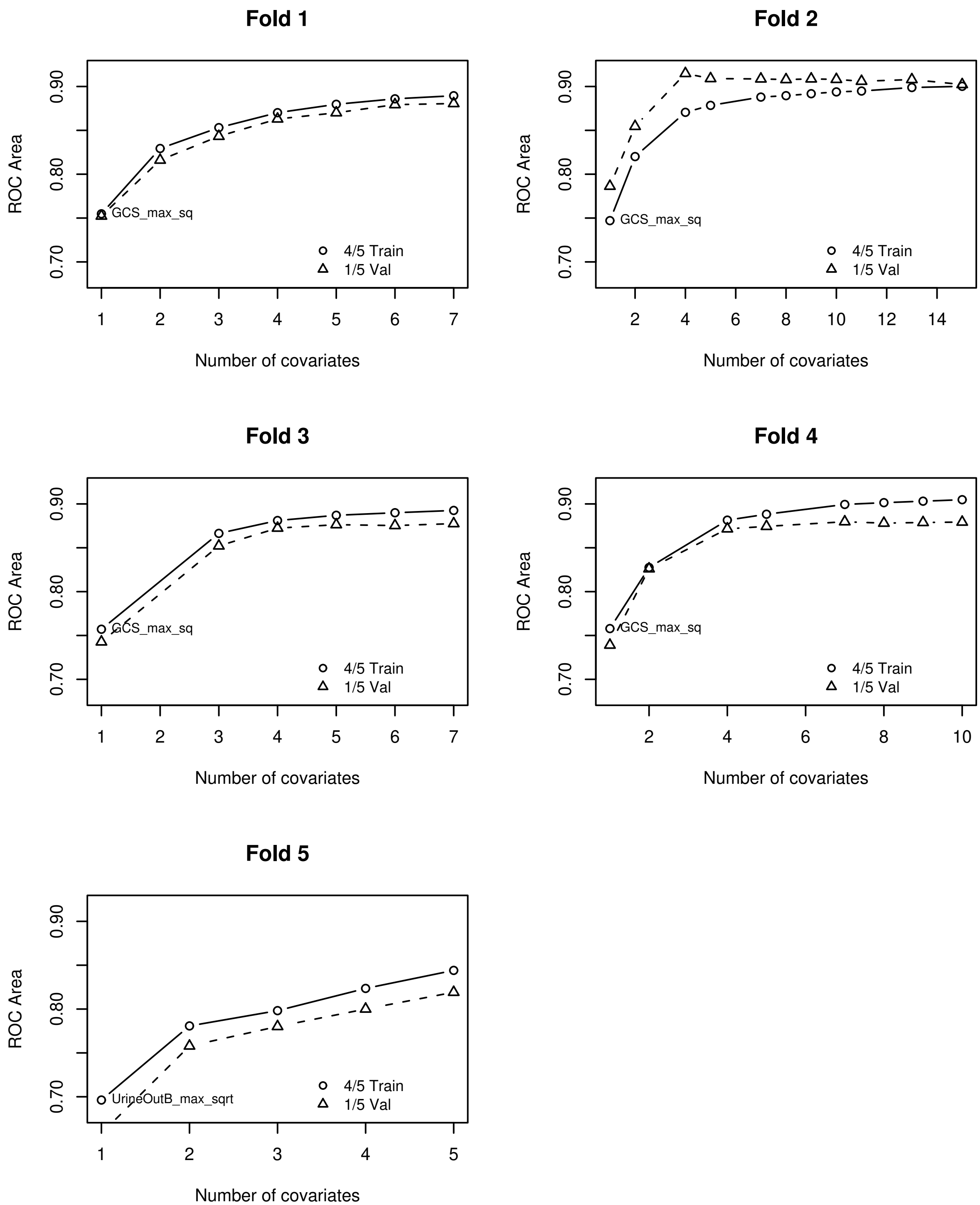


Figure C-2: DAS2 Model Selection Stage 1: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 2)

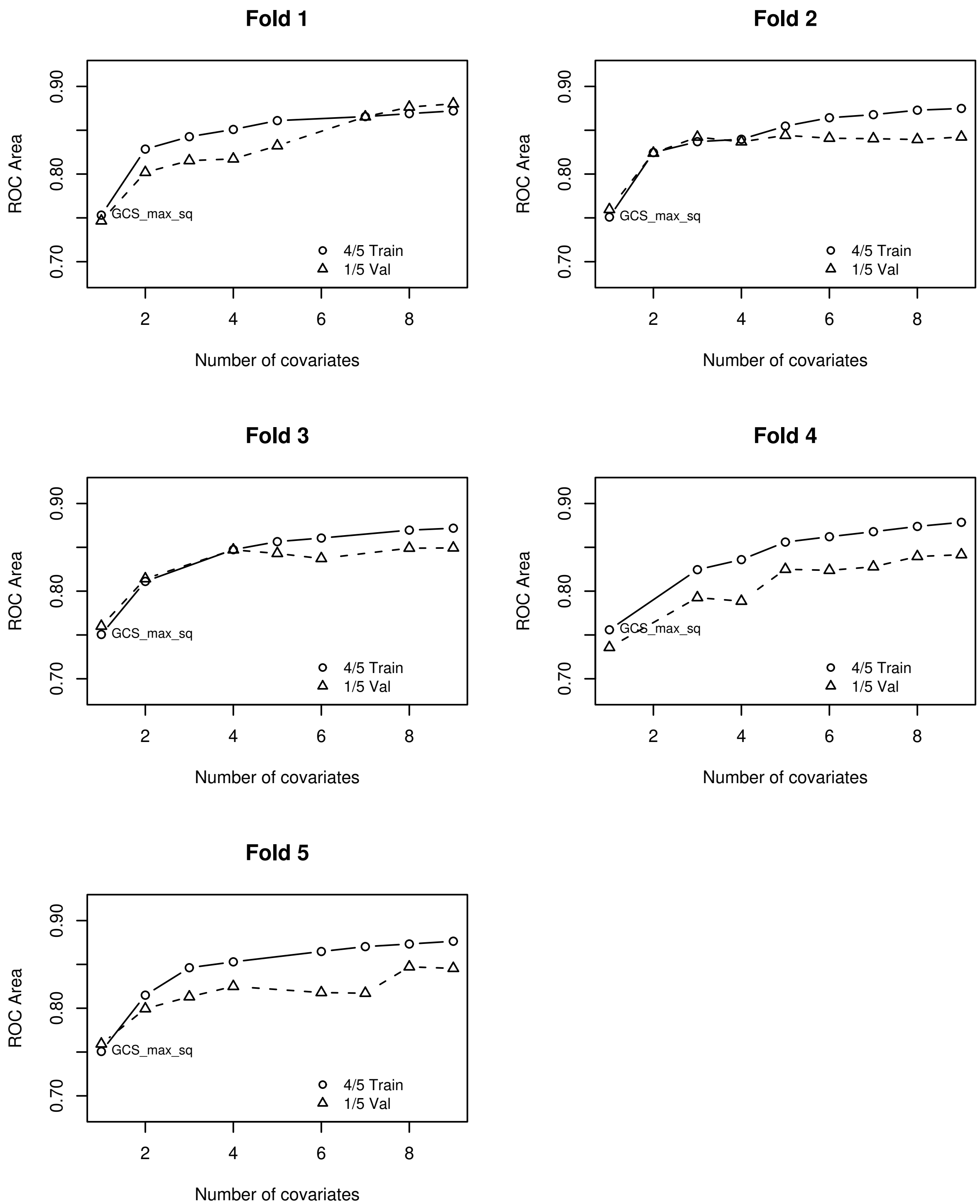


Figure C-3: `DAS3` Model Selection Stage 1: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 3)

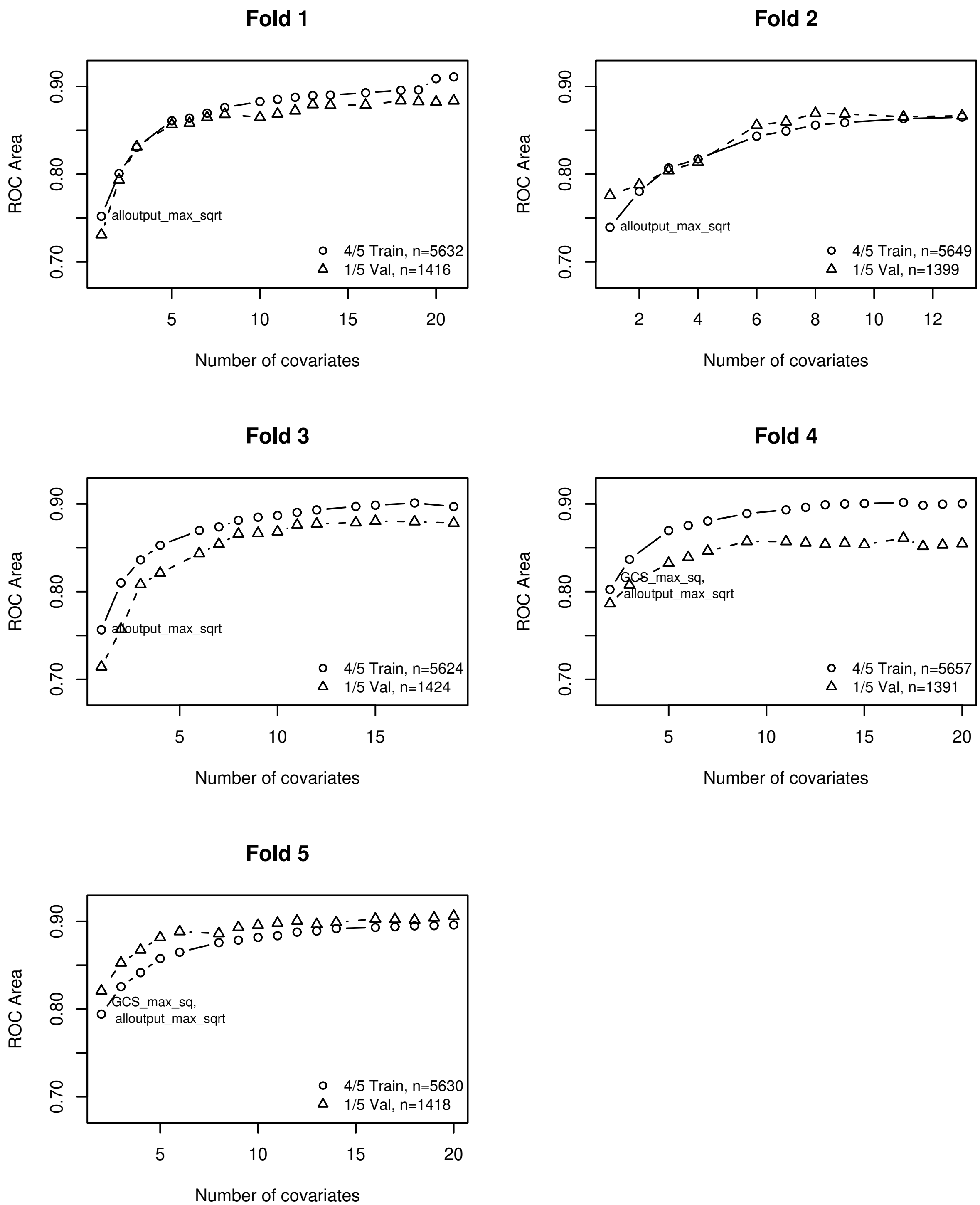


Figure C-4: `DAS1` Model Selection Stage 2: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 1)

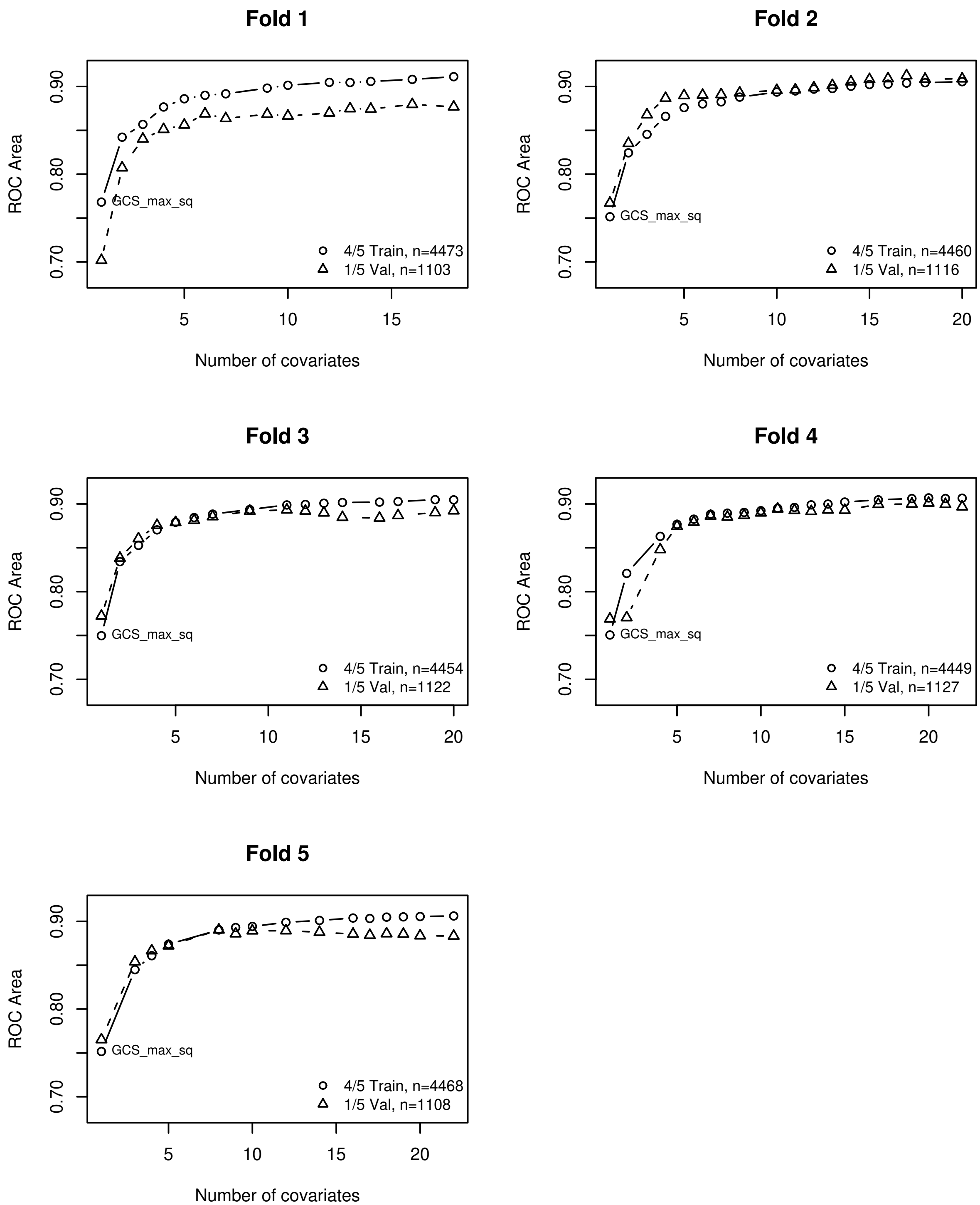


Figure C-5: `DAS2` Model Selection Stage 2: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 2)

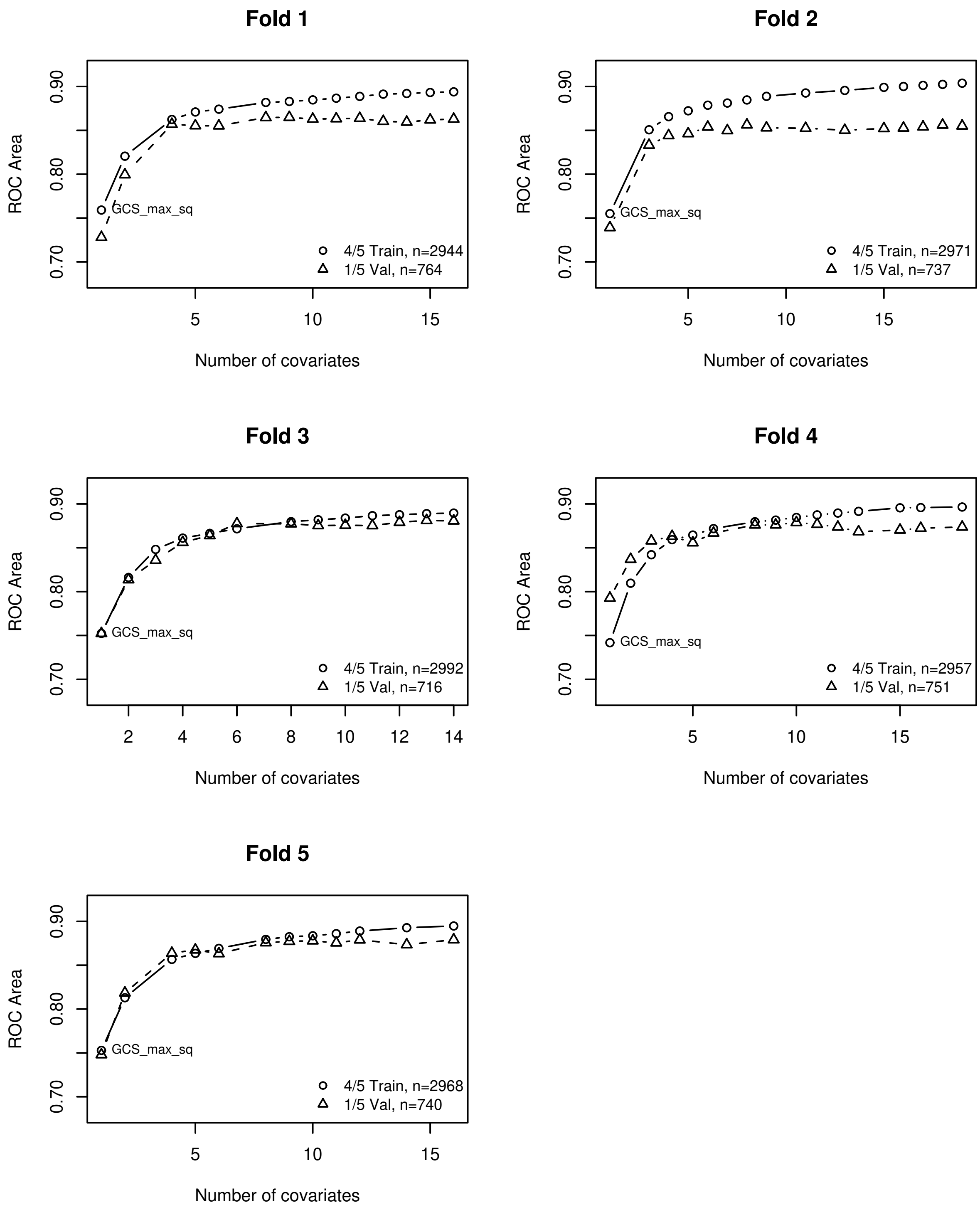


Figure C-6: `DAS3` Model Selection Stage 2: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 3)

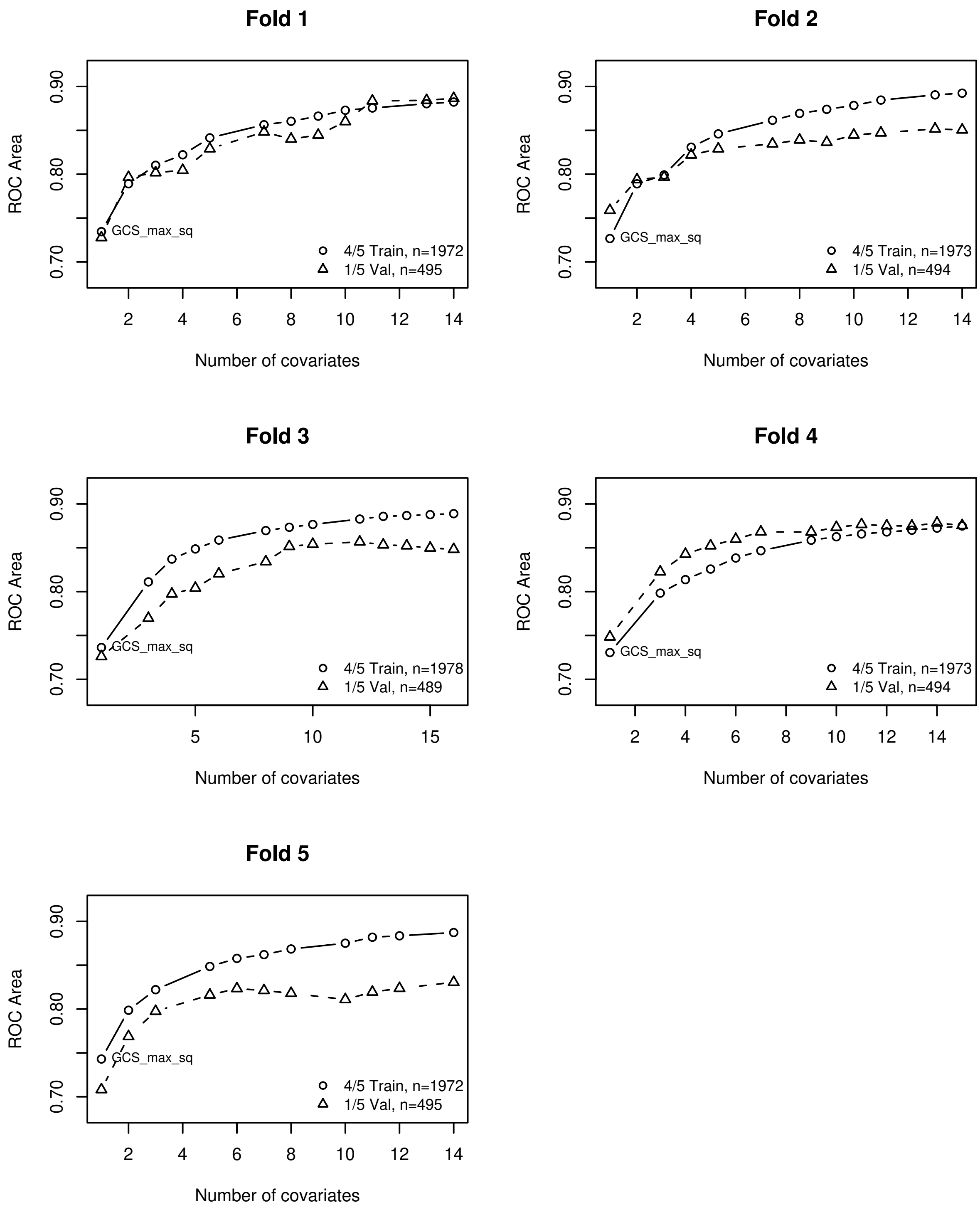


Figure C-7: `DAS4` Model Selection Stage 2: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 4)

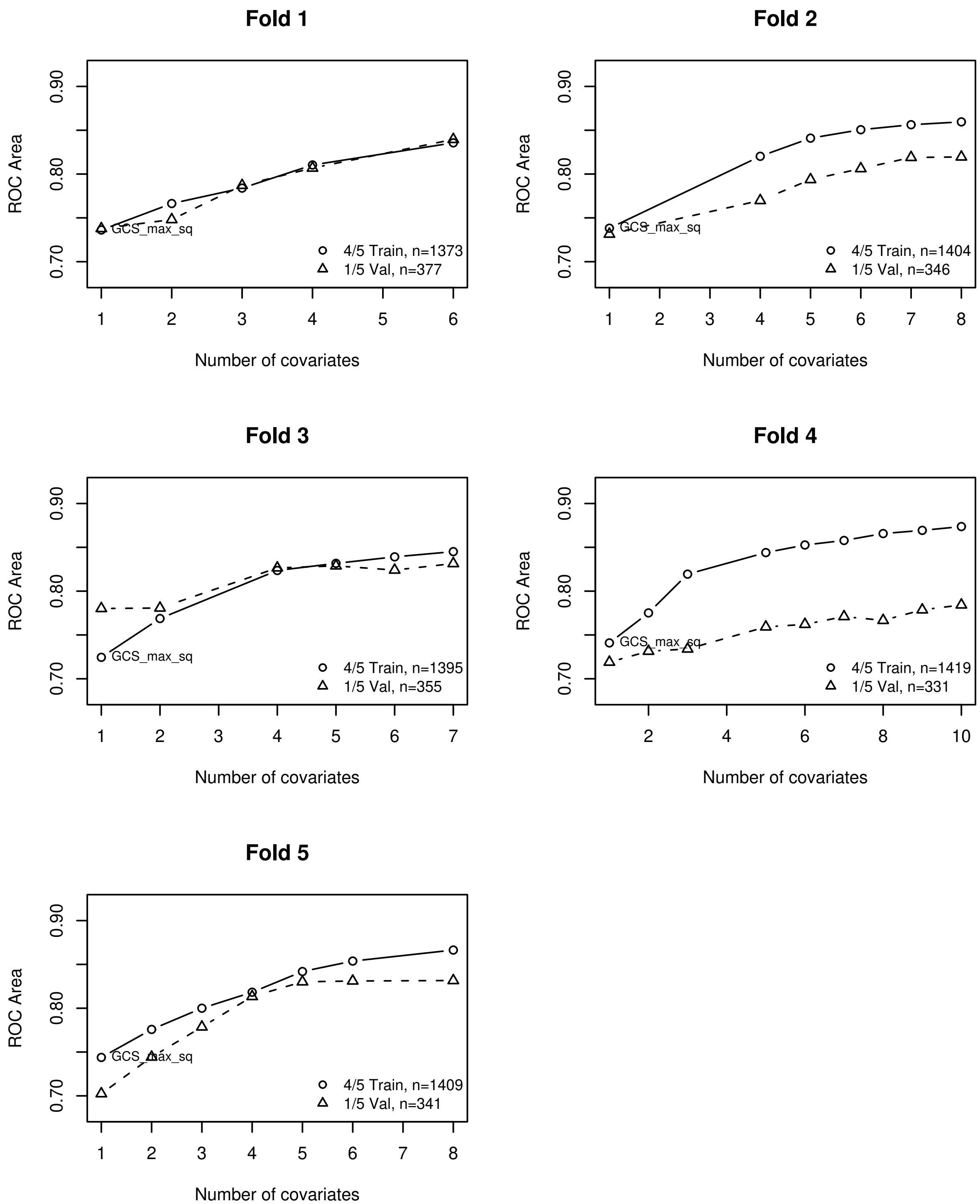


Figure C-8: DAS5 Model Selection Stage 2: Sensitivity to Number of Covariates on Each Cross Validation Fold (Day 5)

Appendix D

Hosmer-Lemeshow Tests for DASn Models

In this appendix, Hosmer-Lemeshow goodness of fit tests are provided for each DASn model.

Table D.1: DAS1: Hosmer-Lemeshow Goodness of Fit Test: Risk Deciles

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-3	[0.00027,0.01090)	0.004	5	7.9	1905	1902.1	1910
4	[0.01090,0.01925)	0.015	5	9.5	631	626.5	636
5	[0.01925,0.03284)	0.025	8	16.2	628	619.8	636
6	[0.03284,0.05405)	0.042	29	26.9	608	610.1	637
7	[0.05405,0.08823)	0.07	53	44.2	583	591.8	636
8	[0.08823,0.16135)	0.119	77	75.8	560	561.2	637
9	[0.16135,0.33751)	0.237	167	150.5	469	485.5	636
10	[0.33751,0.99911]	0.58	356	369	280	267	636

$\chi^2 = 13.03,\ d.f. = 6; p = 0.043$

Table D.2: `DAS2`: Hosmer-Lemeshow Goodness of Fit Test: Risk Deciles

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-3	[0.00031,0.00842)	0.004	3	6.8	1551	1547.2	1554
4	[0.00842,0.01375)	0.011	3	5.6	515	512.4	518
5	[0.01375,0.02284)	0.018	9	9.3	509	508.7	518
6	[0.02284,0.03880)	0.03	13	15.5	505	502.5	518
7	[0.03880,0.07282)	0.054	29	28	489	490	518
8	[0.07282,0.14273)	0.102	68	52.8	450	465.2	518
9	[0.14273,0.32484)	0.213	110	110.5	408	407.5	518
10	[0.32484,0.99953]	0.572	289	295.6	228	221.4	517

$$\chi^2 = 9.07,\ d.f. = 6;\ p = 0.170$$

Table D.3: `DAS3`: Hosmer-Lemeshow Goodness of Fit Test: Risk Deciles

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-3	[0.000385,0.00962)	0.004	5	4.3	1053	1053.7	1058
4	[0.009622,0.01846)	0.014	4	4.8	349	348.2	353
5	[0.018458,0.03251)	0.025	3	8.7	349	343.3	352
6	[0.032506,0.05560)	0.043	11	15.1	342	337.9	353
7	[0.055603,0.10060)	0.076	32	26.8	321	326.2	353
8	[0.100602,0.18451)	0.136	50	47.7	302	304.3	352
9	[0.184515,0.38466)	0.268	105	94.5	248	258.5	353
10	[0.384663,0.99884]	0.614	208	216	144	136	352

$$\chi^2 = 8.84,\ d.f. = 6;\ p = 0.183$$

Table D.4: `DAS4`: Hosmer-Lemeshow Goodness of Fit Test: Risk Deciles

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-4	[0.00035,0.0279)	0.009	11	8.8	930	932.2	941
5	[0.02790,0.0499)	0.038	9	8.9	226	226.1	235
6	[0.04994,0.0808)	0.064	16	15.1	219	219.9	235
7	[0.08081,0.1373)	0.105	19	24.7	216	210.3	235
8	[0.13727,0.2399)	0.183	51	42.9	184	192.1	235
9	[0.23992,0.4357)	0.324	72	76.2	163	158.8	235
10	[0.43567,0.9987]	0.674	157	158.5	78	76.5	235

$$\chi^2 = 4.36,\ d.f. = 5;\ p = 0.499$$

Table D.5: `DAS5`: Hosmer-Lemeshow Goodness of Fit Test: Risk Deciles

			Died		Survived		
Decile	Prob.Range	Prob.	Obs.	Exp.	Obs.	Exp.	Total
1-3	[0.000203,0.02133)	0.009	4	4.5	503	502.5	507
4	[0.021331,0.03864)	0.029	3	4.9	166	164.1	169
5	[0.038638,0.07023)	0.054	13	9.1	156	159.9	169
6	[0.070234,0.11305)	0.09	12	15.3	157	153.7	169
7	[0.113049,0.17858)	0.142	23	24.1	146	144.9	169
8	[0.178583,0.29531)	0.231	41	39.1	128	129.9	169
9	[0.295315,0.51441)	0.383	65	64.8	104	104.2	169
10	[0.514414,0.99608]	0.706	120	119.3	49	49.7	169

$\chi^2 = 3.59,\ d.f. = 6;\ p = 0.732$

Appendix E

RAS Individual Patient Risk Profiles

This appendix contains examples of the models developed in this thesis applied to 20 randomly selected patients who expired and 20 random patients who survived. The real-time acuity score (RAS) is discussed Chapter in 5. The secondary outcome models (SSOM, PWM, AKIM, and BPWM) are discussed in Chapter 6. To provide interesting examples, I required that patients have at least 20 RAS predictions to be considered for random selection.

Each patient plot displays the RAS predictions in solid black. If available, the secondary outcome predictions are also indicated: SSOM predictions are displayed with red dashes, PWM predictions are displayed with blue dots, AKIM predictions are displayed with a green dot-dash-dot pattern, and the BPWM predictions are displayed with long cyan dashes. If the outcome of interest for a secondary model is present (e.g., SIRS and HDFR for SSOM), the corresponding line (type and color) is shown along the x-axis for the duration of the outcome.

E.1 Expired Patients

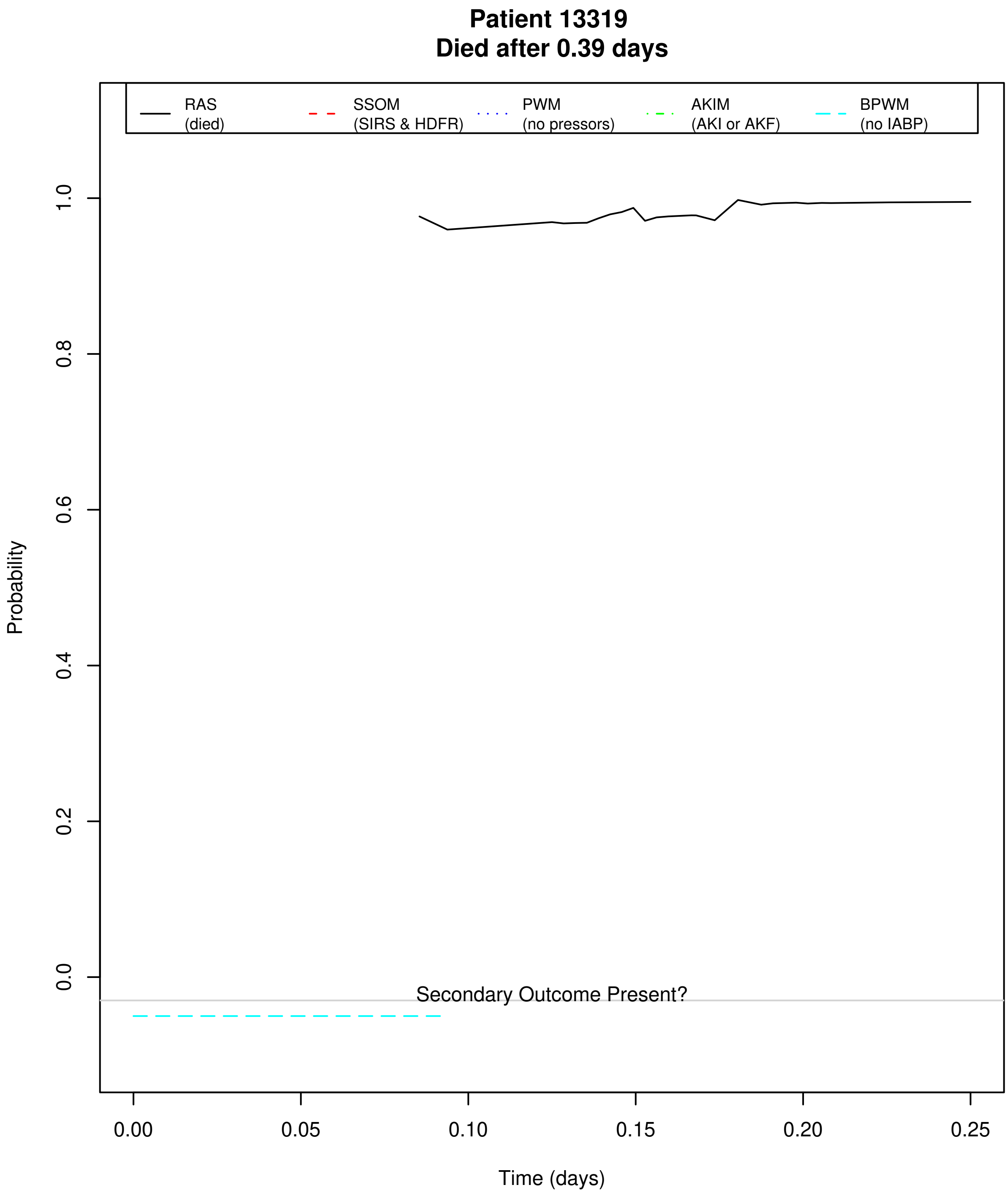


Figure E-1: Model outputs for patient 13319. The presence of secondary outcomes is marked along the x-axis.

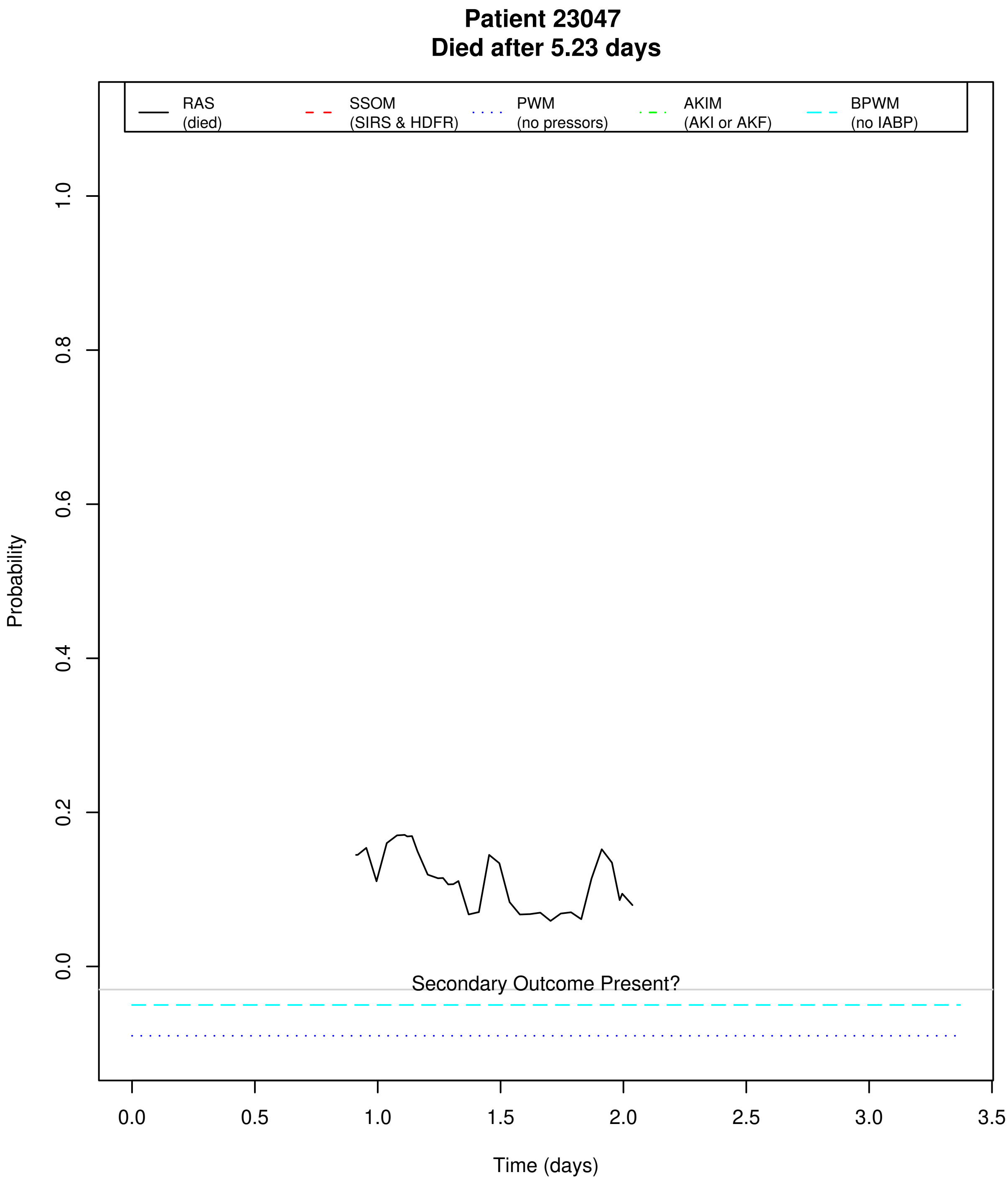


Figure E-2: Model outputs for patient 23047. The presence of secondary outcomes is marked along the x-axis.

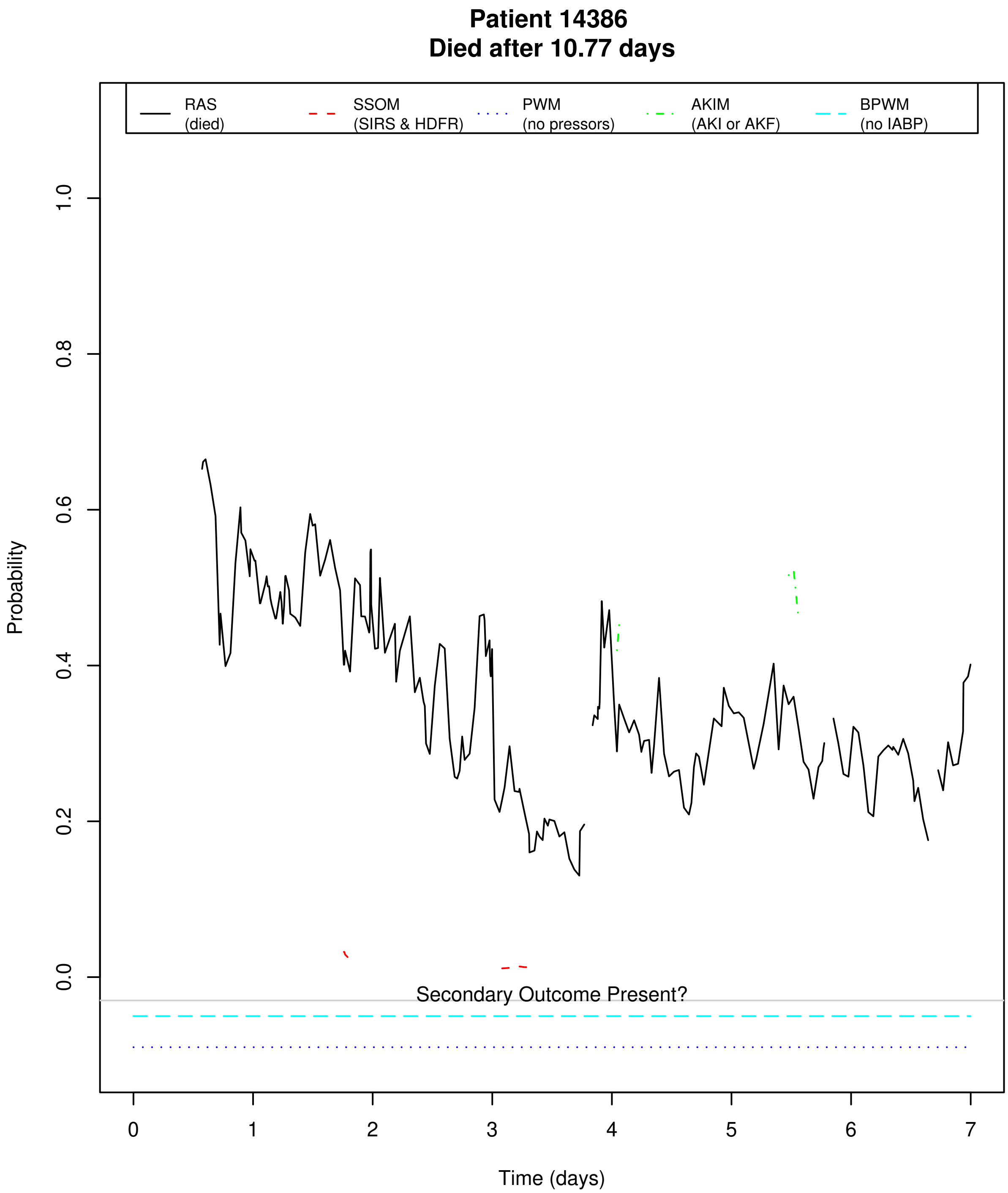


Figure E-3: Model outputs for patient 14386. The presence of secondary outcomes is marked along the x-axis.

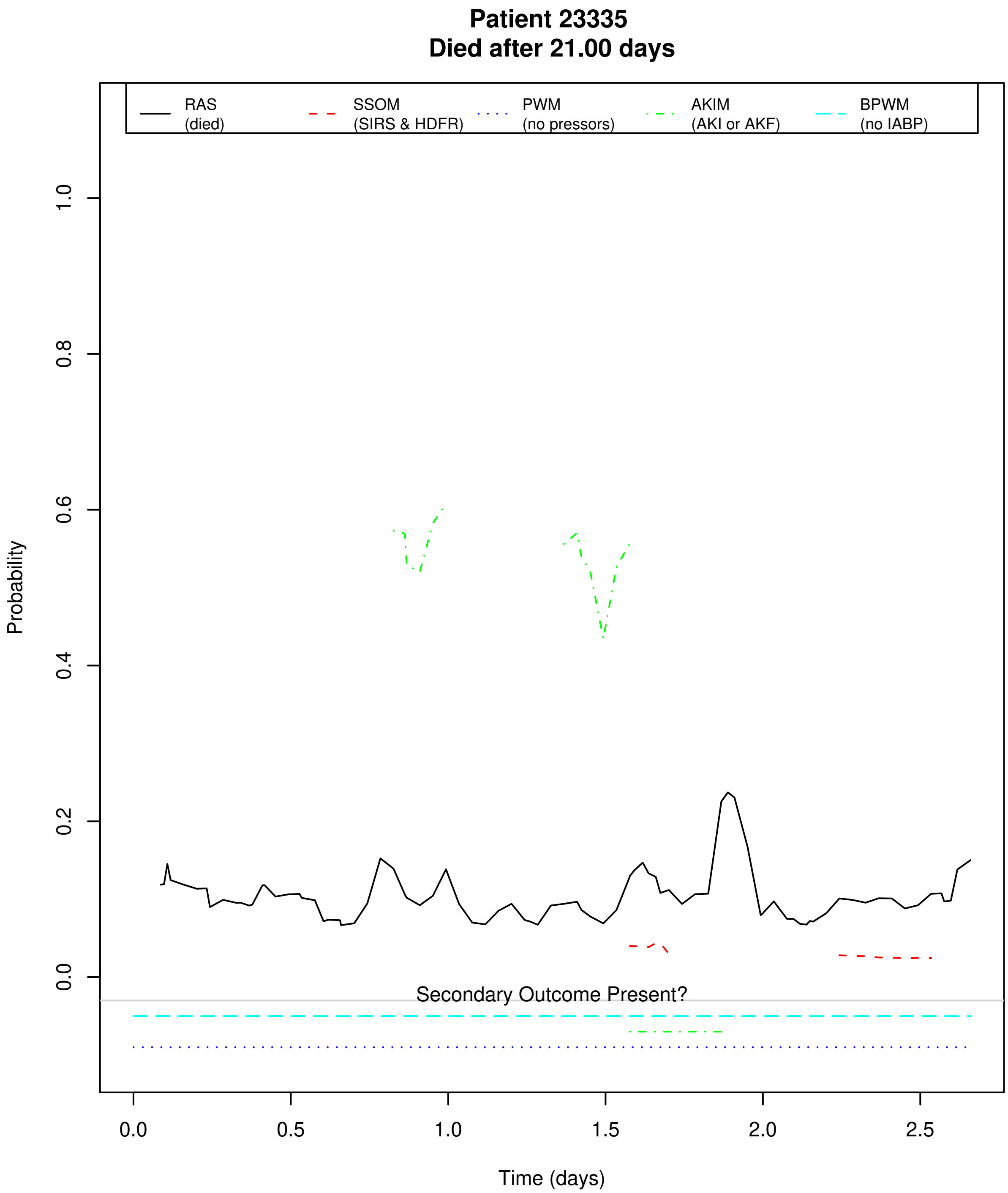


Figure E-4: Model outputs for patient 23335. The presence of secondary outcomes is marked along the x-axis.

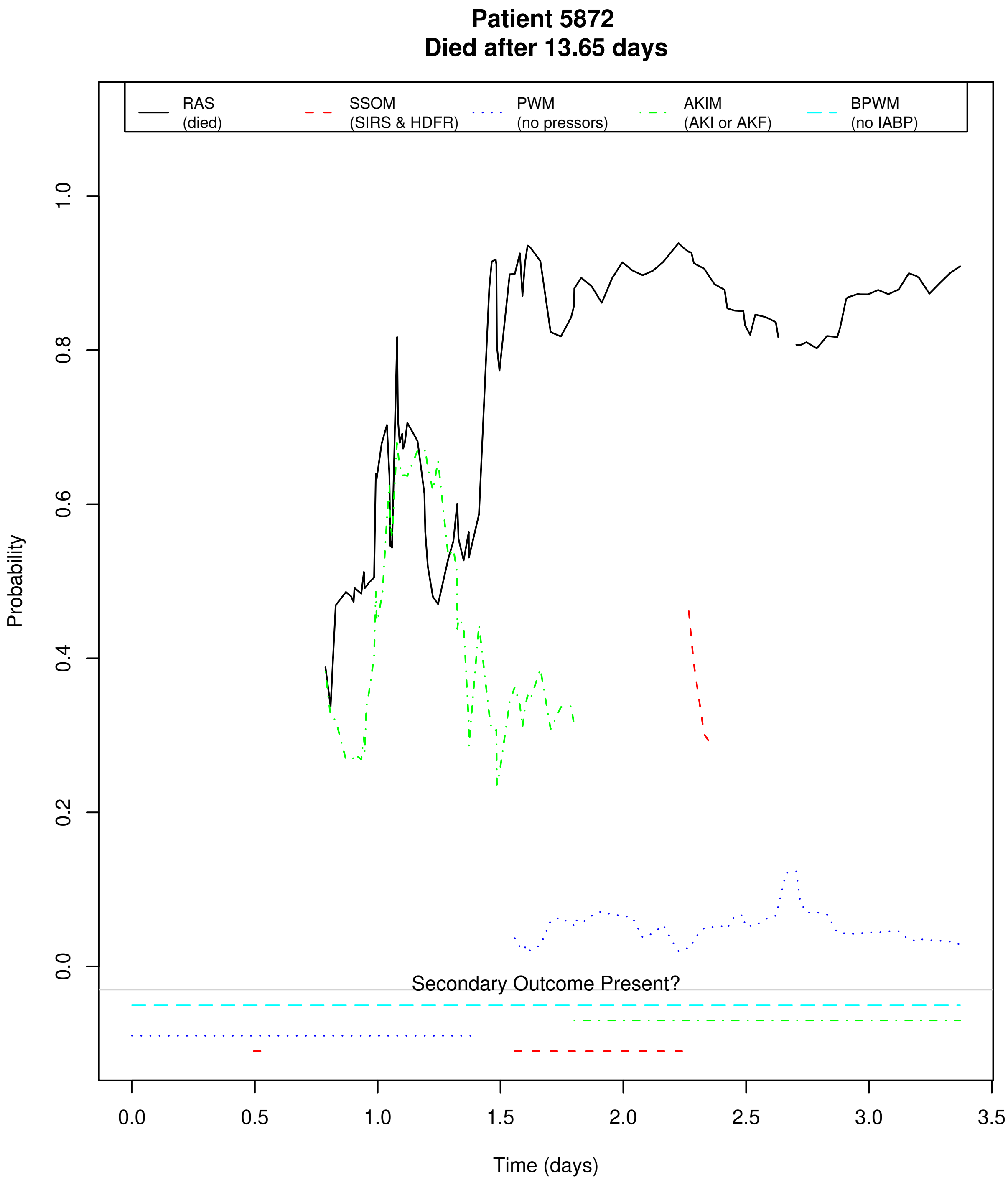


Figure E-5: Model outputs for patient 5872. The presence of secondary outcomes is marked along the x-axis.

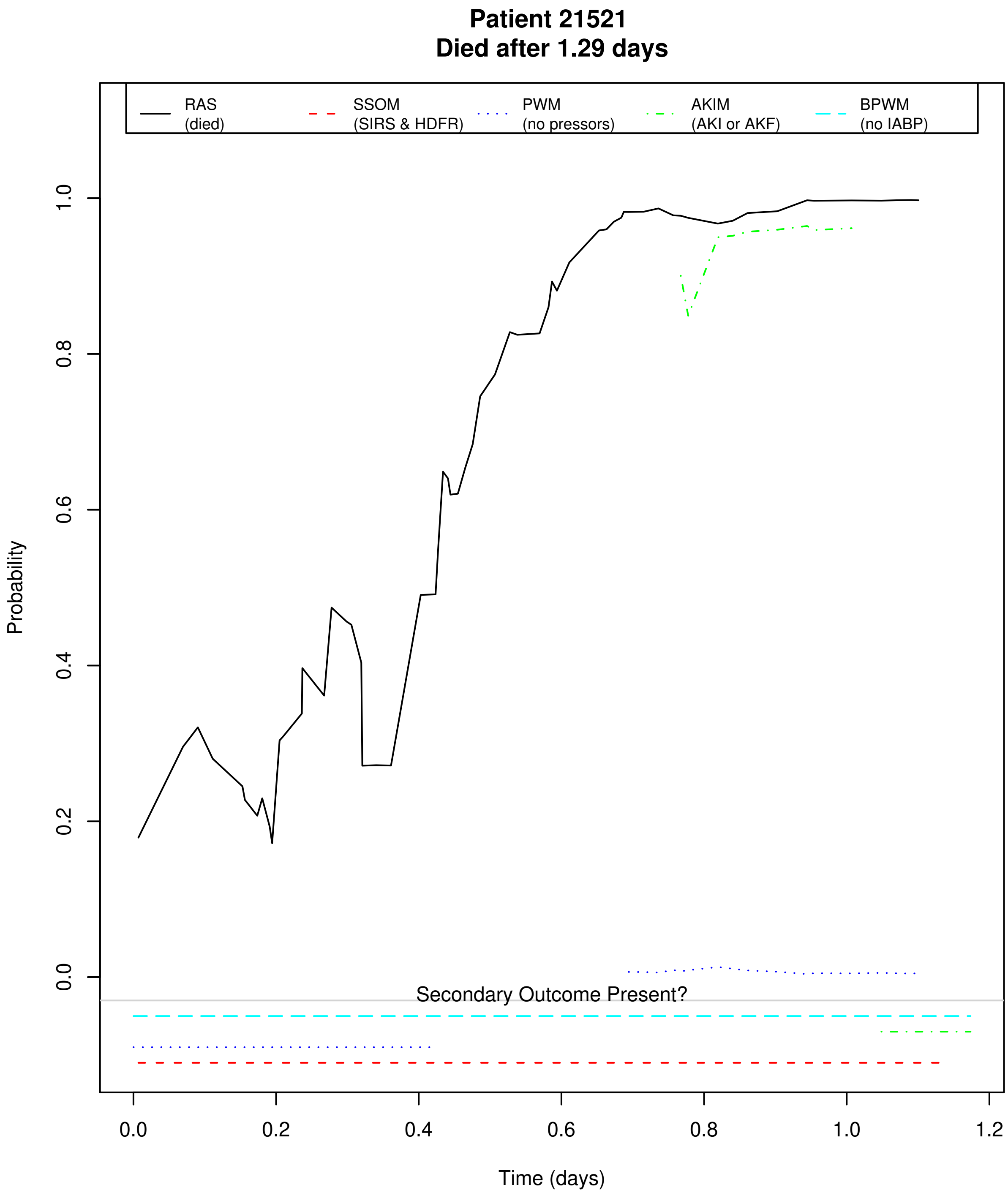


Figure E-6: Model outputs for patient 21521. The presence of secondary outcomes is marked along the x-axis.

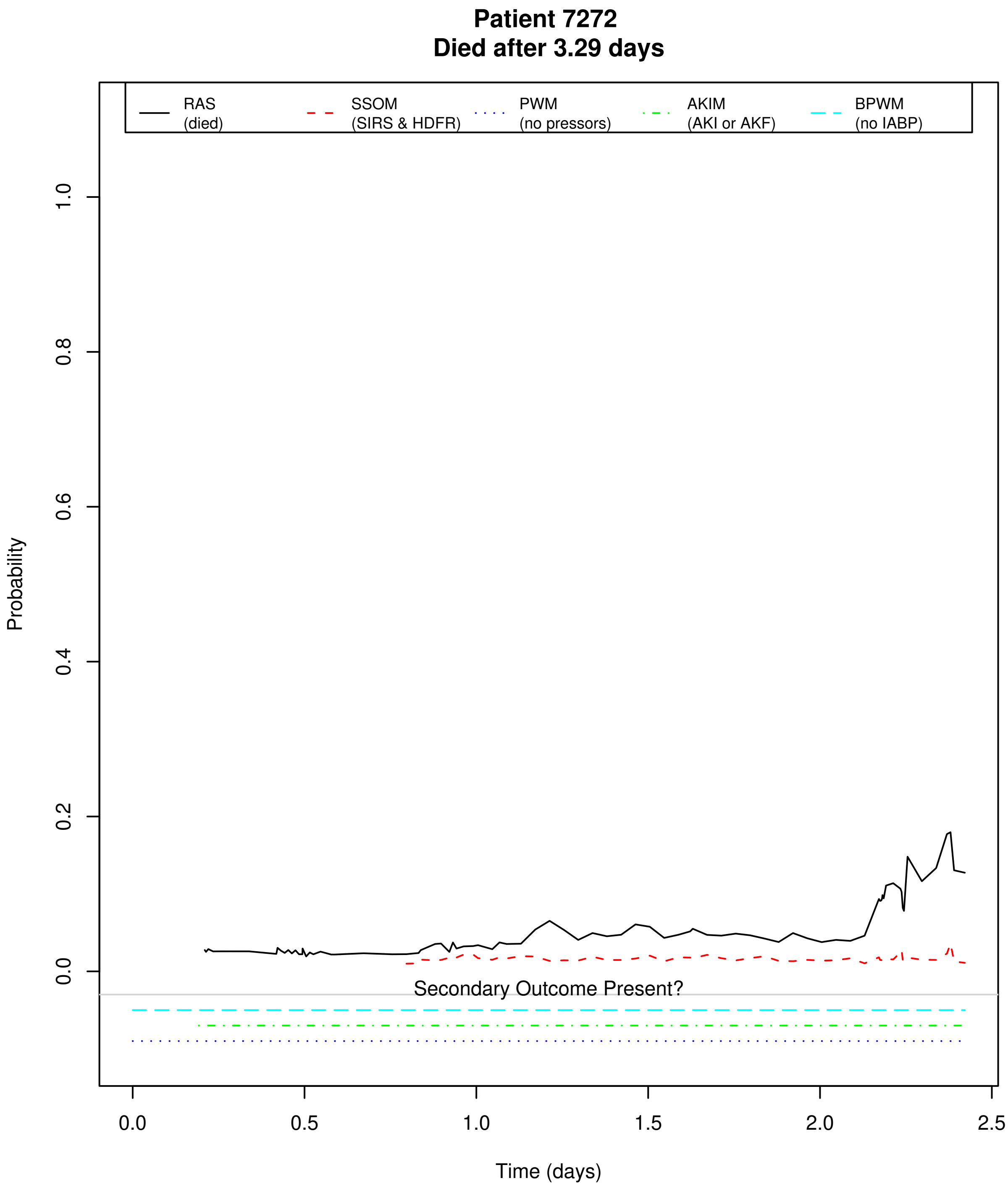


Figure E-7: Model outputs for patient 7272. The presence of secondary outcomes is marked along the x-axis.

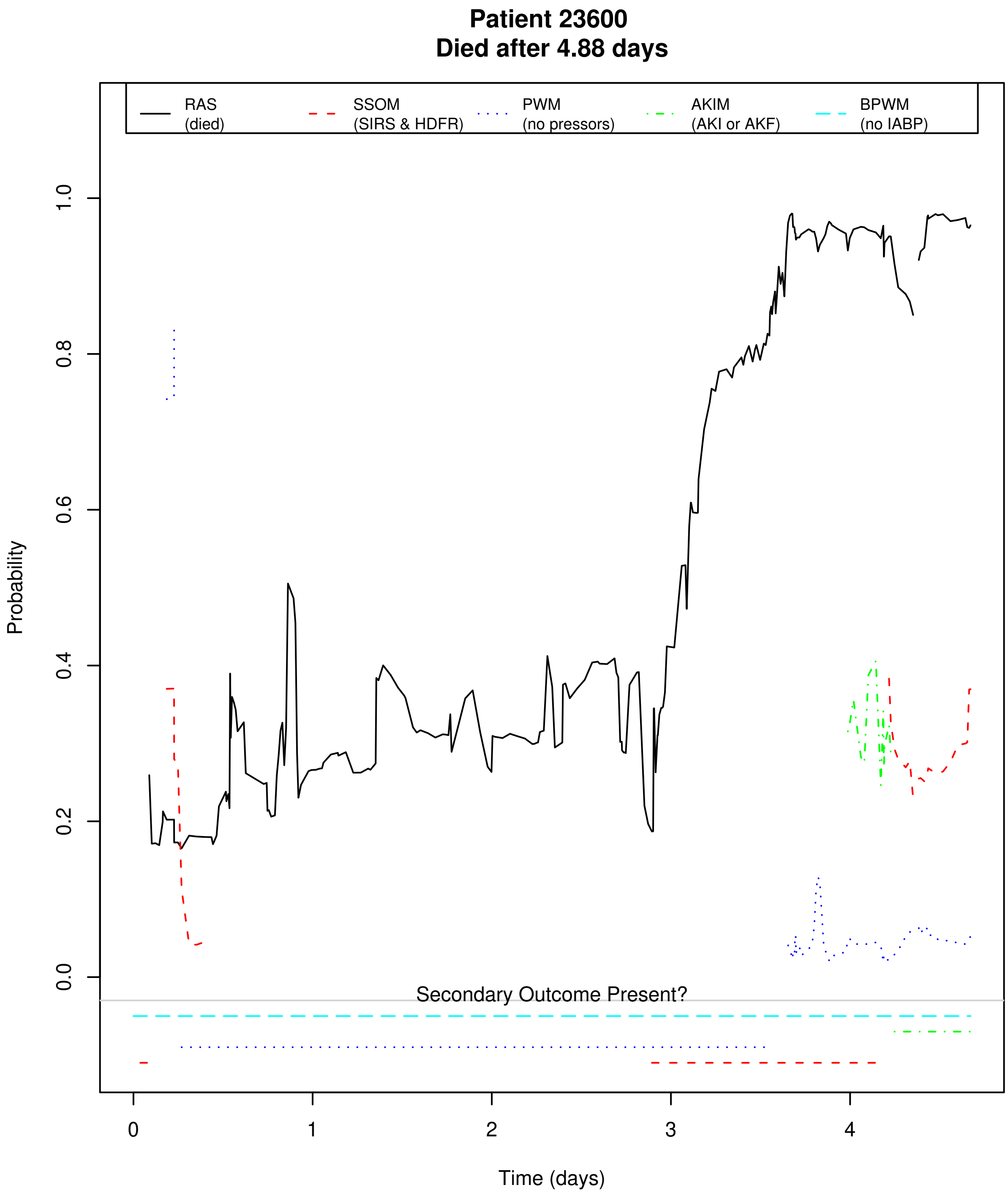


Figure E-8: Model outputs for patient 23600. The presence of secondary outcomes is marked along the x-axis.

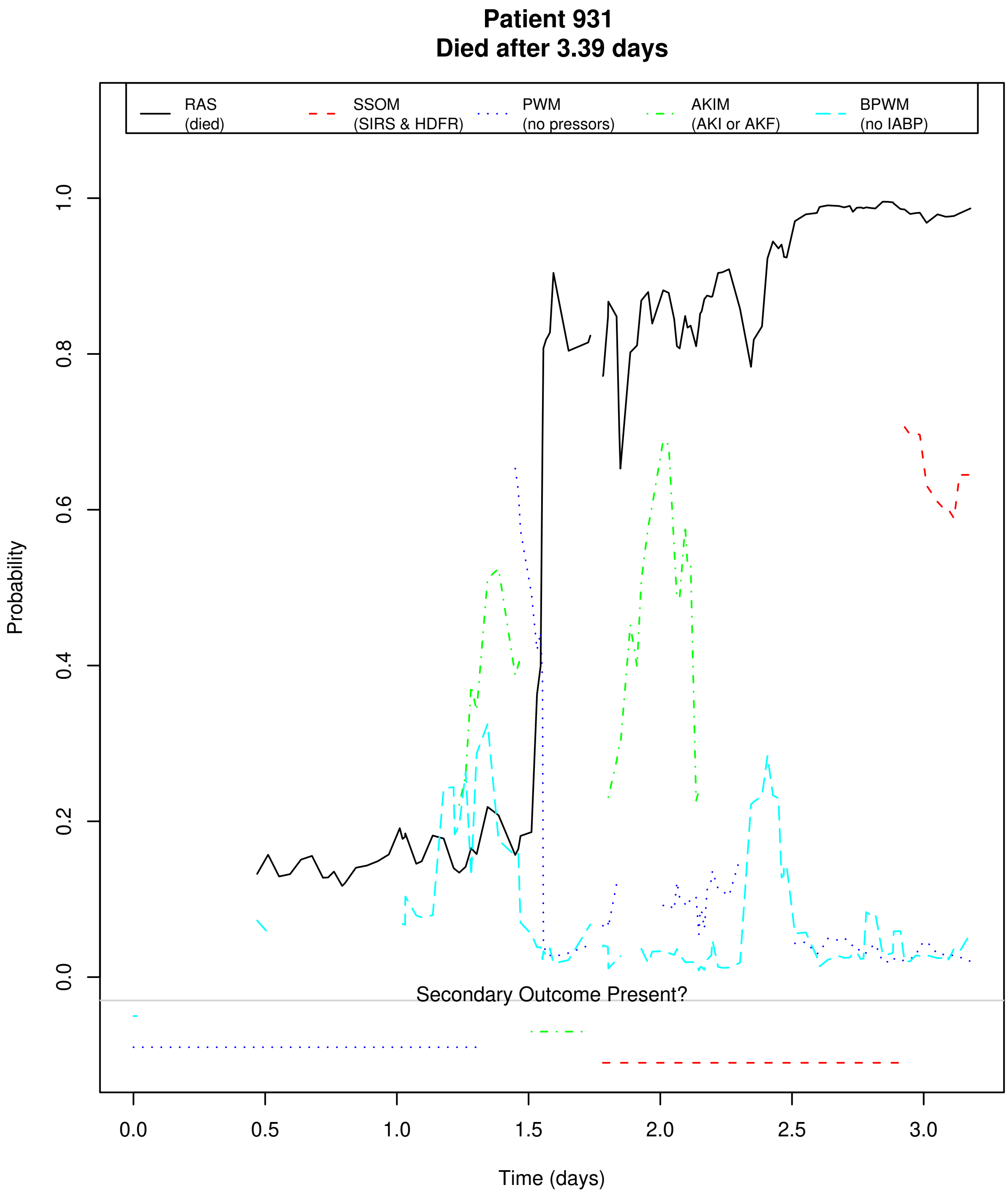


Figure E-9: Model outputs for patient 931. The presence of secondary outcomes is marked along the x-axis.

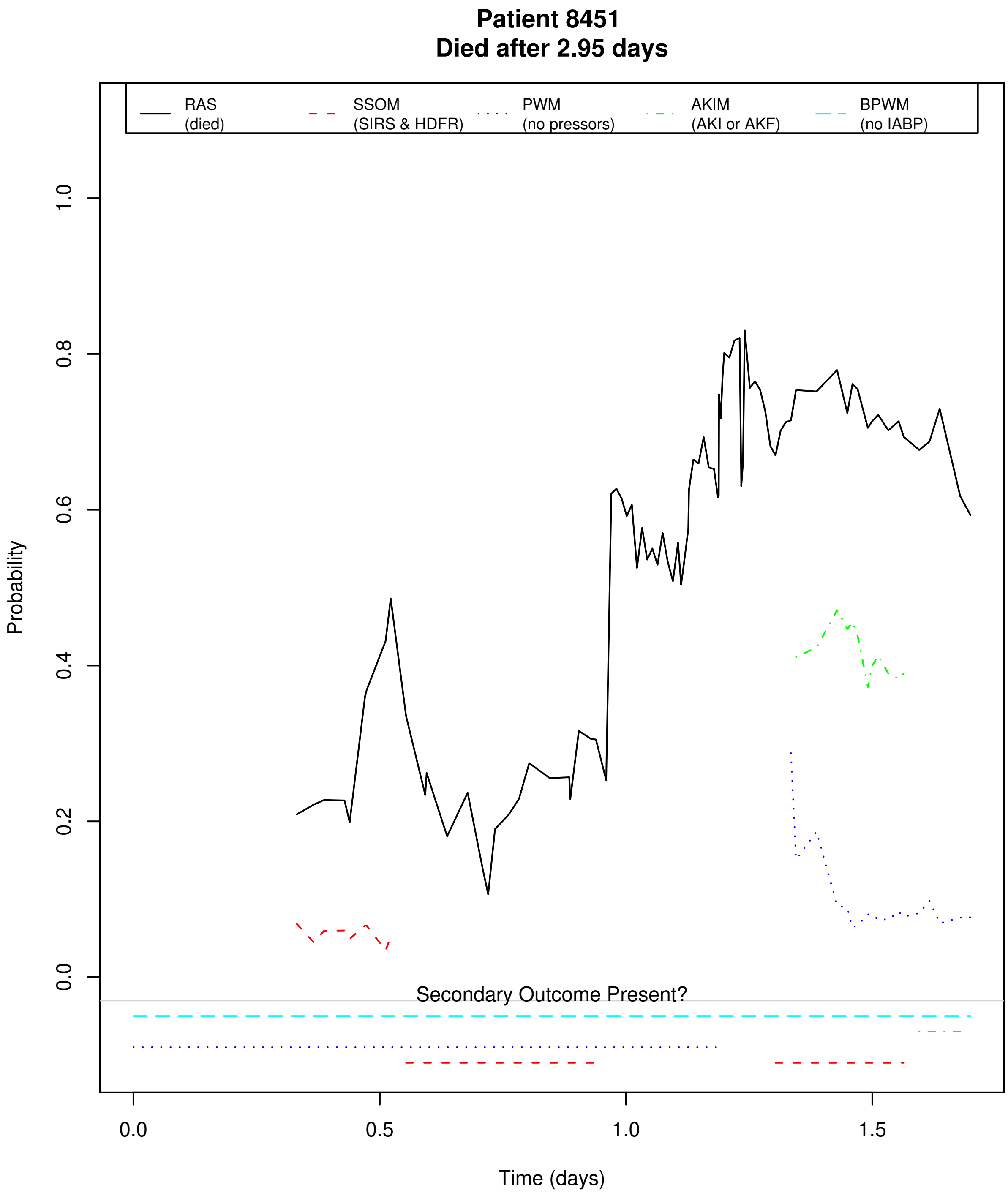


Figure E-10: Model outputs for patient 8451. The presence of secondary outcomes is marked along the x-axis.

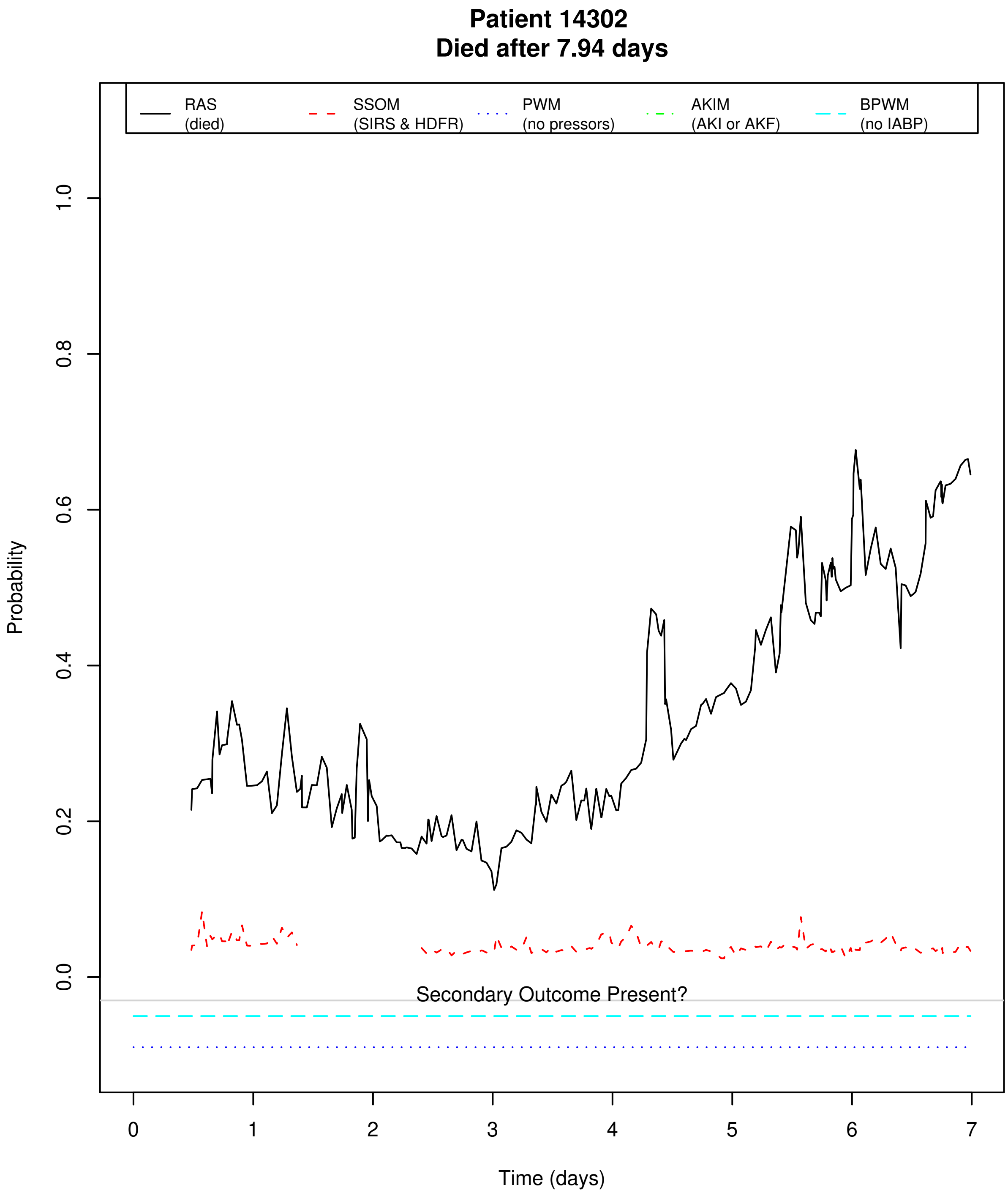


Figure E-11: Model outputs for patient 14302. The presence of secondary outcomes is marked along the x-axis.

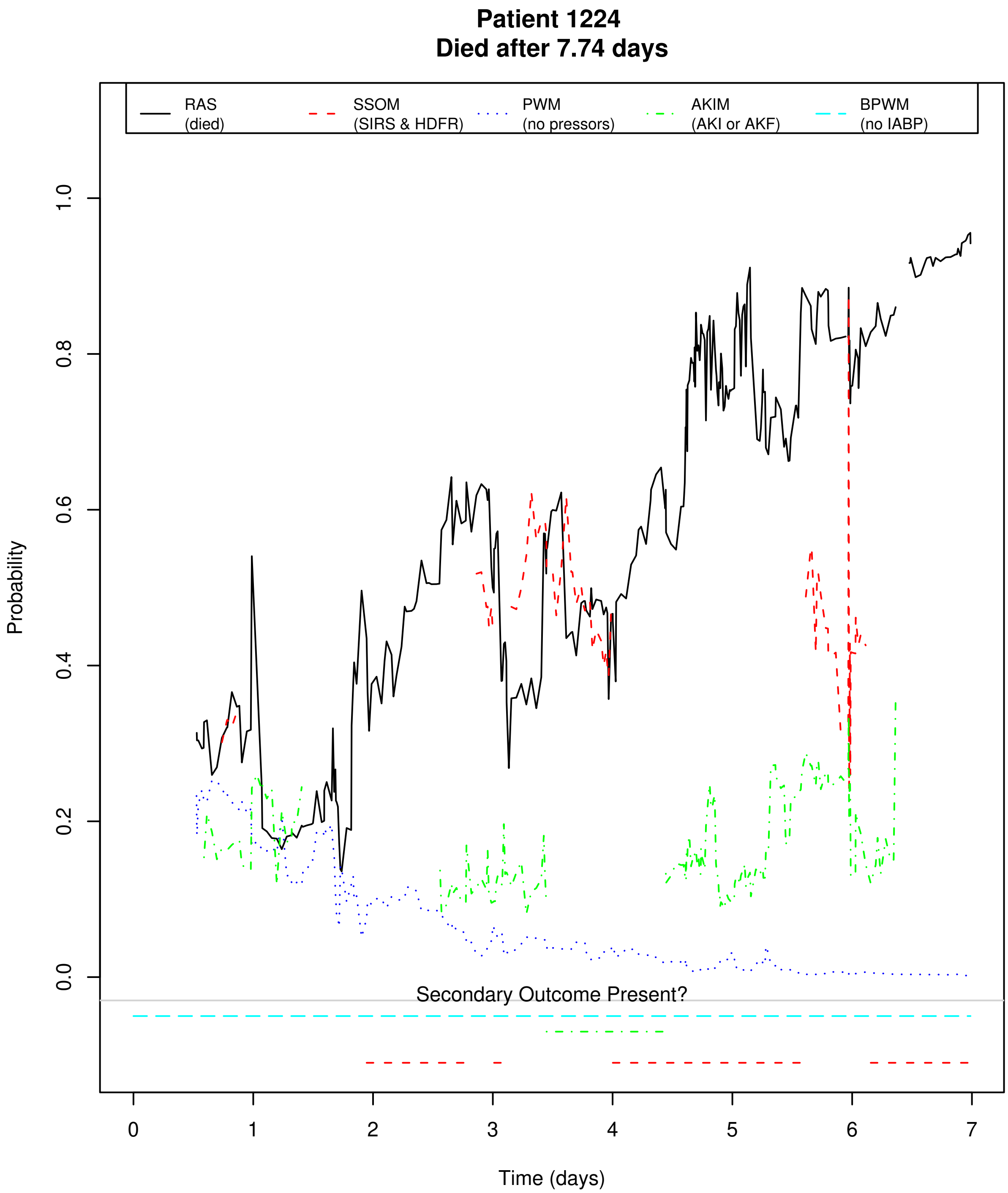


Figure E-12: Model outputs for patient 1224. The presence of secondary outcomes is marked along the x-axis.

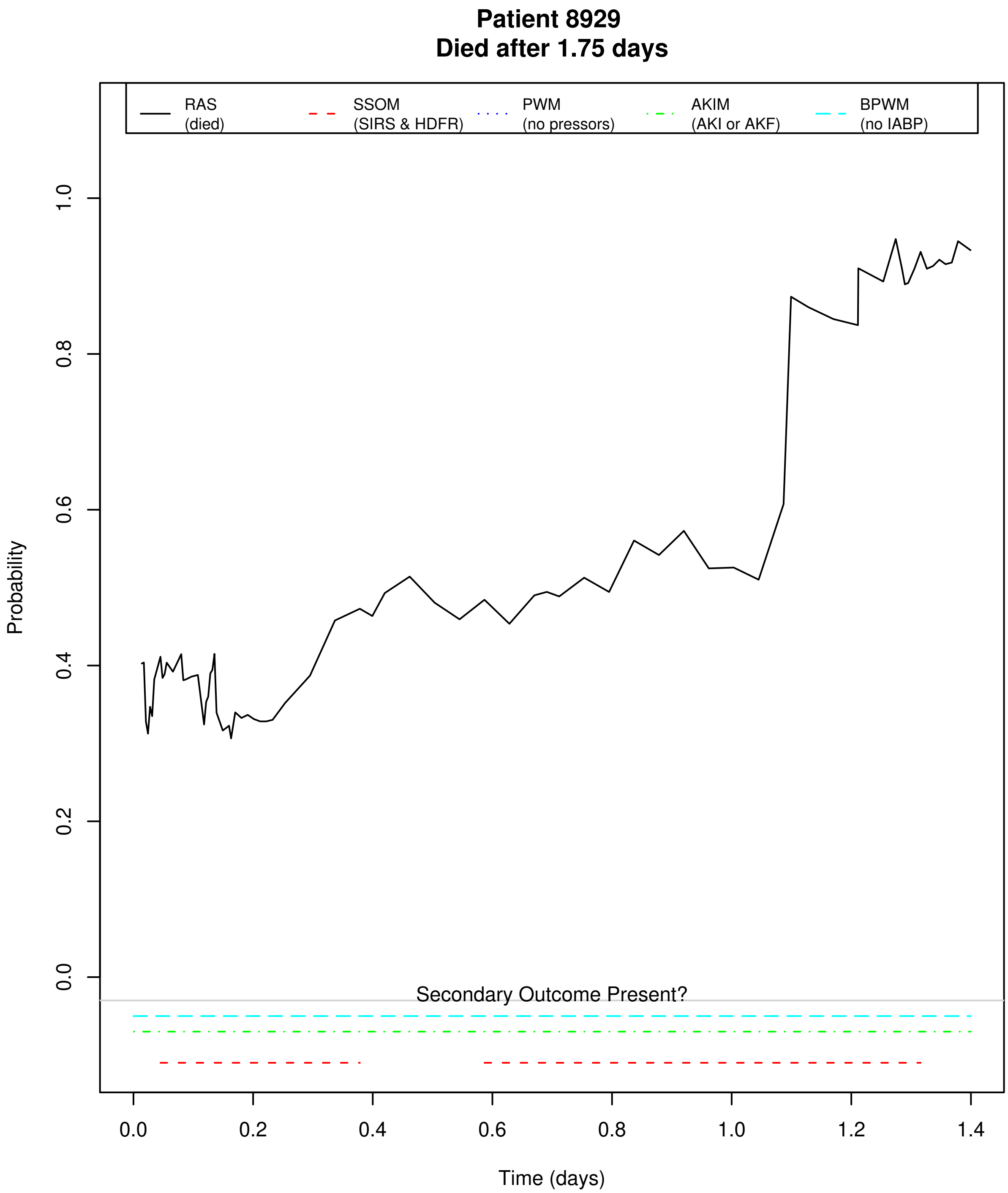


Figure E-13: Model outputs for patient 8929. The presence of secondary outcomes is marked along the x-axis.

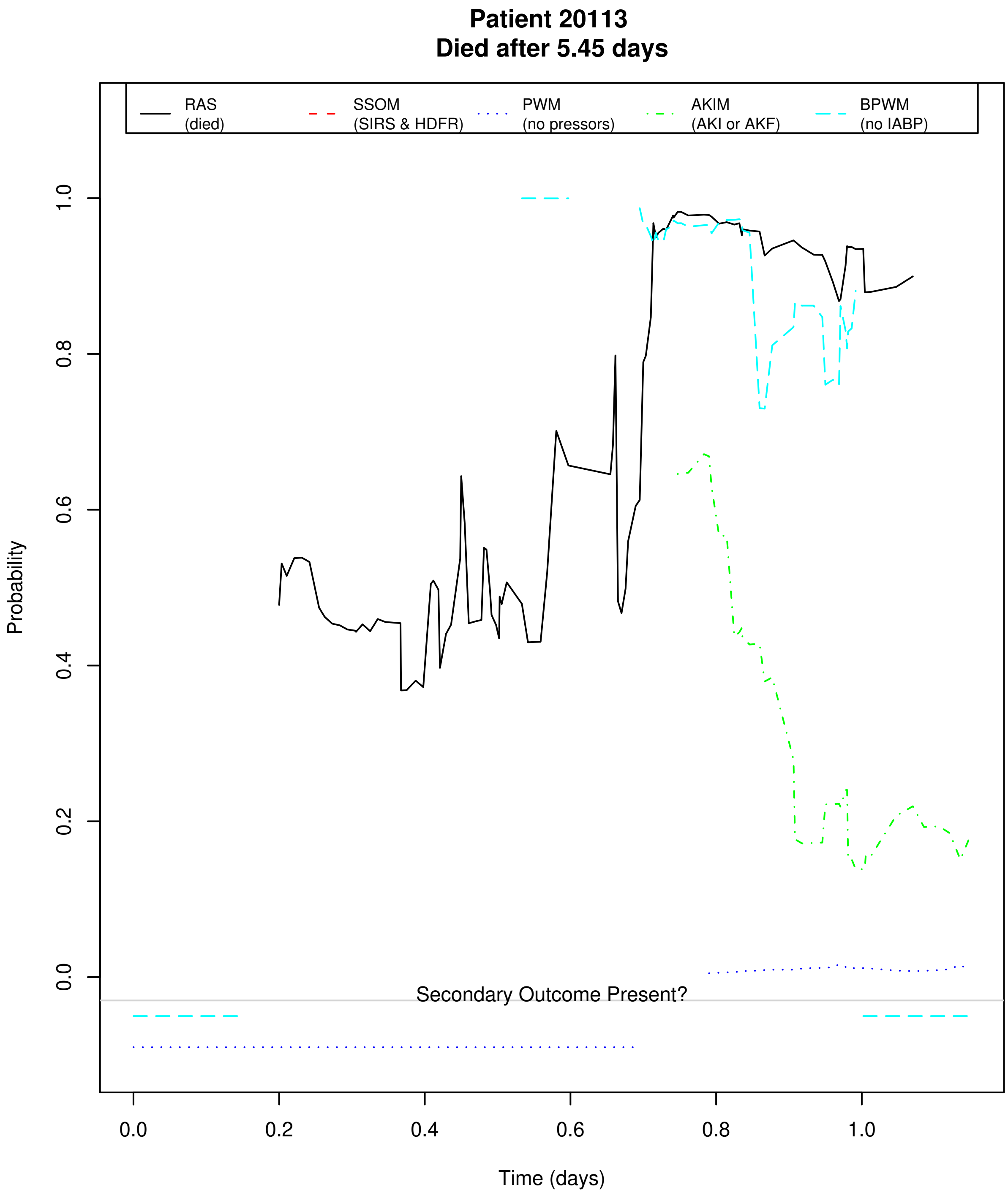


Figure E-14: Model outputs for patient 20113. The presence of secondary outcomes is marked along the x-axis.

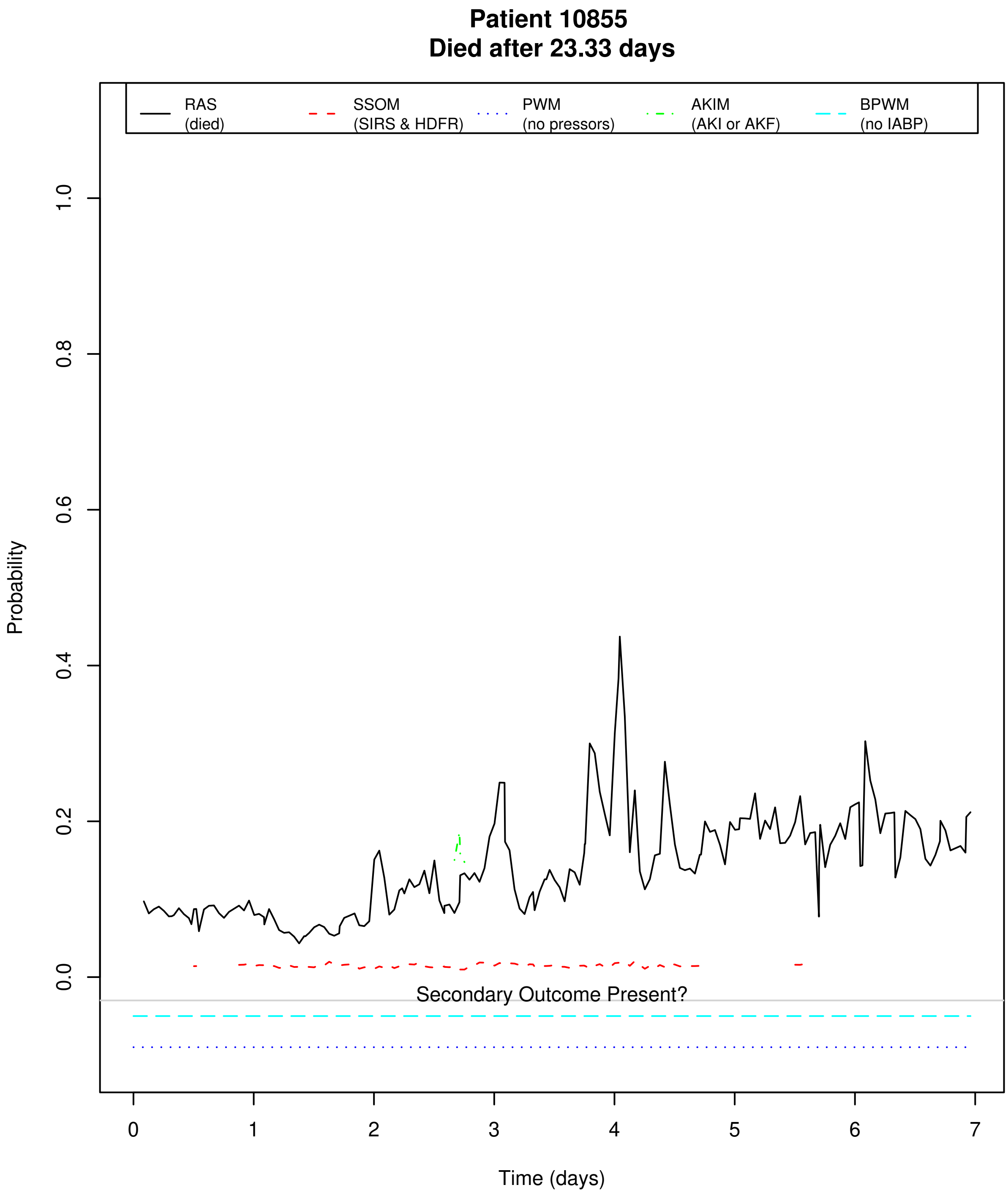


Figure E-15: Model outputs for patient 10855. The presence of secondary outcomes is marked along the x-axis.

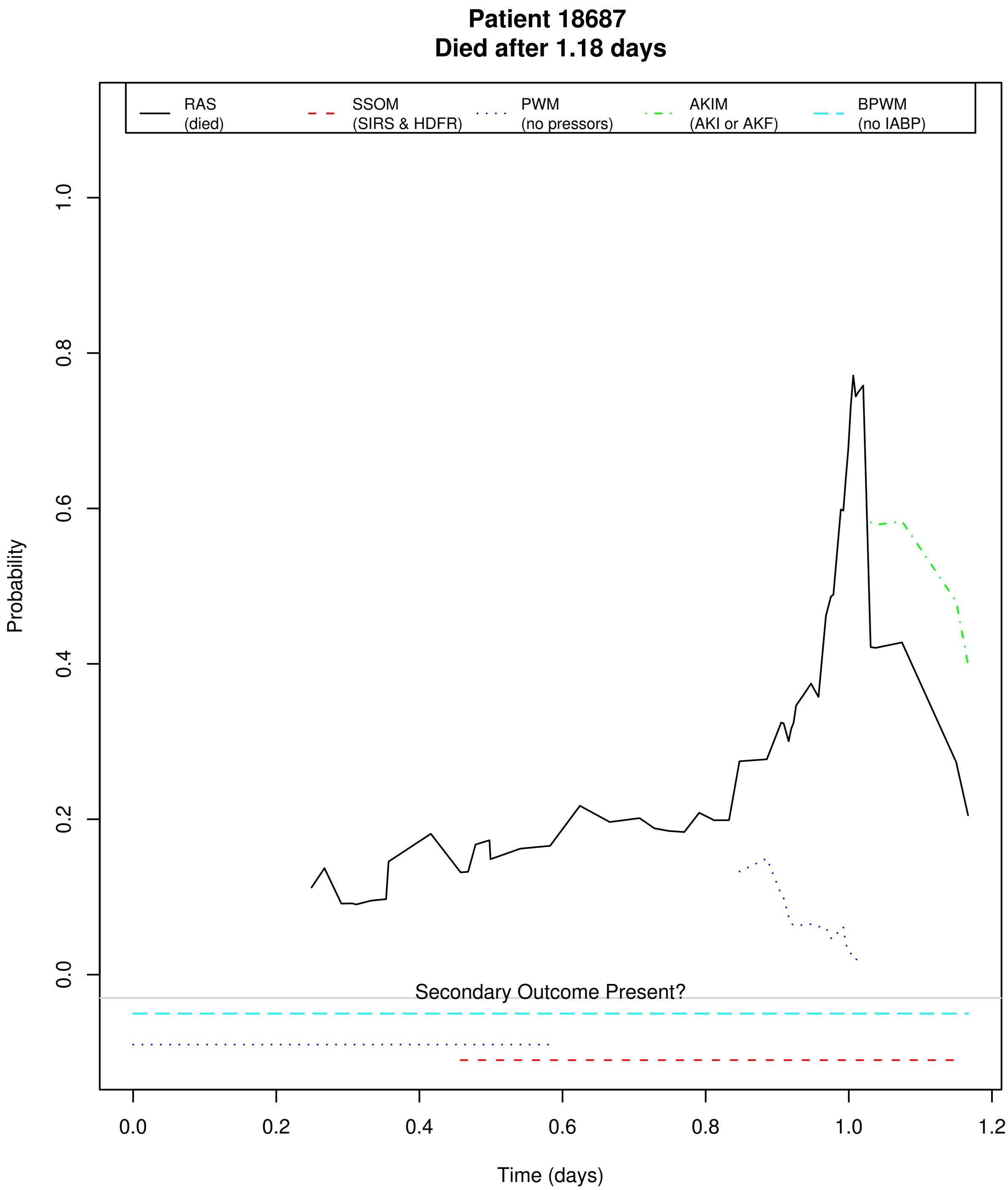


Figure E-16: Model outputs for patient 18687. The presence of secondary outcomes is marked along the x-axis.

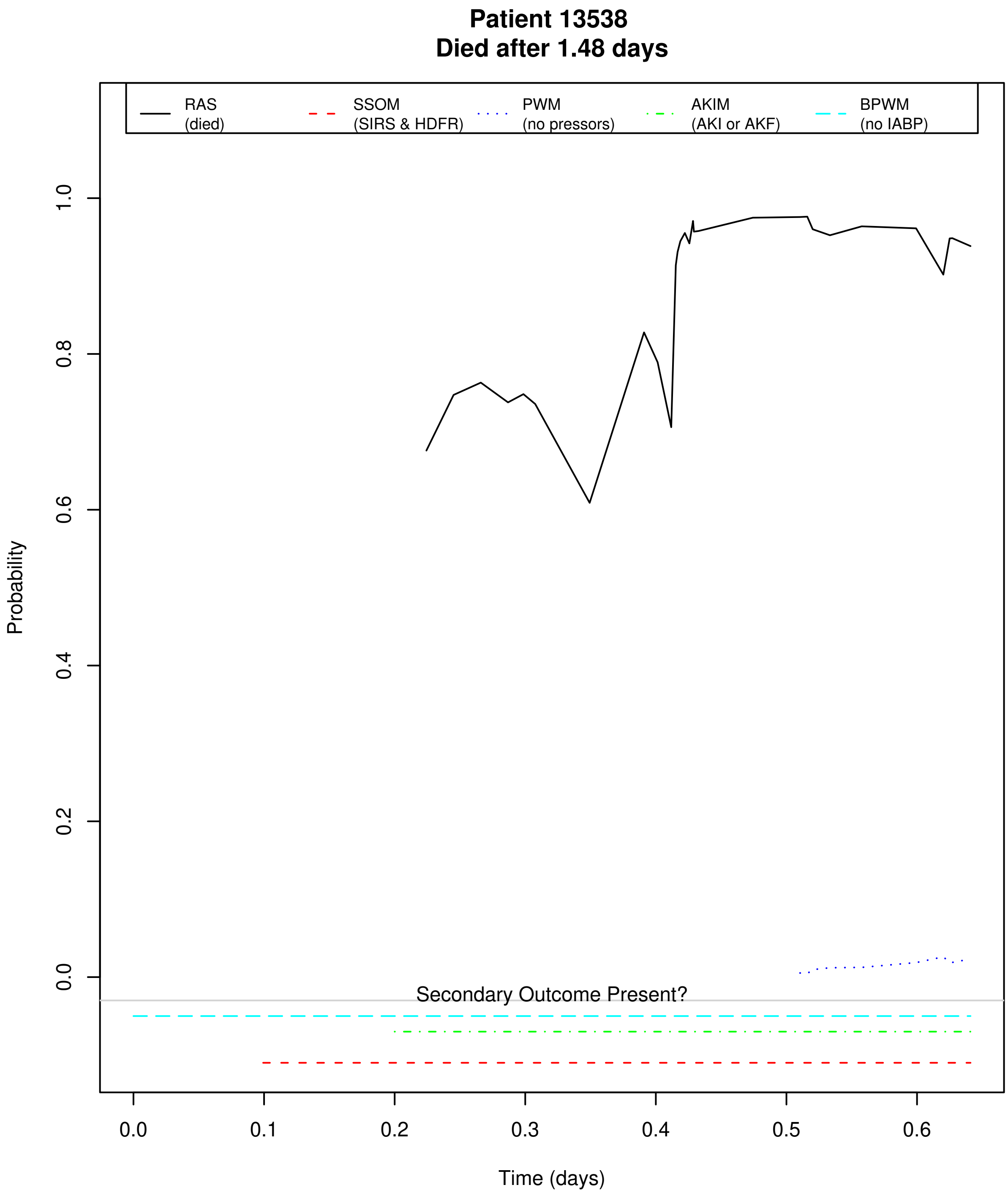


Figure E-17: Model outputs for patient 13538. The presence of secondary outcomes is marked along the x-axis.

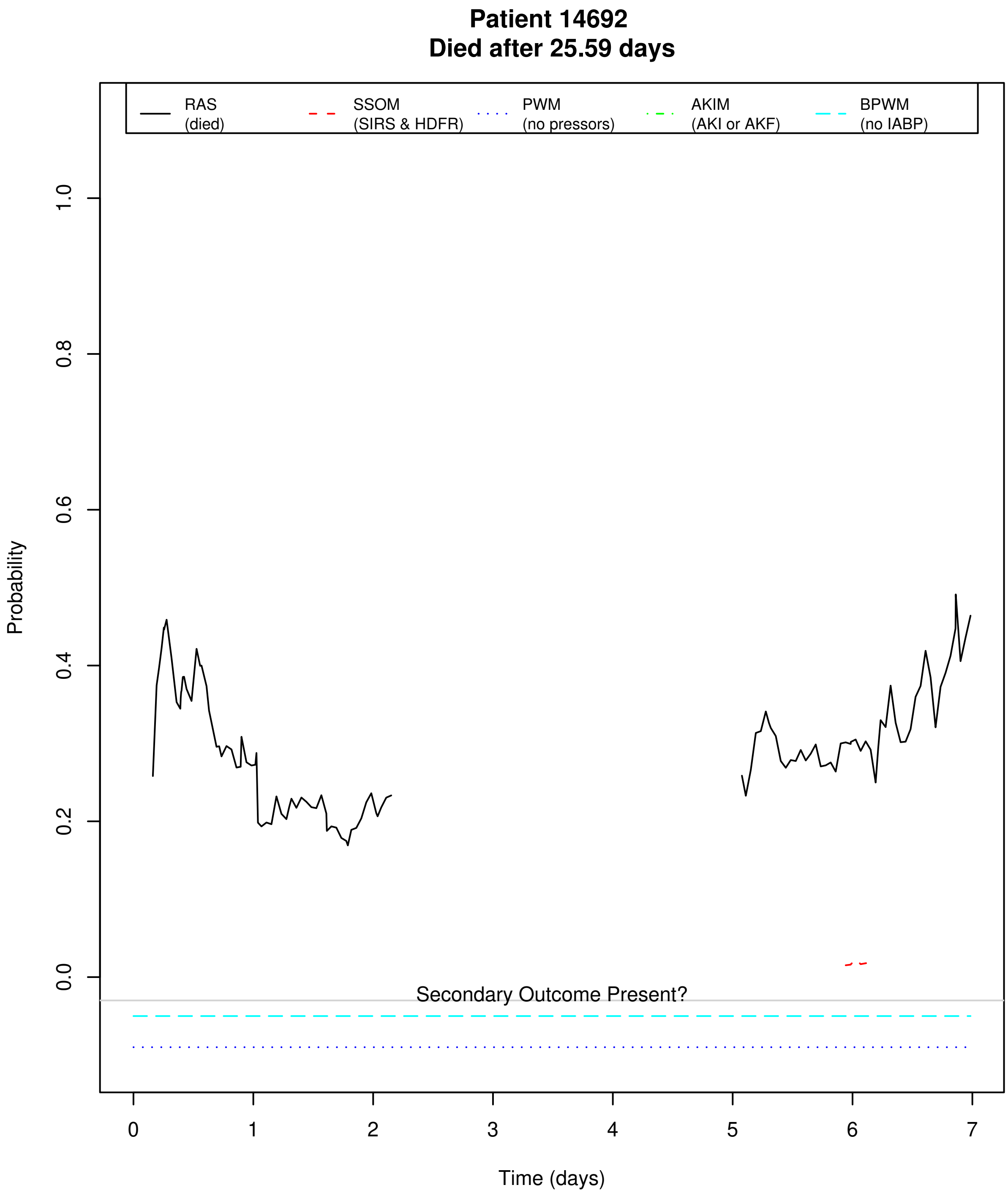


Figure E-18: Model outputs for patient 14692. The presence of secondary outcomes is marked along the x-axis.

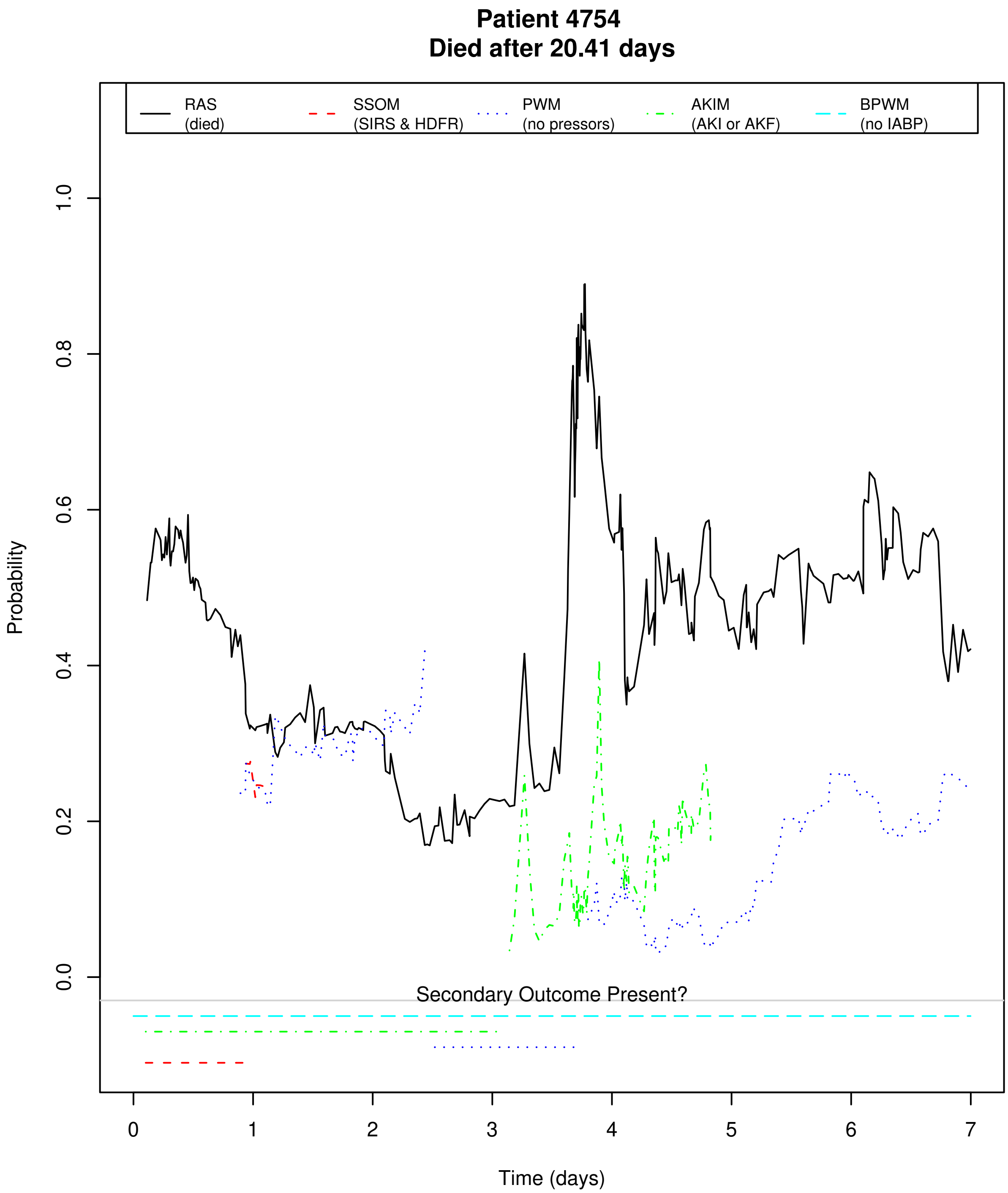


Figure E-19: Model outputs for patient 4754. The presence of secondary outcomes is marked along the x-axis.

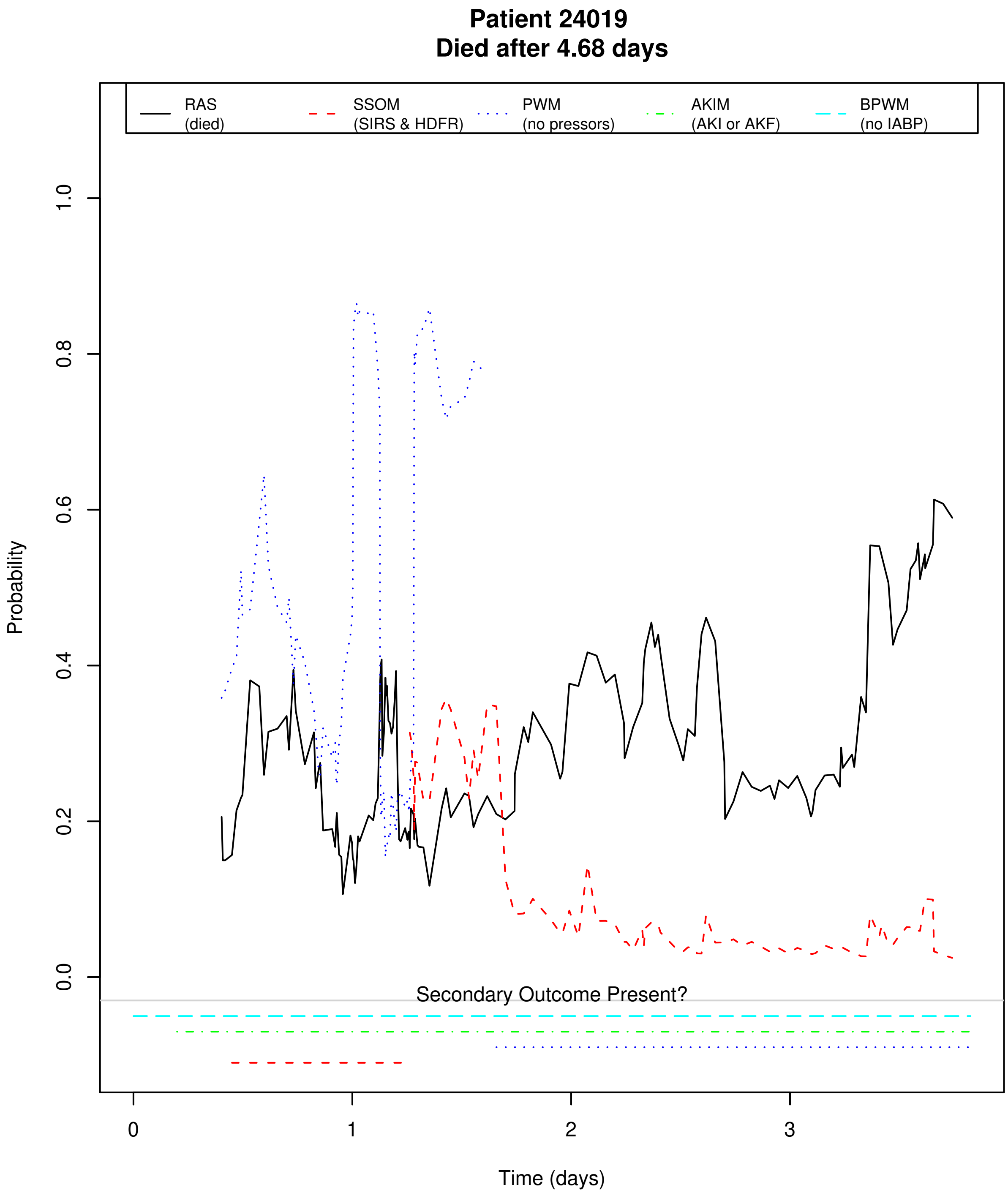


Figure E-20: Model outputs for patient 24019. The presence of secondary outcomes is marked along the x-axis.

E.2 Survived Patients

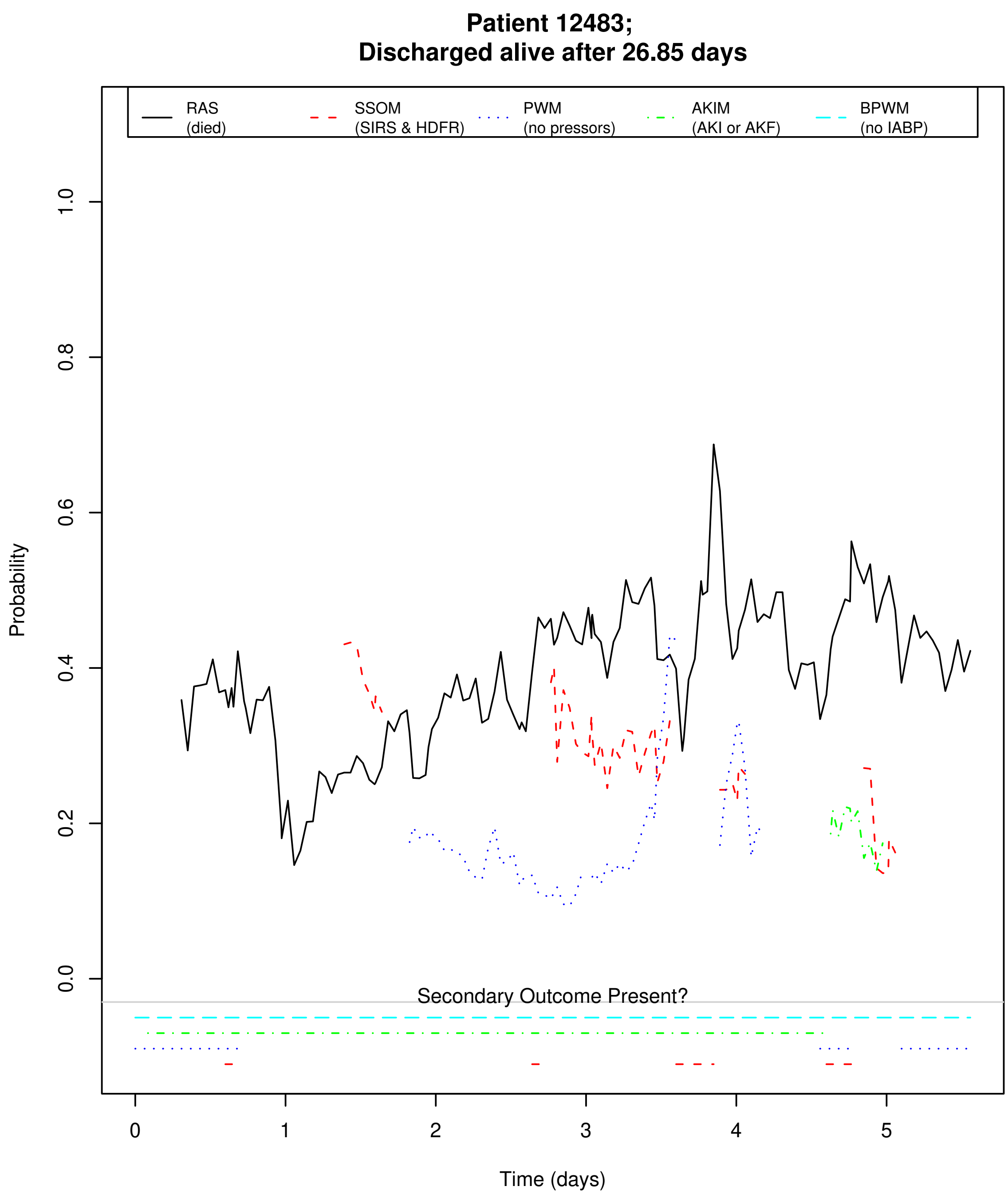


Figure E-21: Model outputs for patient 12483. The presence of secondary outcomes is marked along the x-axis.

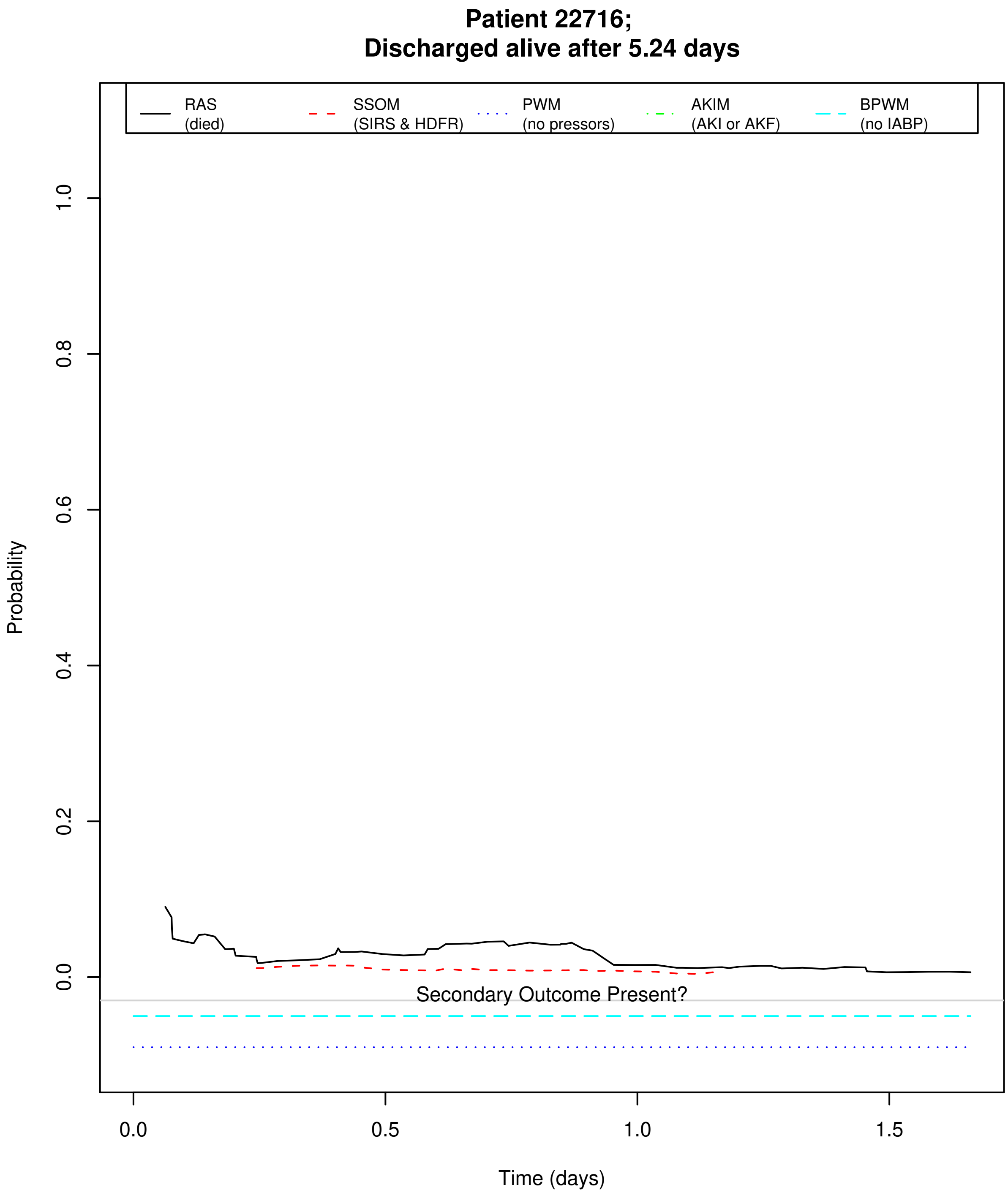


Figure E-22: Model outputs for patient 22716. The presence of secondary outcomes is marked along the x-axis.

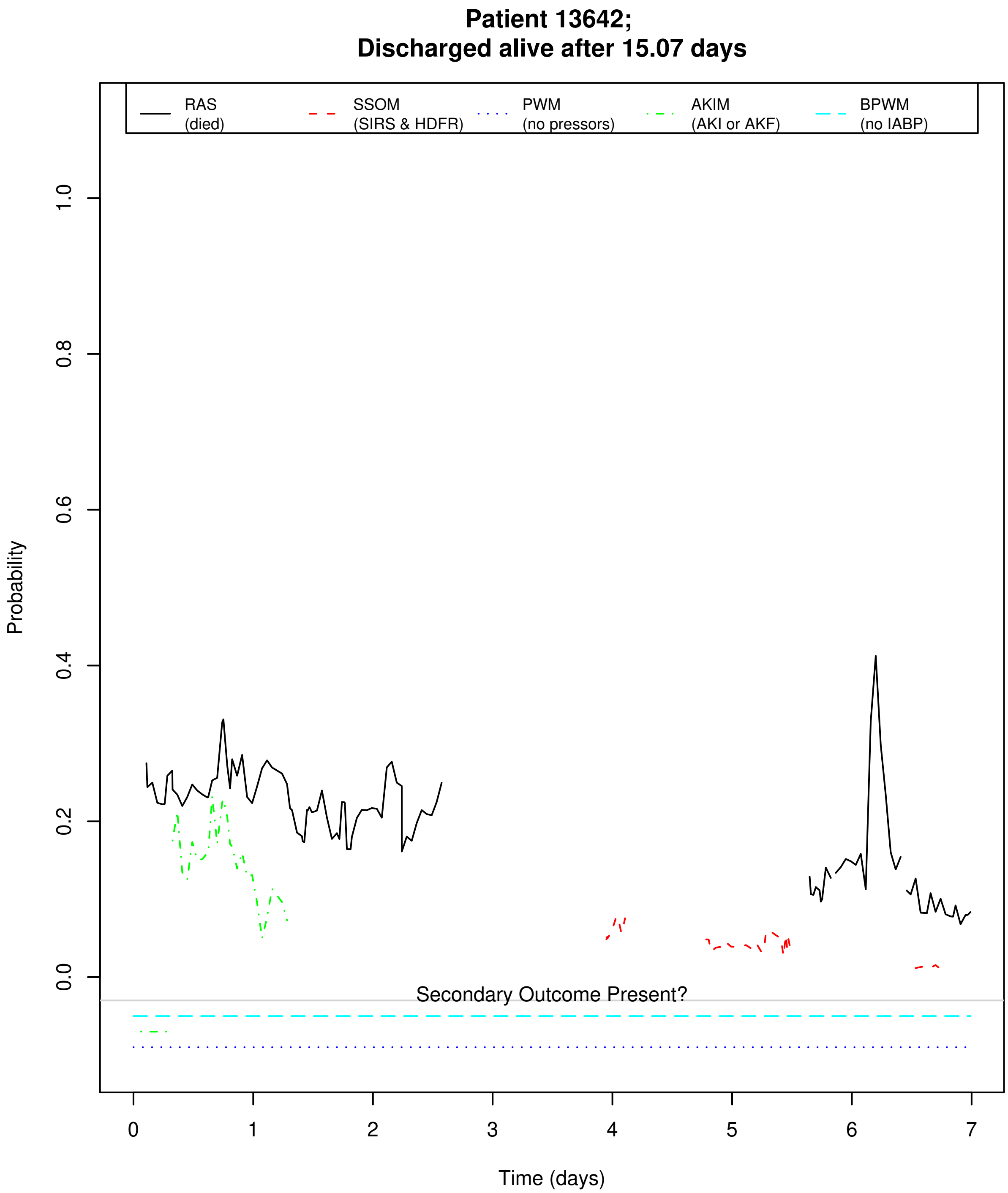


Figure E-23: Model outputs for patient 13642. The presence of secondary outcomes is marked along the x-axis.

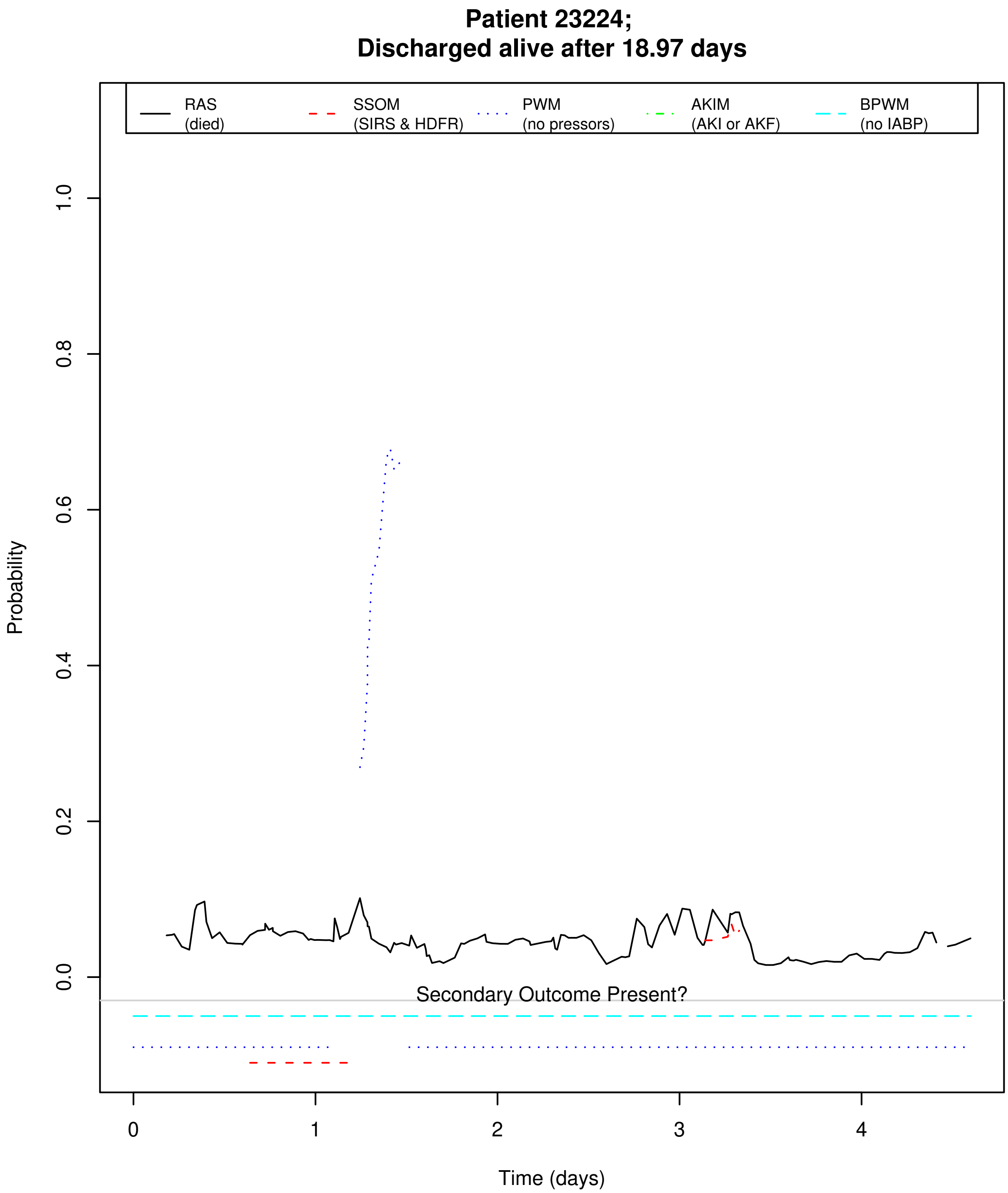


Figure E-24: Model outputs for patient 23224. The presence of secondary outcomes is marked along the x-axis.

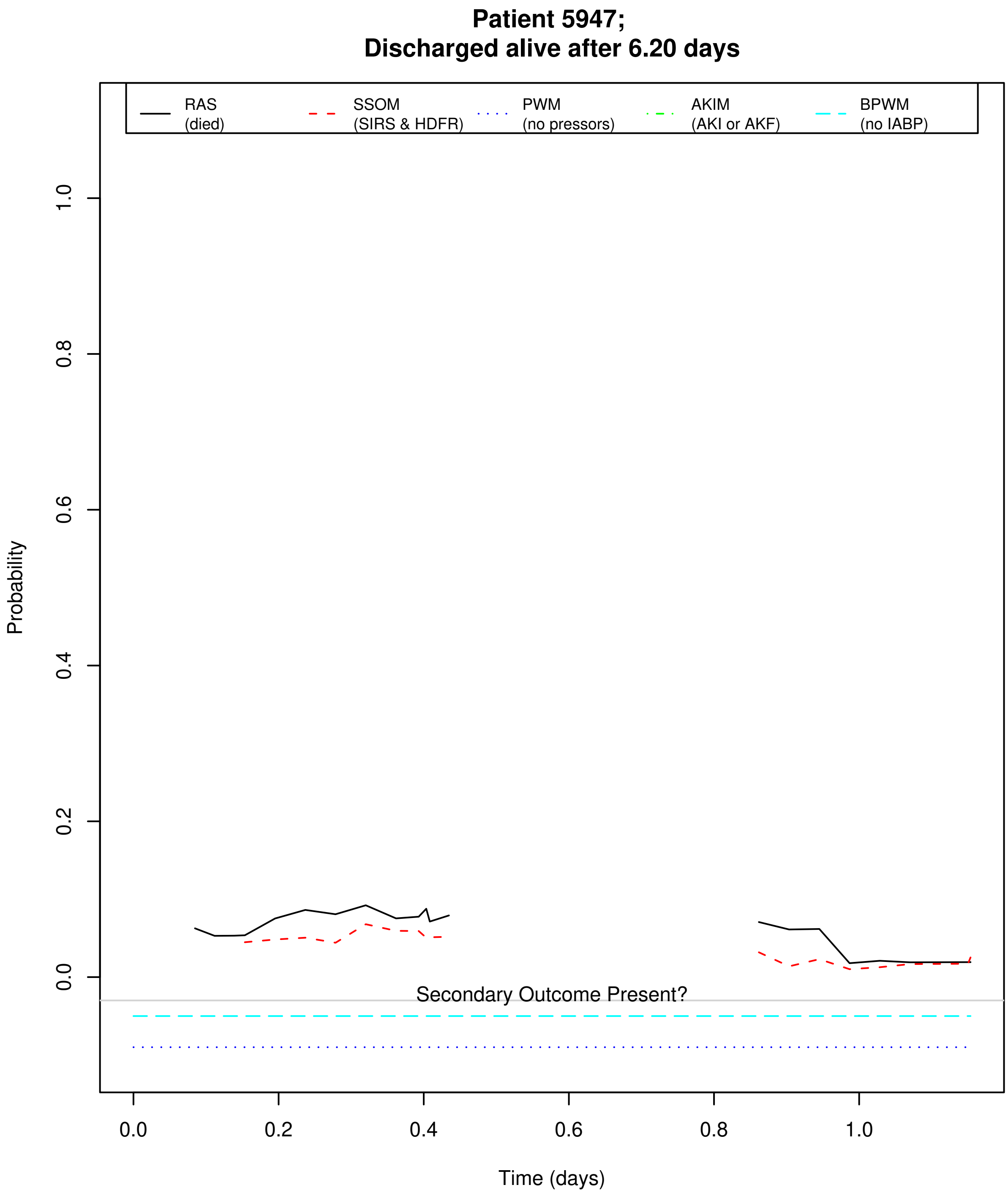


Figure E-25: Model outputs for patient 5947. The presence of secondary outcomes is marked along the x-axis.

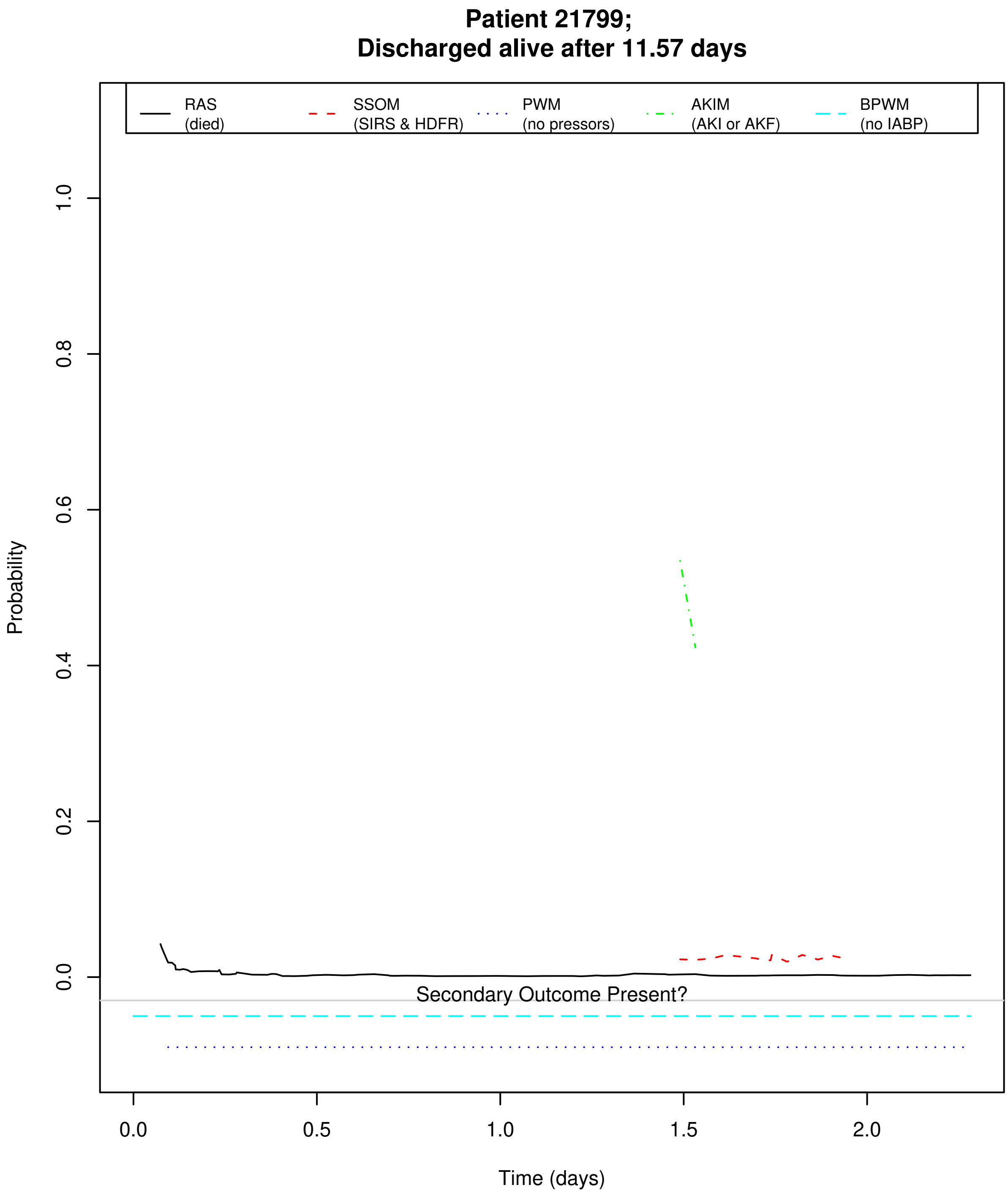


Figure E-26: Model outputs for patient 21799. The presence of secondary outcomes is marked along the x-axis.

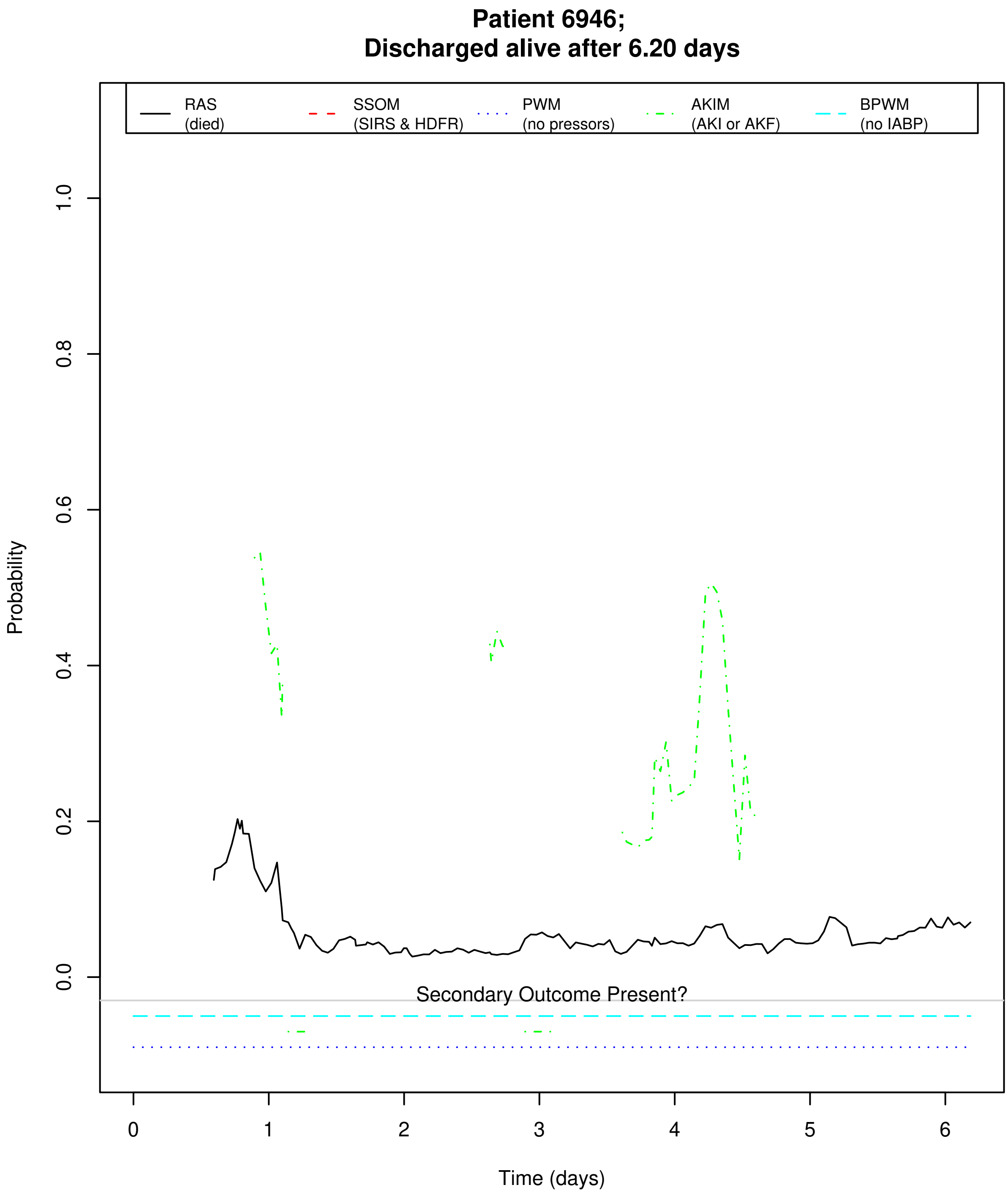


Figure E-27: Model outputs for patient 6946. The presence of secondary outcomes is marked along the x-axis.

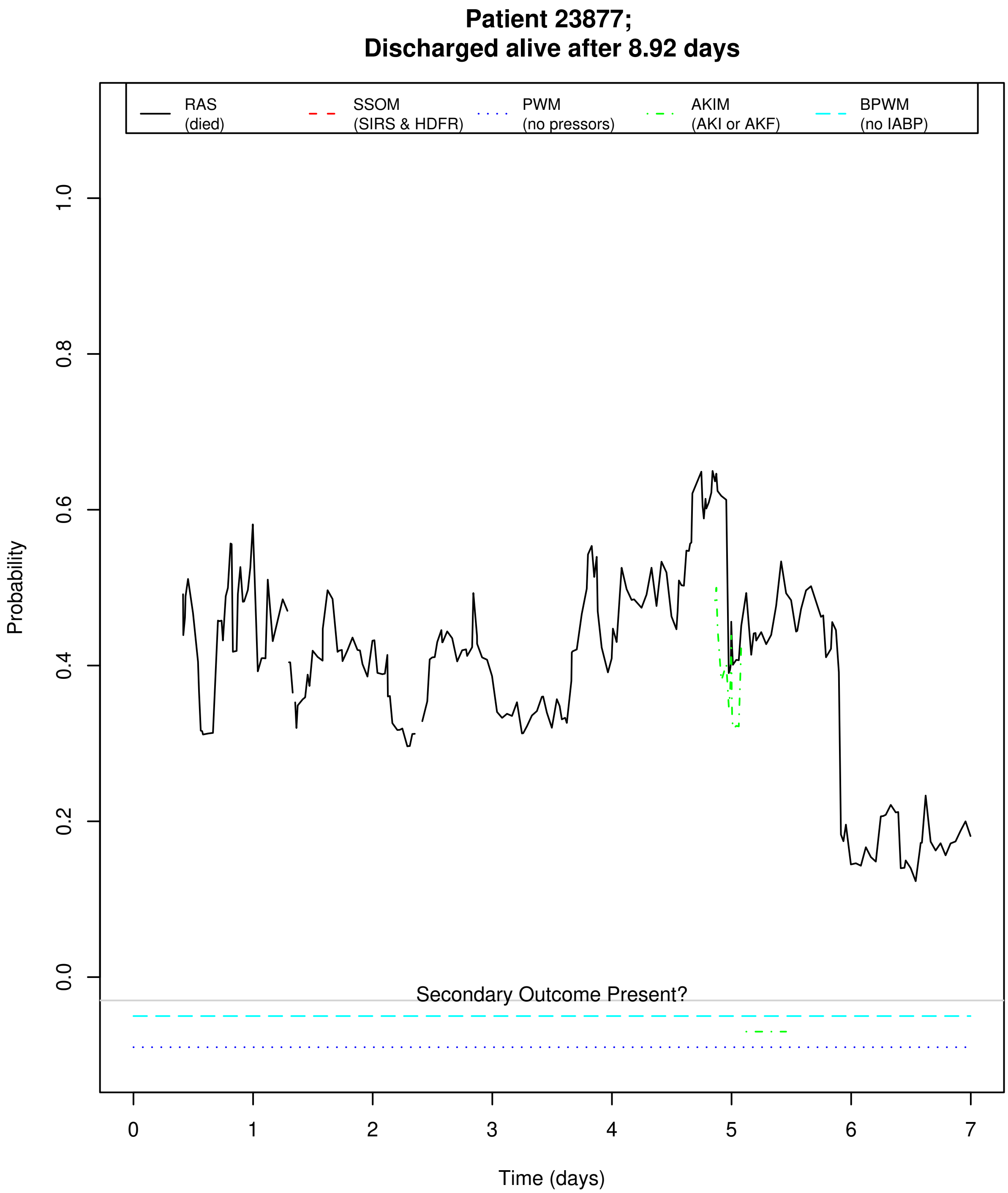


Figure E-28: Model outputs for patient 23877. The presence of secondary outcomes is marked along the x-axis.

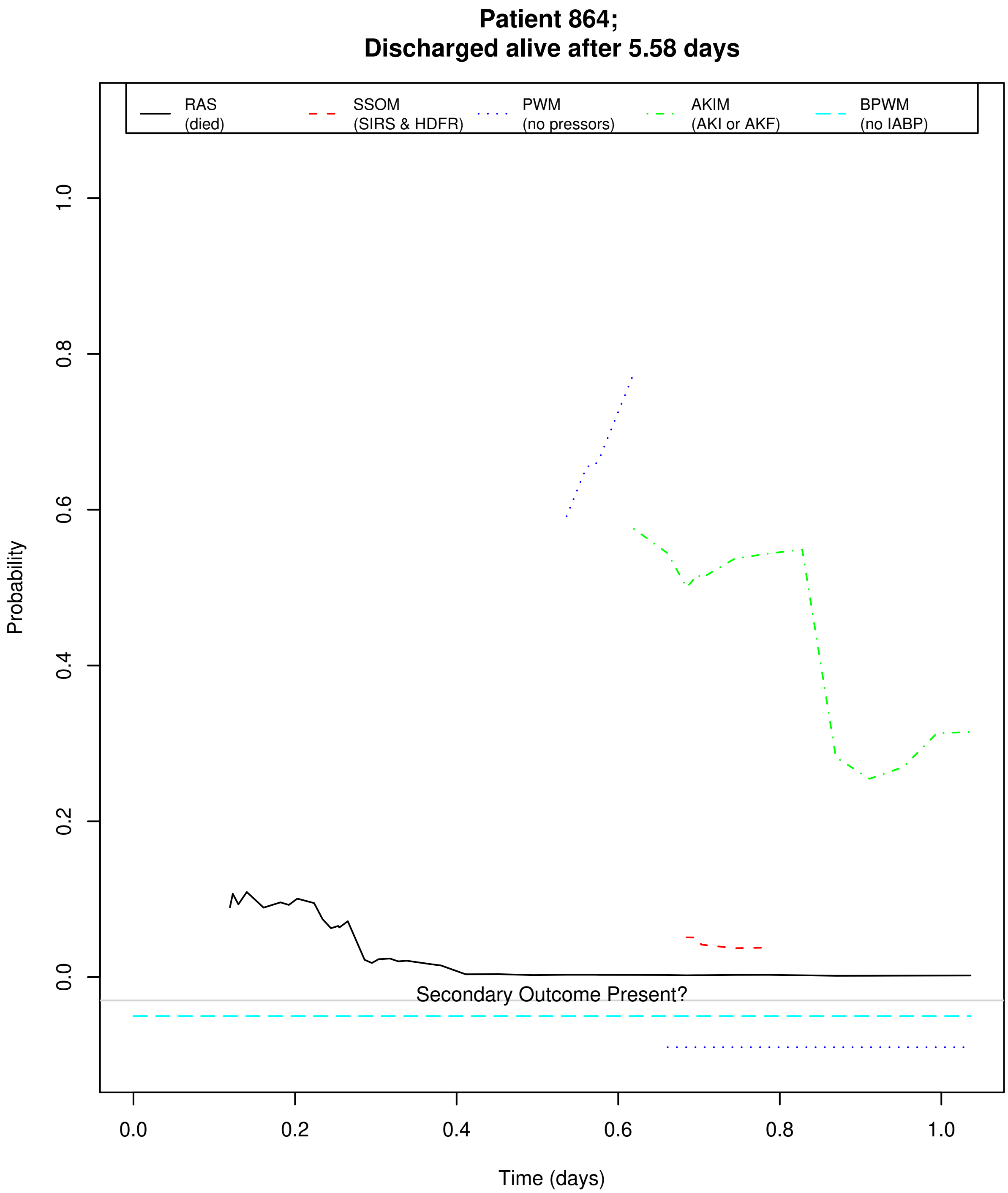


Figure E-29: Model outputs for patient 864. The presence of secondary outcomes is marked along the x-axis.

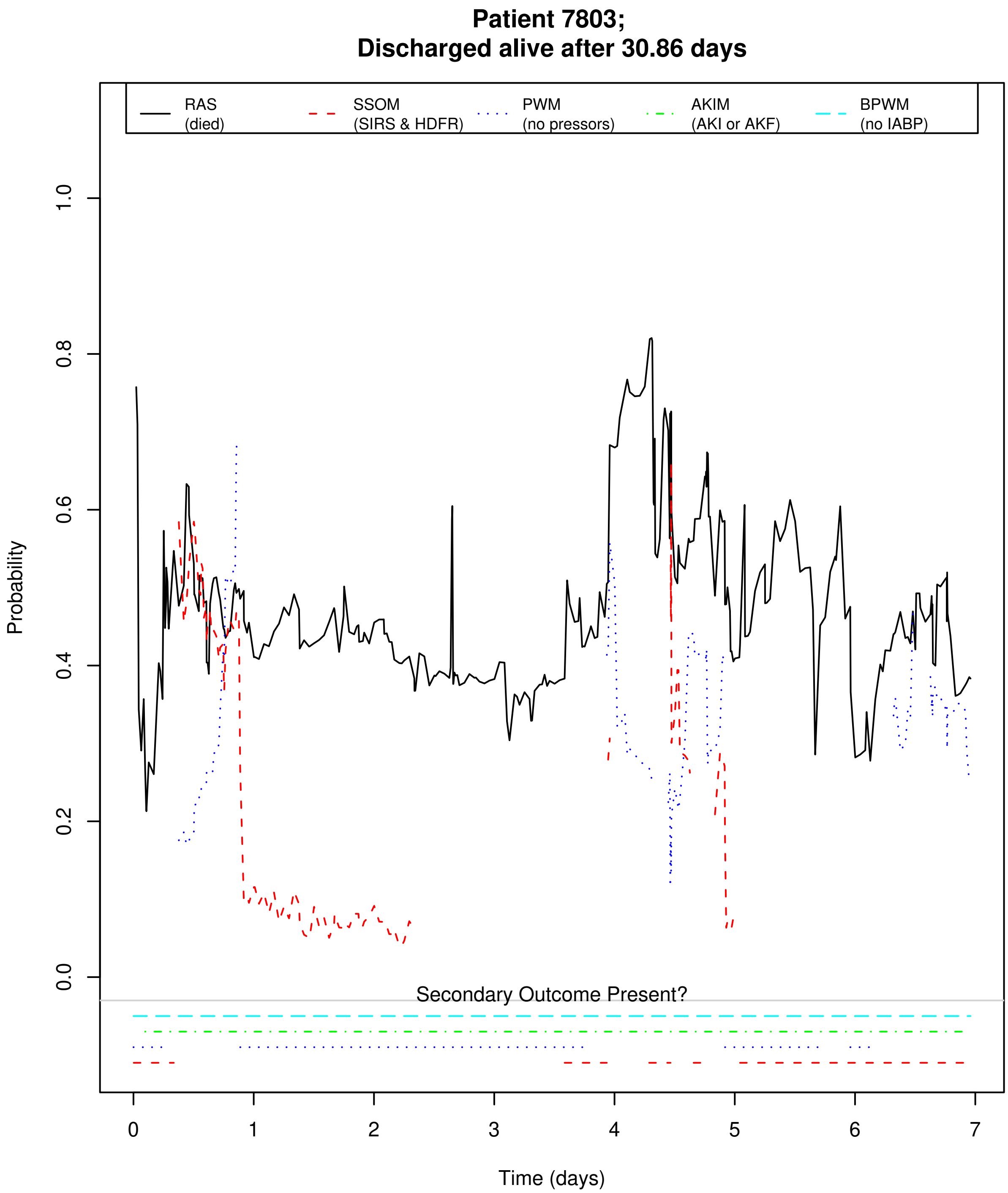


Figure E-30: Model outputs for patient 7803. The presence of secondary outcomes is marked along the x-axis.

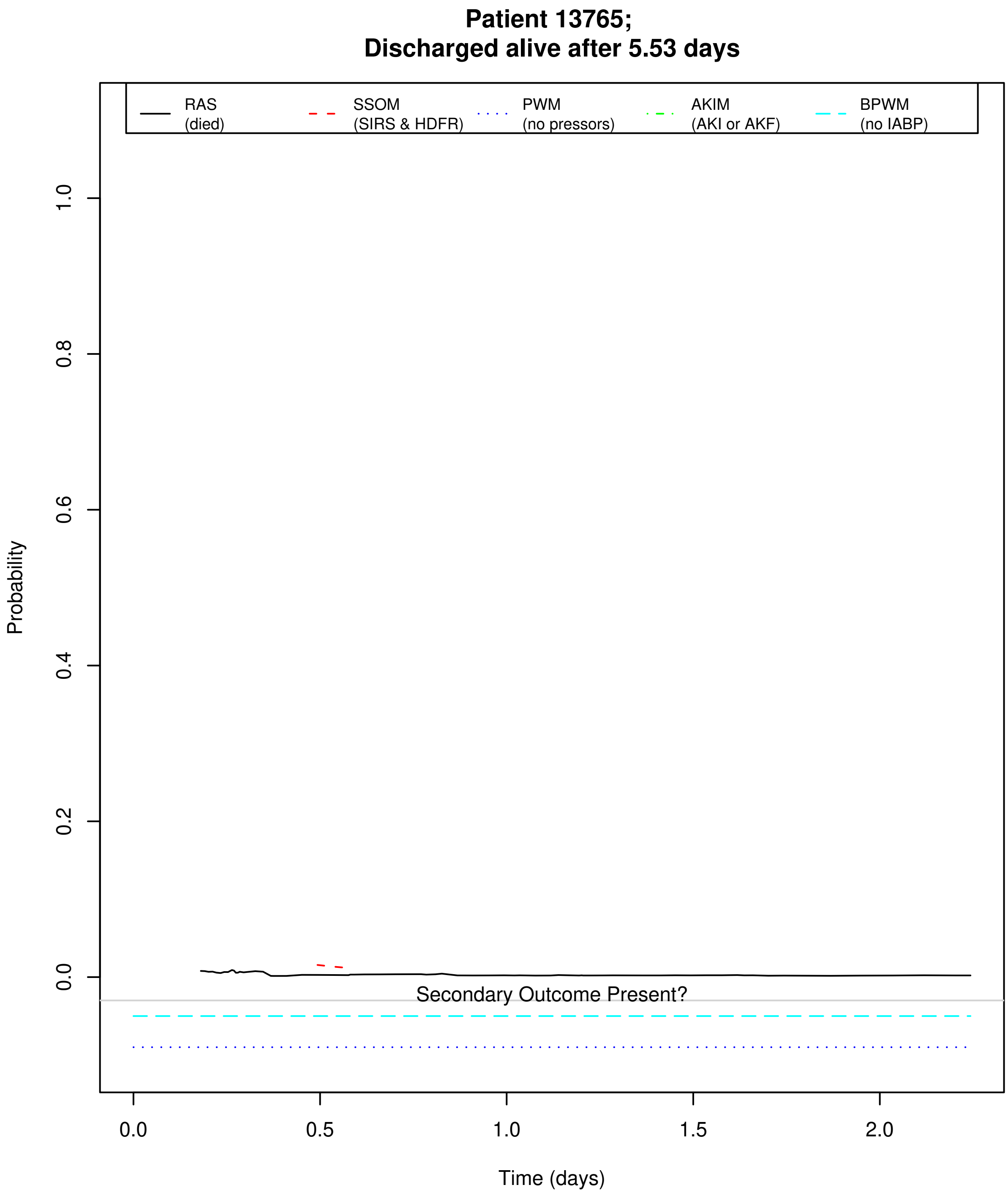


Figure E-31: Model outputs for patient 13765. The presence of secondary outcomes is marked along the x-axis.

End:

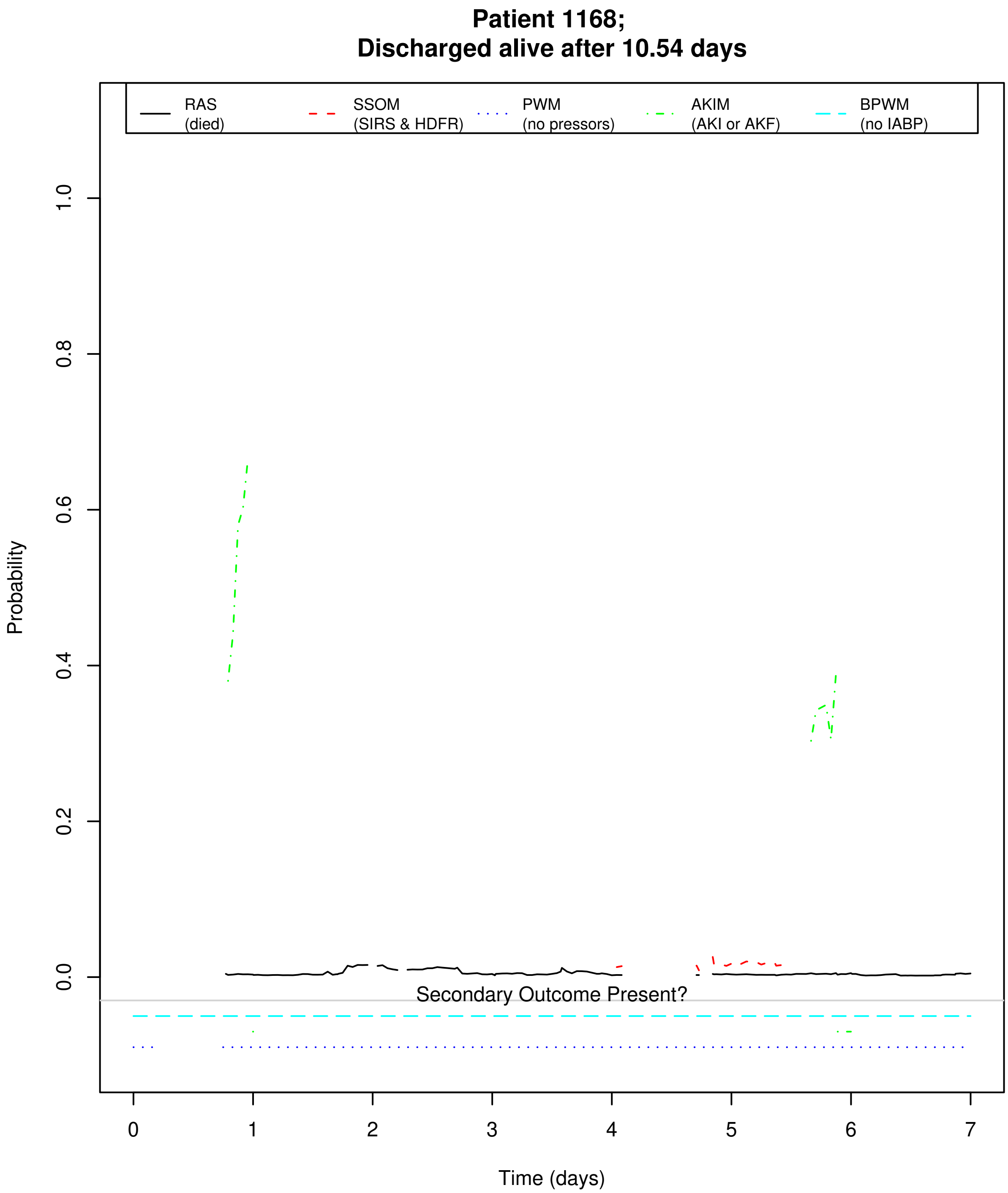


Figure E-32: Model outputs for patient 1168. The presence of secondary outcomes is marked along the x-axis.

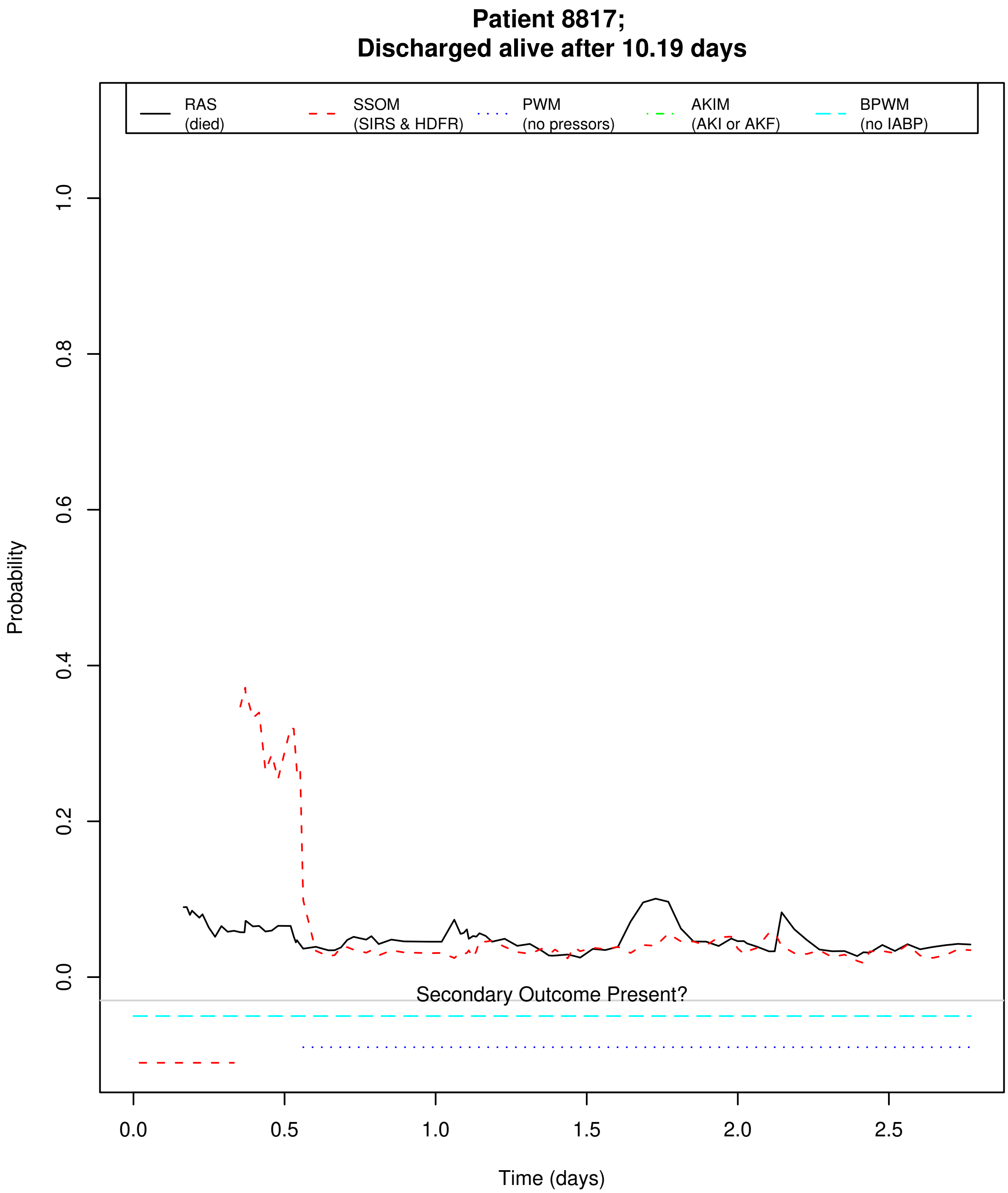


Figure E-33: Model outputs for patient 8817. The presence of secondary outcomes is marked along the x-axis.

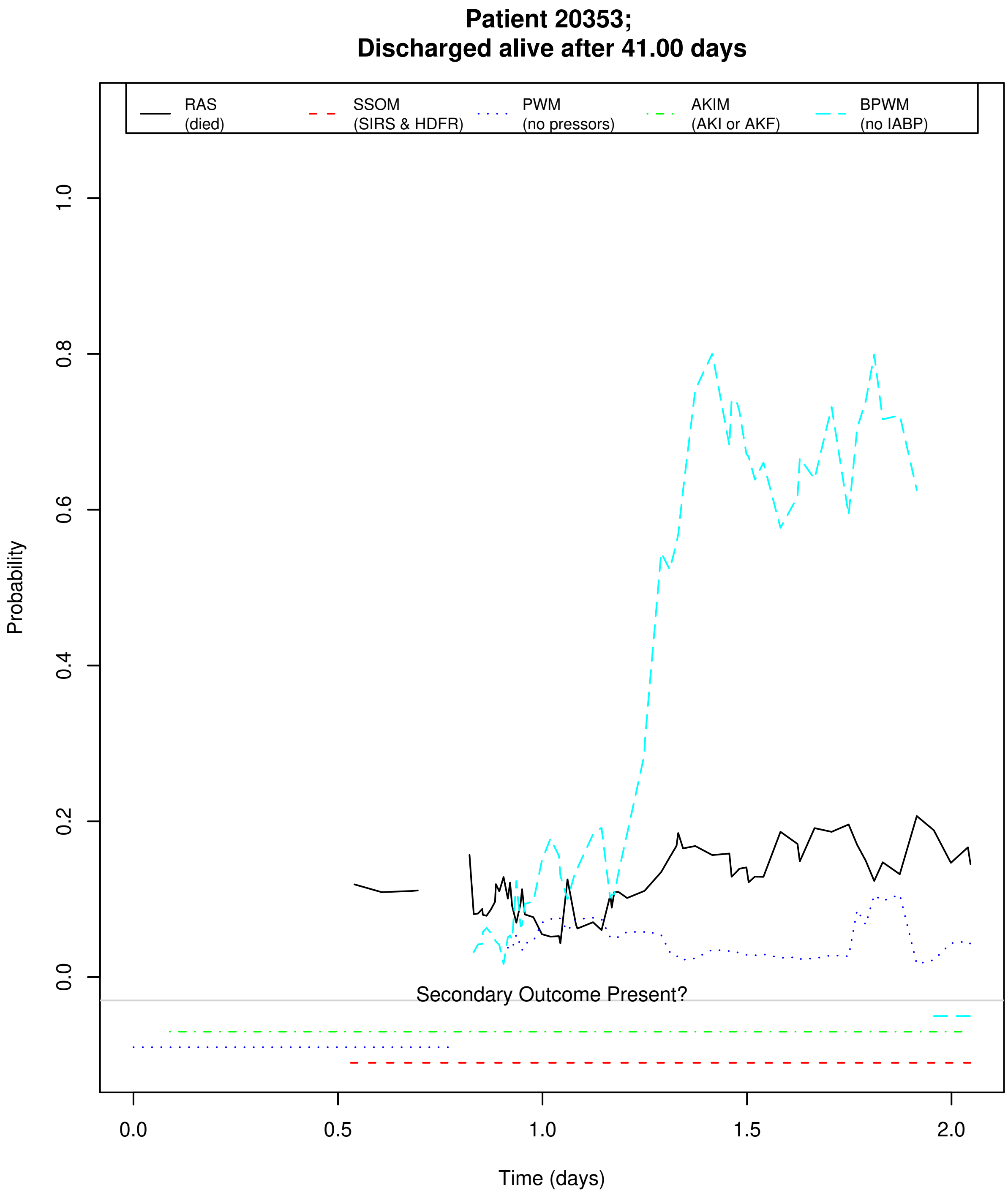


Figure E-34: Model outputs for patient 20353. The presence of secondary outcomes is marked along the x-axis.

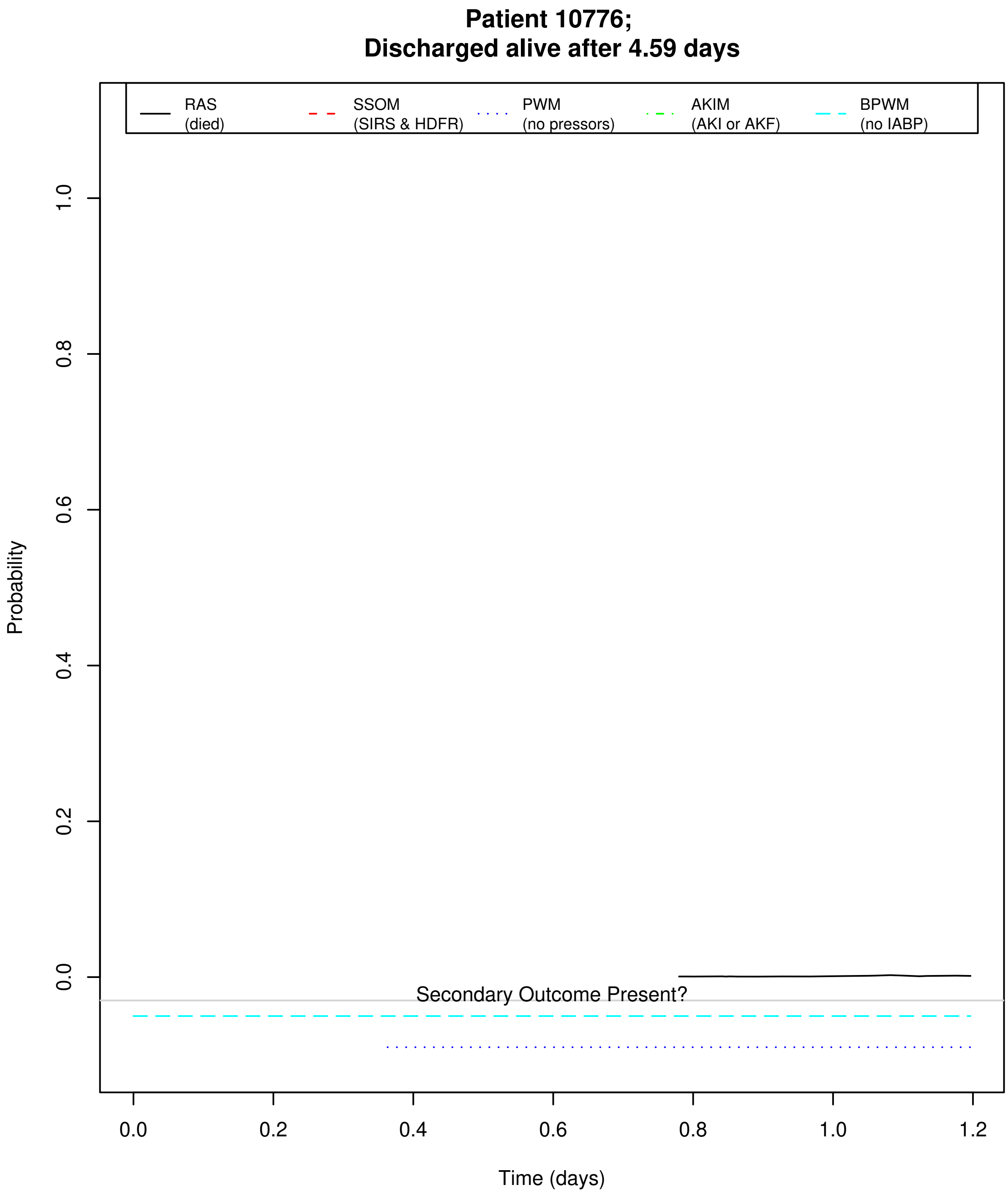


Figure E-35: Model outputs for patient 10776. The presence of secondary outcomes is marked along the x-axis.

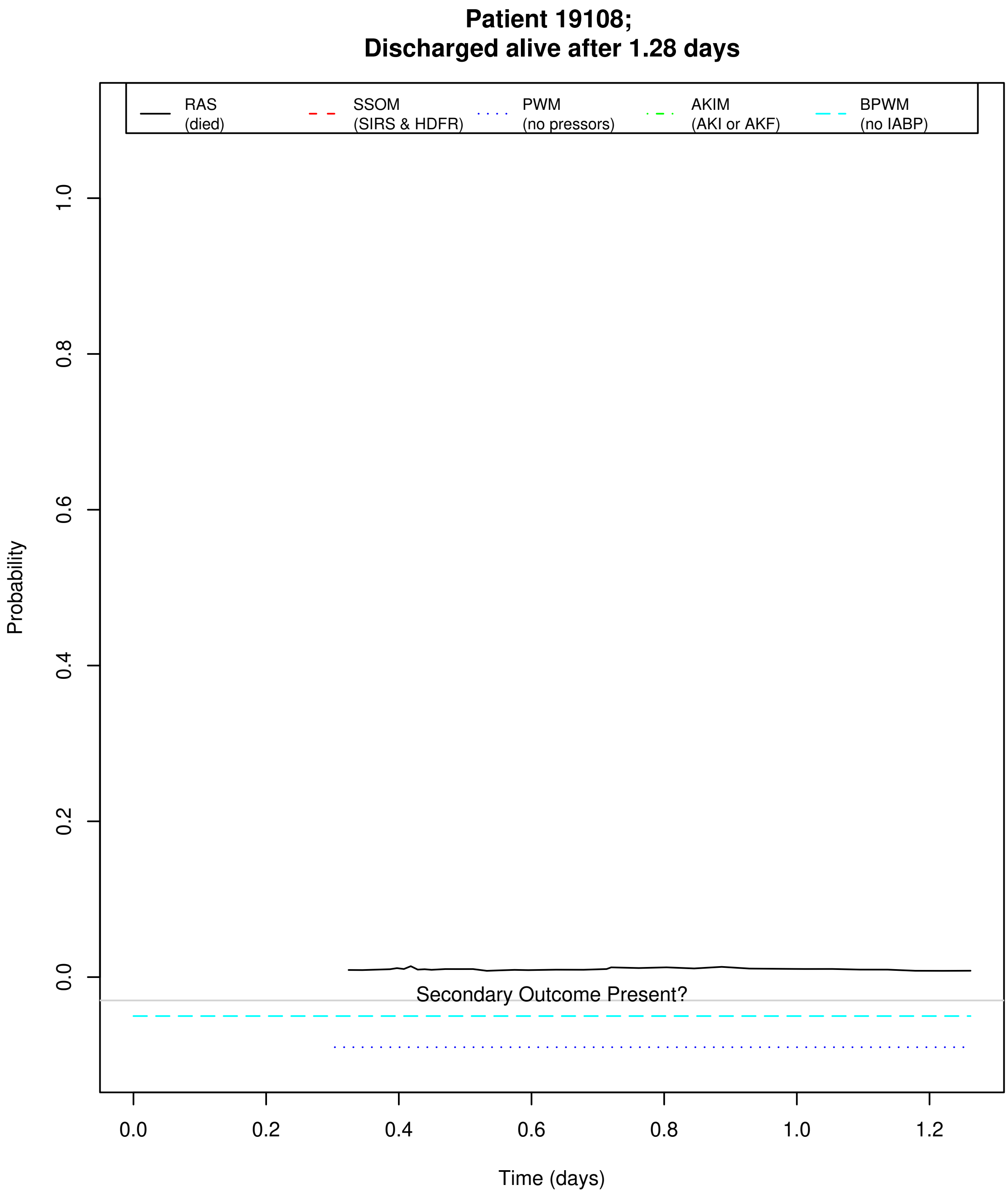


Figure E-36: Model outputs for patient 19108. The presence of secondary outcomes is marked along the x-axis.

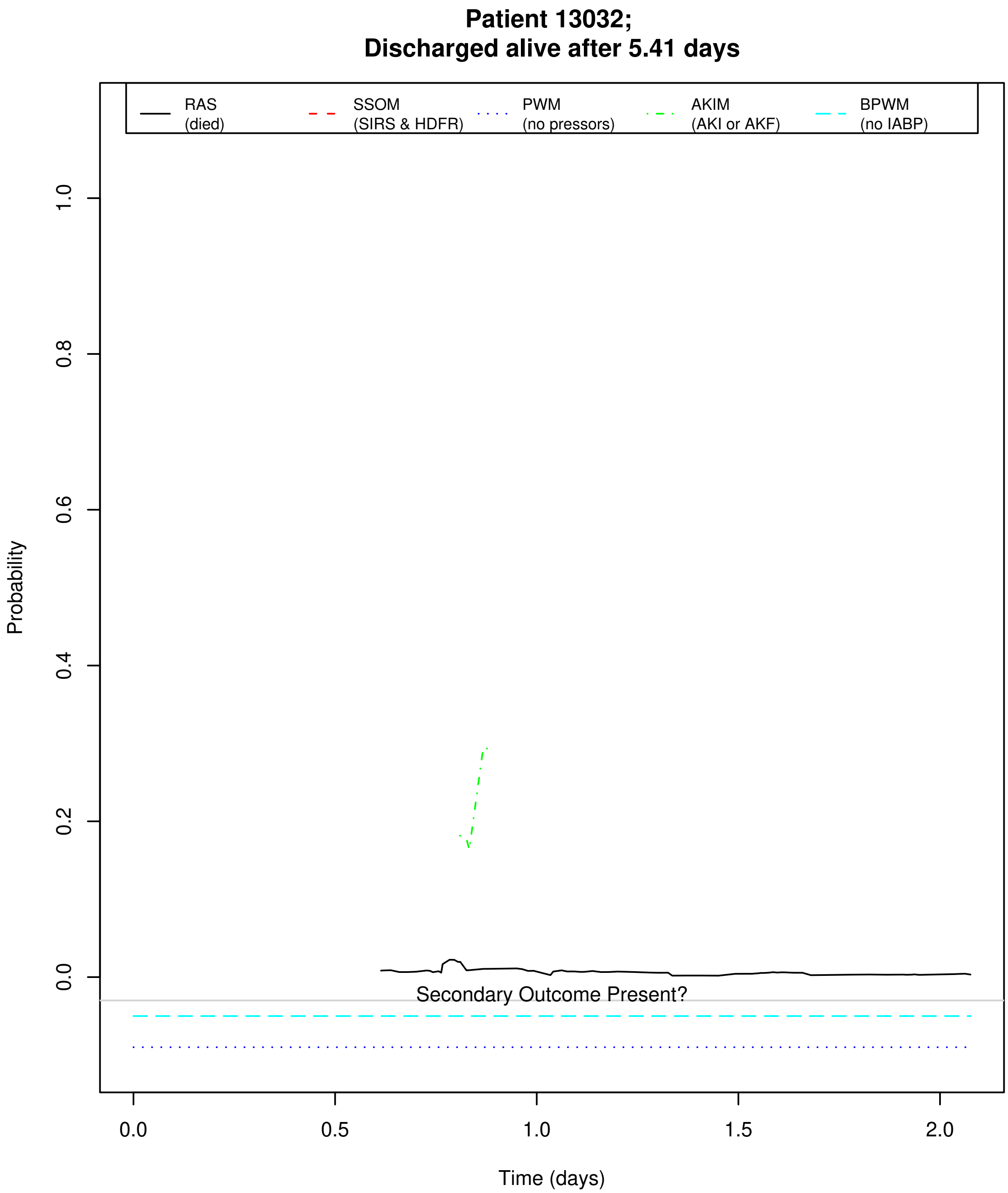


Figure E-37: Model outputs for patient 13032. The presence of secondary outcomes is marked along the x-axis.

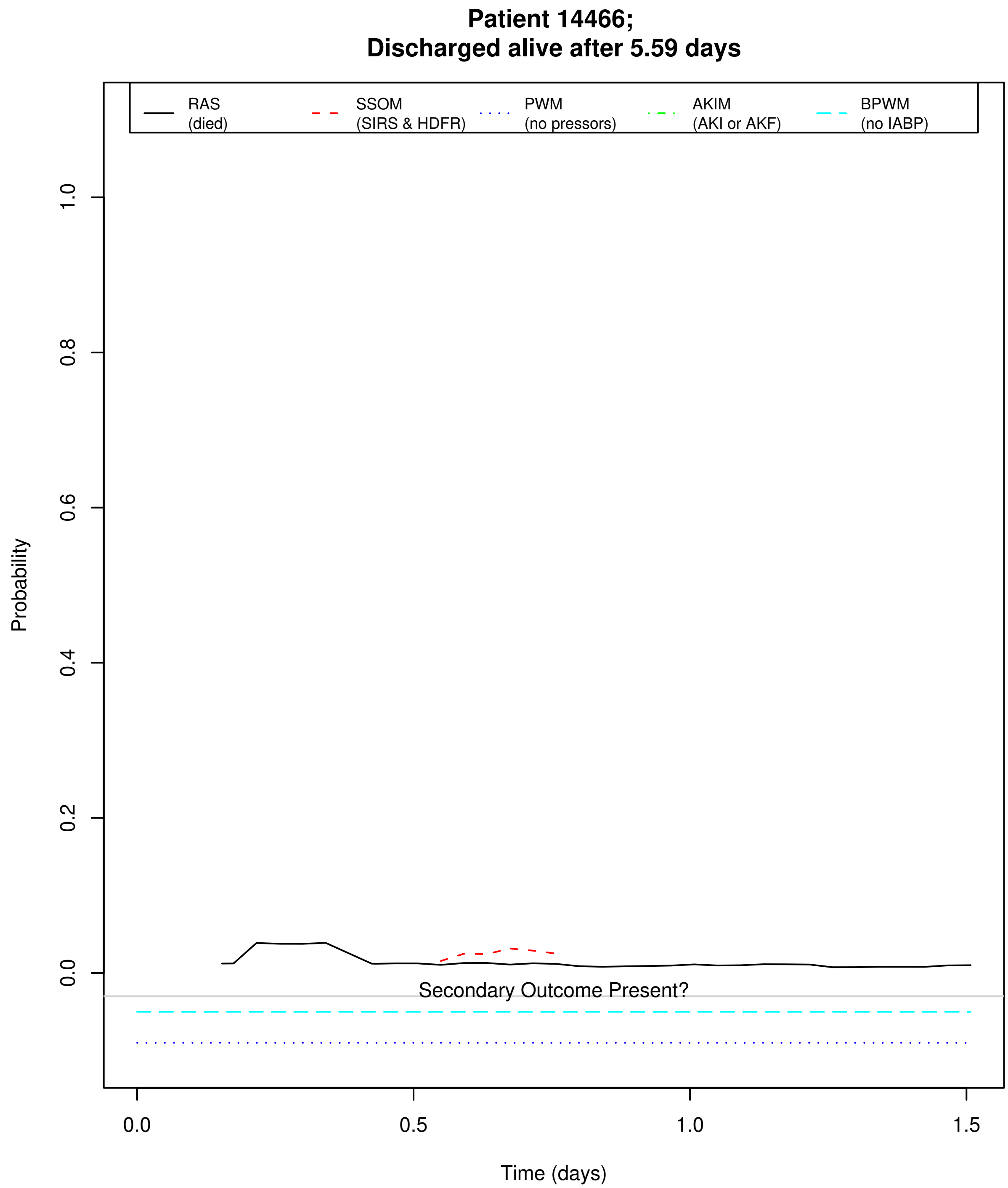


Figure E-38: Model outputs for patient 14466. The presence of secondary outcomes is marked along the x-axis.

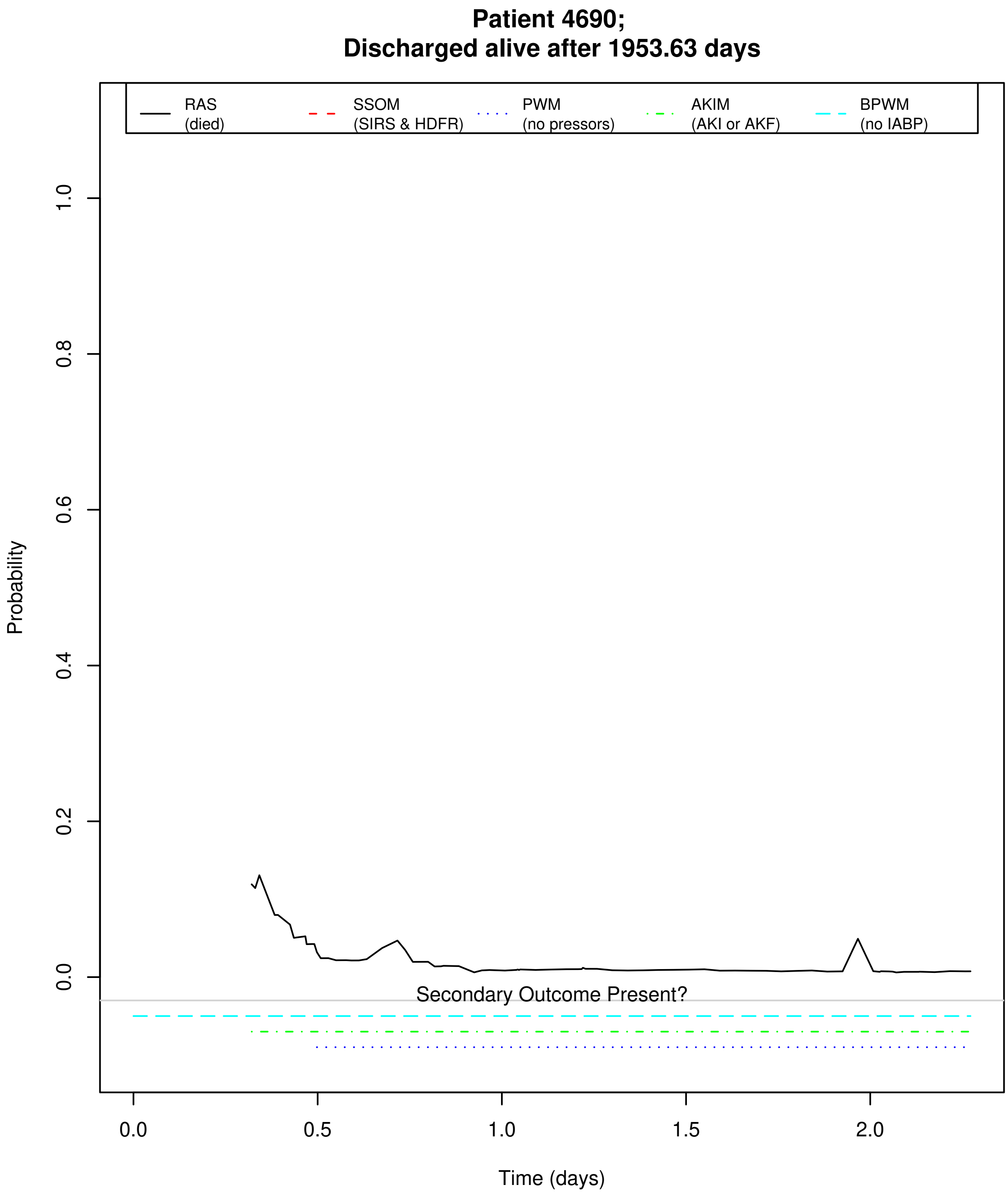


Figure E-39: Model outputs for patient 4690. The presence of secondary outcomes is marked along the x-axis.

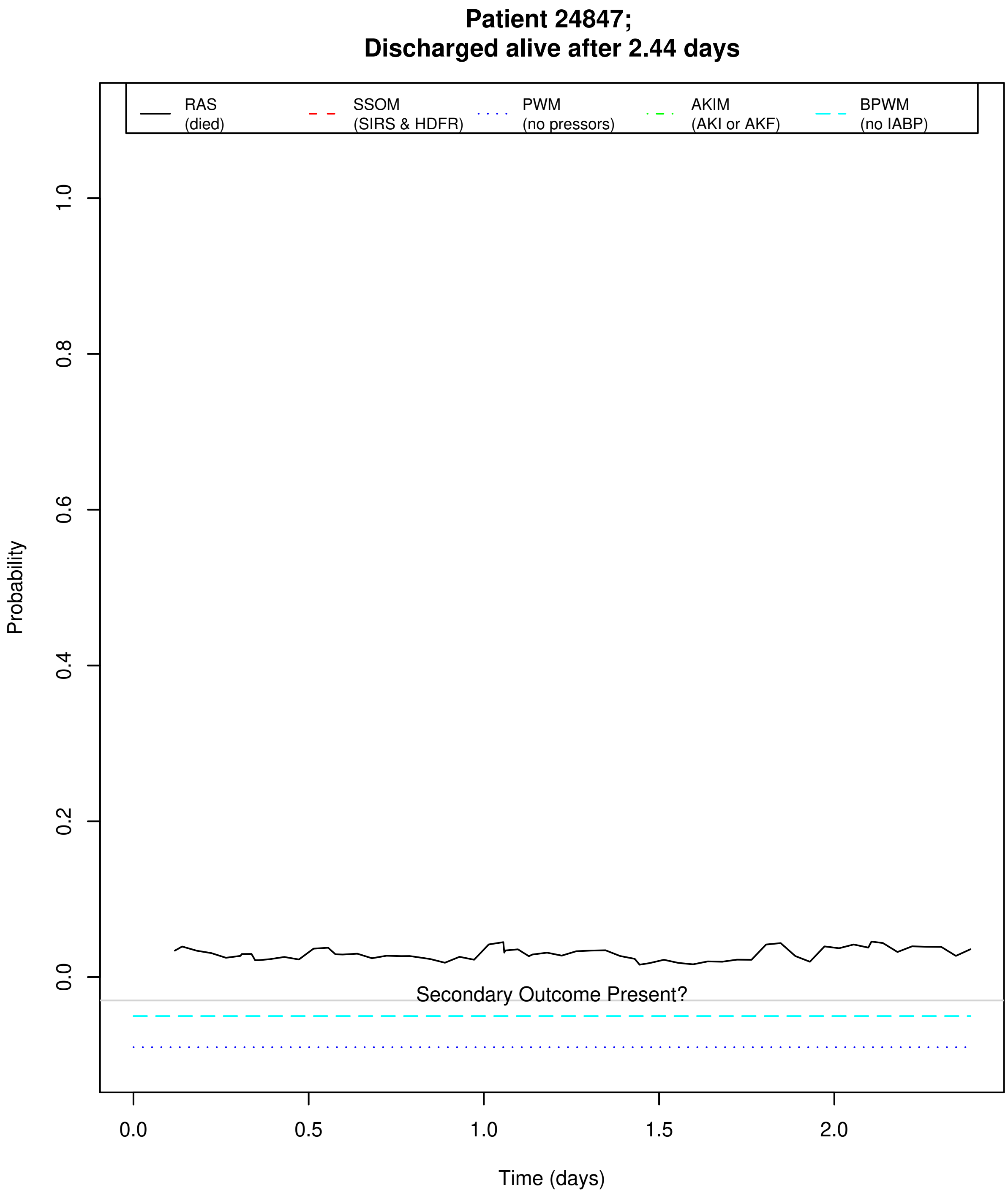


Figure E-40: Model outputs for patient 24847. The presence of secondary outcomes is marked along the x-axis.

Appendix F

Secondary Outcome Model Selection

This appendix provides cross validation performance plots that detail the feature selection process used for each secondary outcome model discussed in Chapter 6. Plots showing the cross validation performance on the final set of features are also provided. For plots that show less than 50 covariates, the individual points are shown on the plots. When a plot covers more than 50 covariates, I only show the line that connects the points.

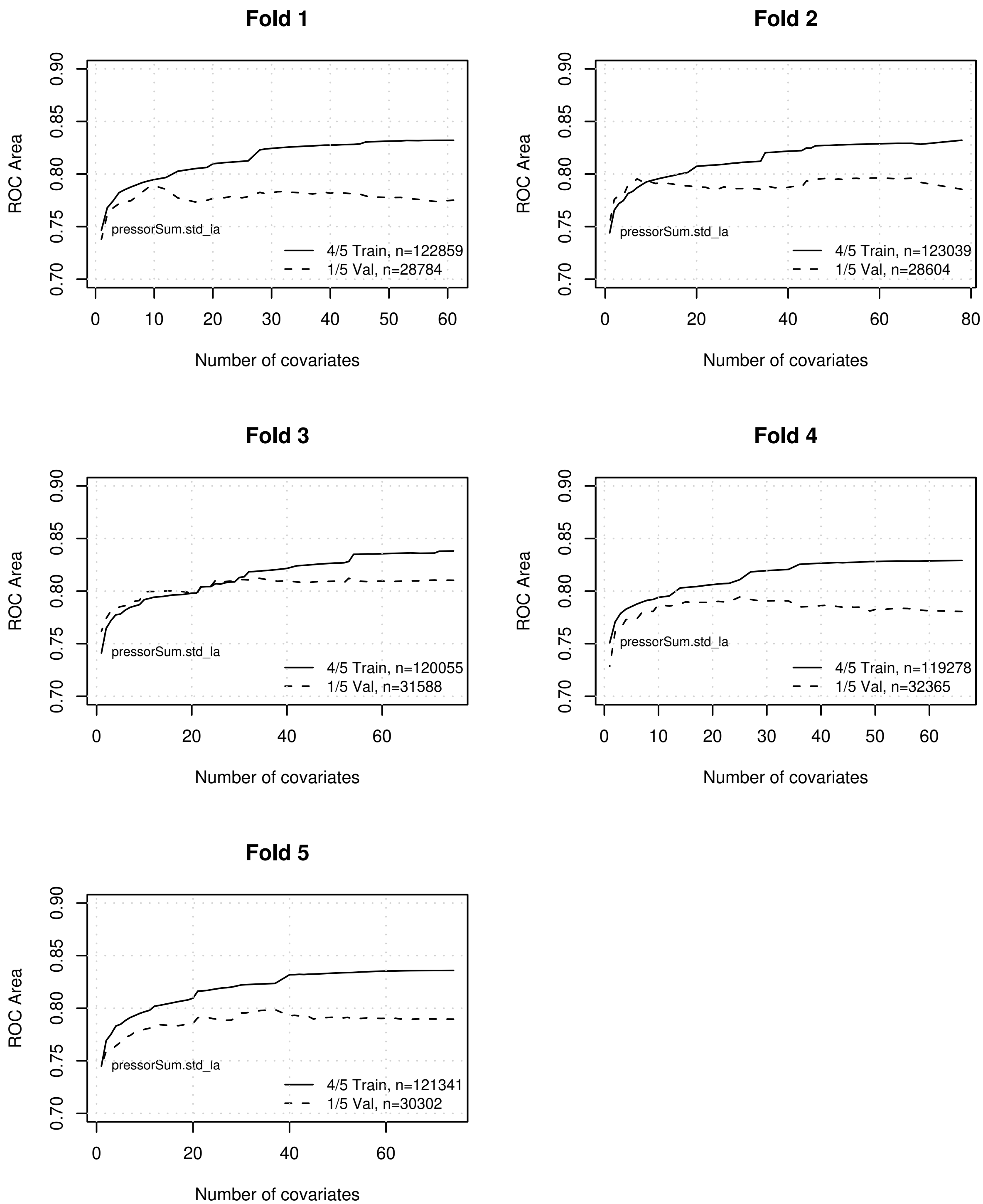


Figure F-1: `PWM` model selection, sensitivity to number of covariates on each cross validation fold (development data)

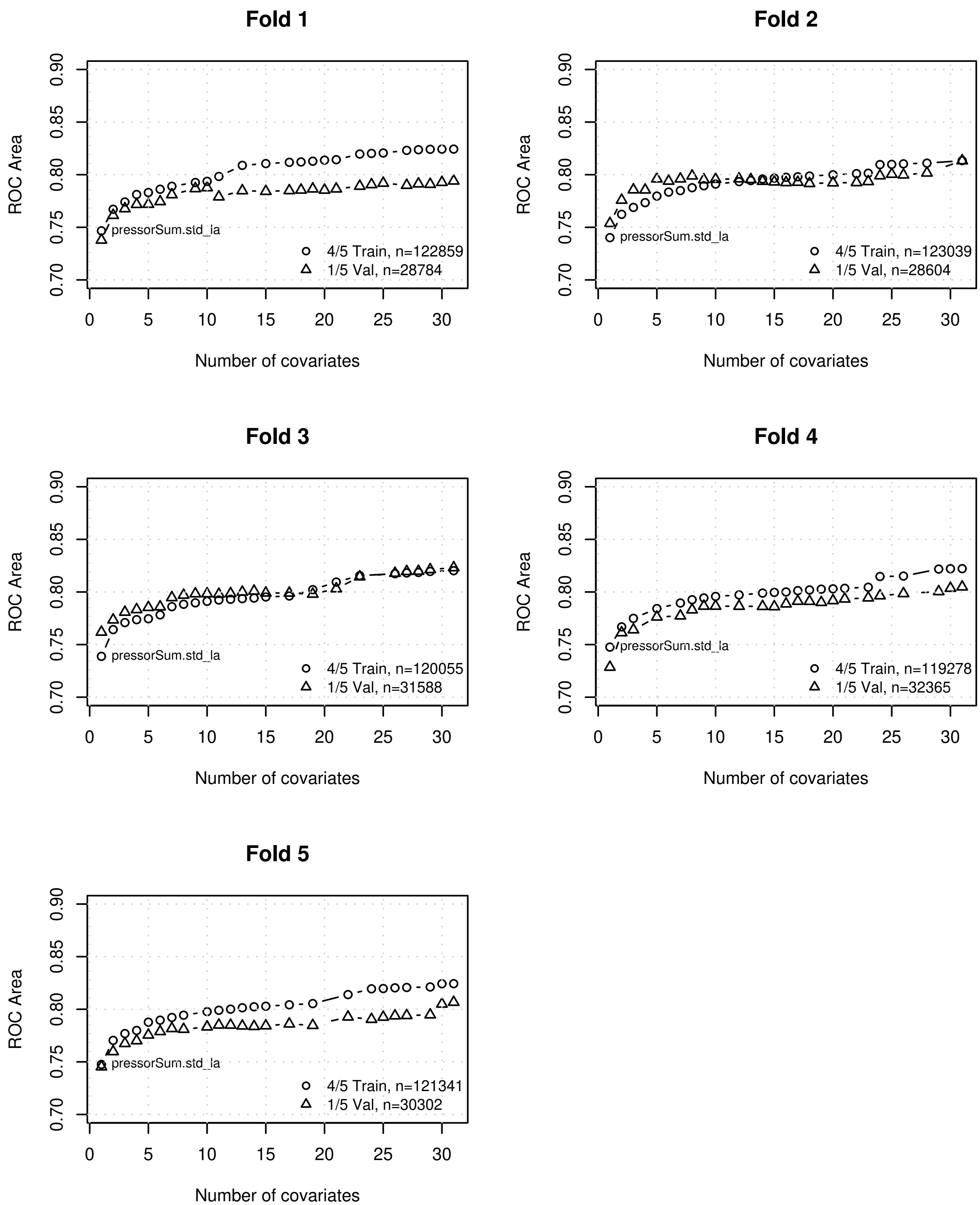


Figure F-2: `PWM` final feature set performance on cross validation folds

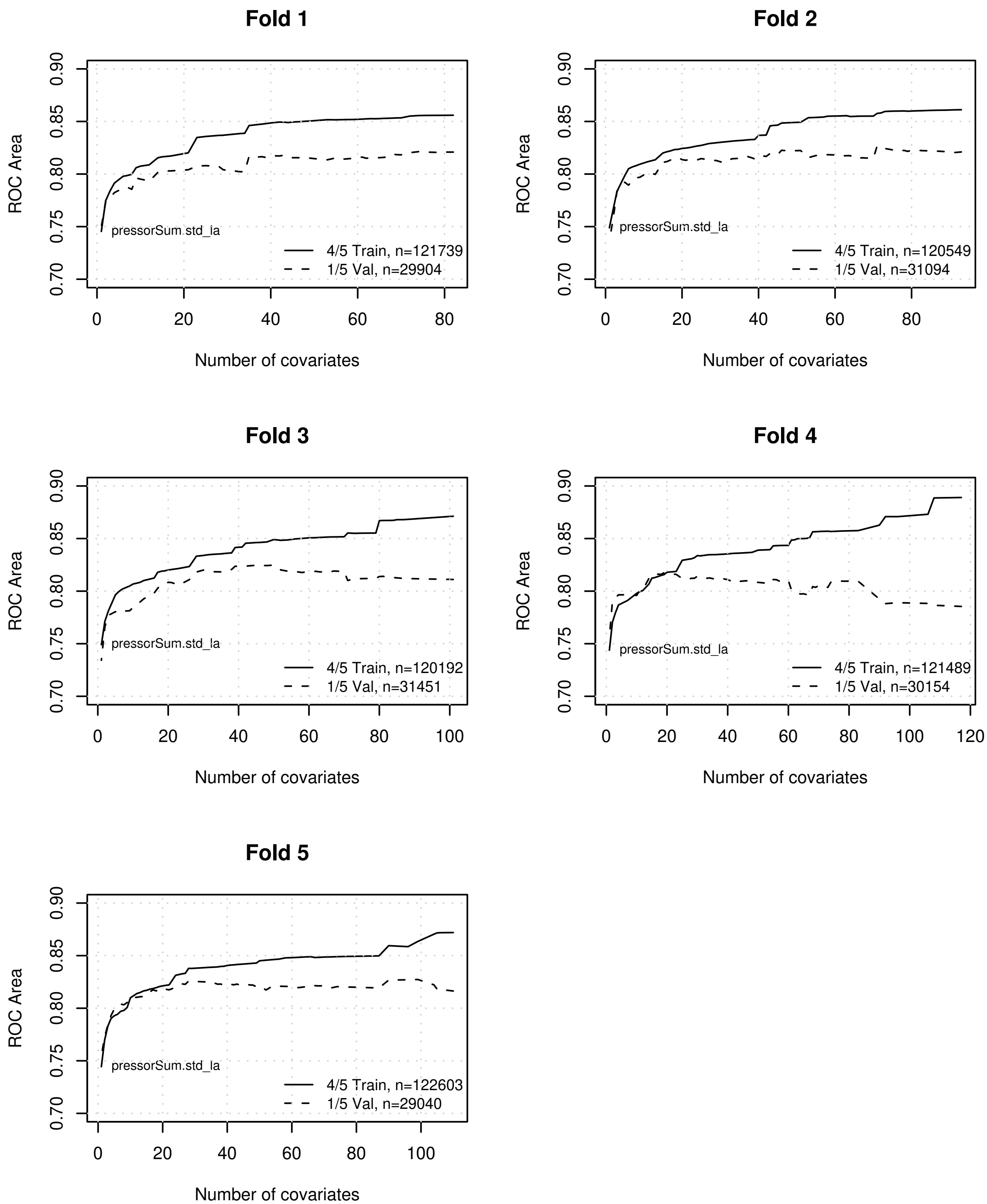


Figure F-3: `PWLM` model selection, sensitivity to number of covariates on each cross validation fold (development data)

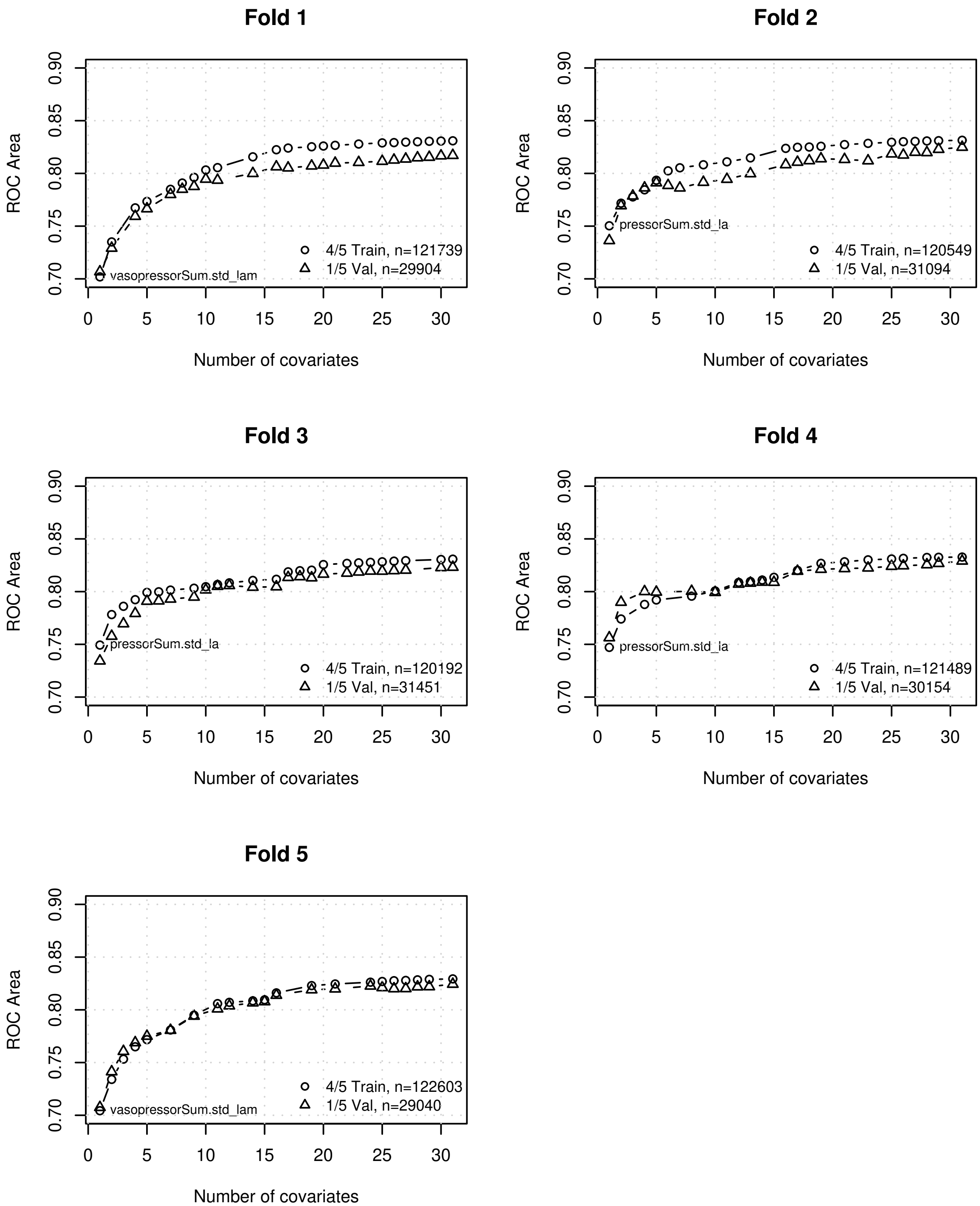


Figure F-4: `PWLM` final feature set performance on cross validation folds

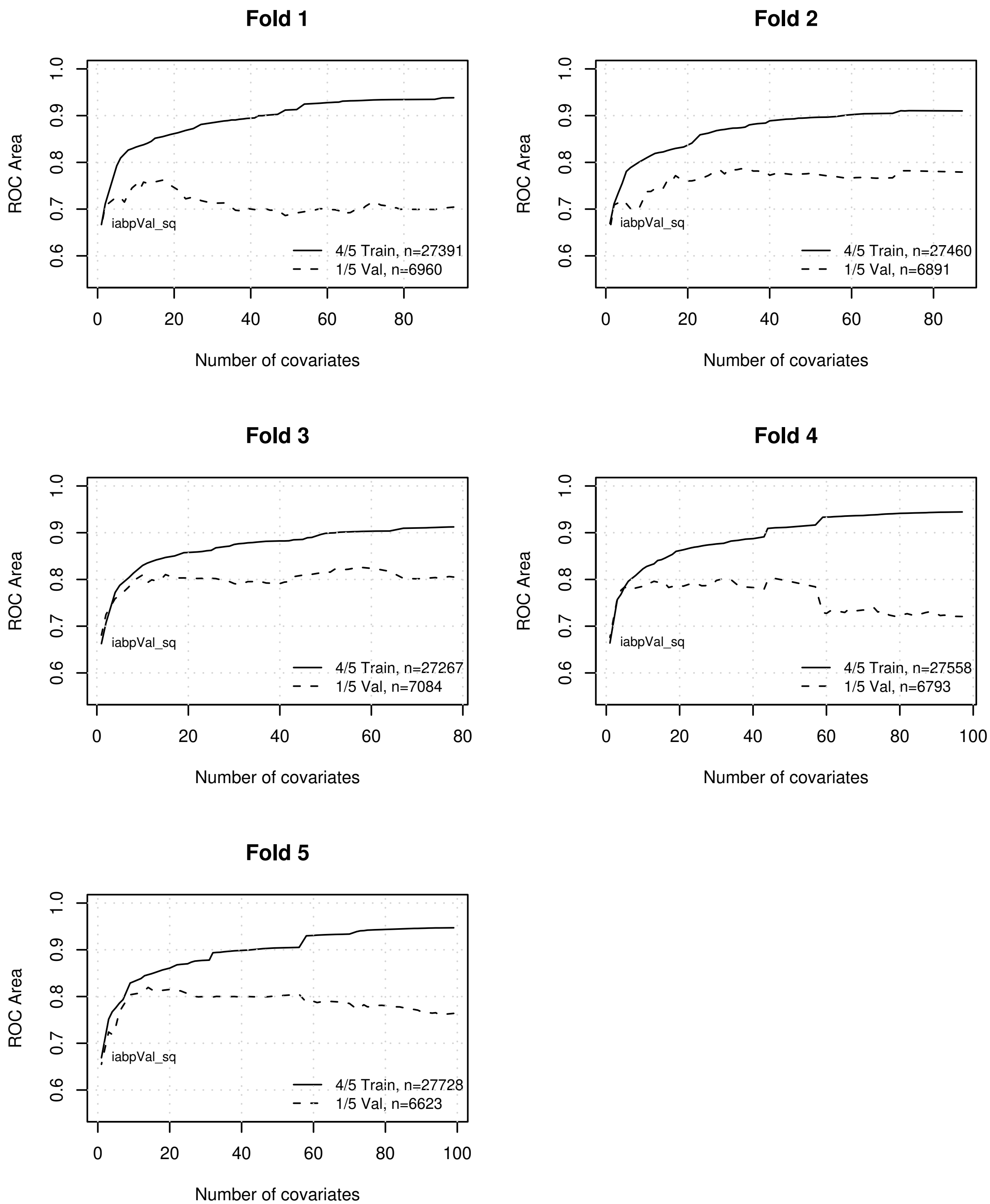


Figure F-5: `BPWM` model selection, sensitivity to number of covariates on each cross validation fold (development data)

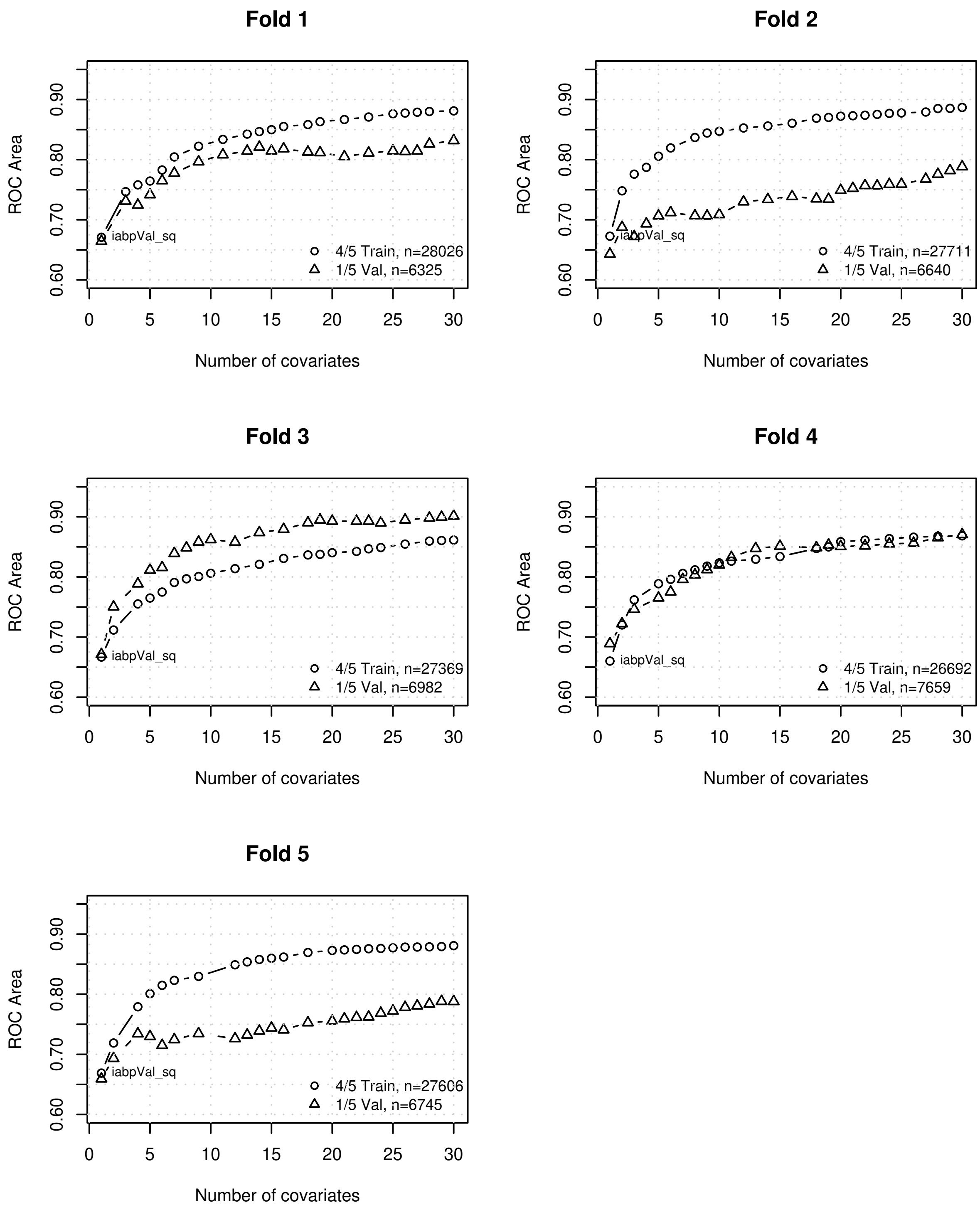


Figure F-6: BPWM final feature set performance on cross validation folds

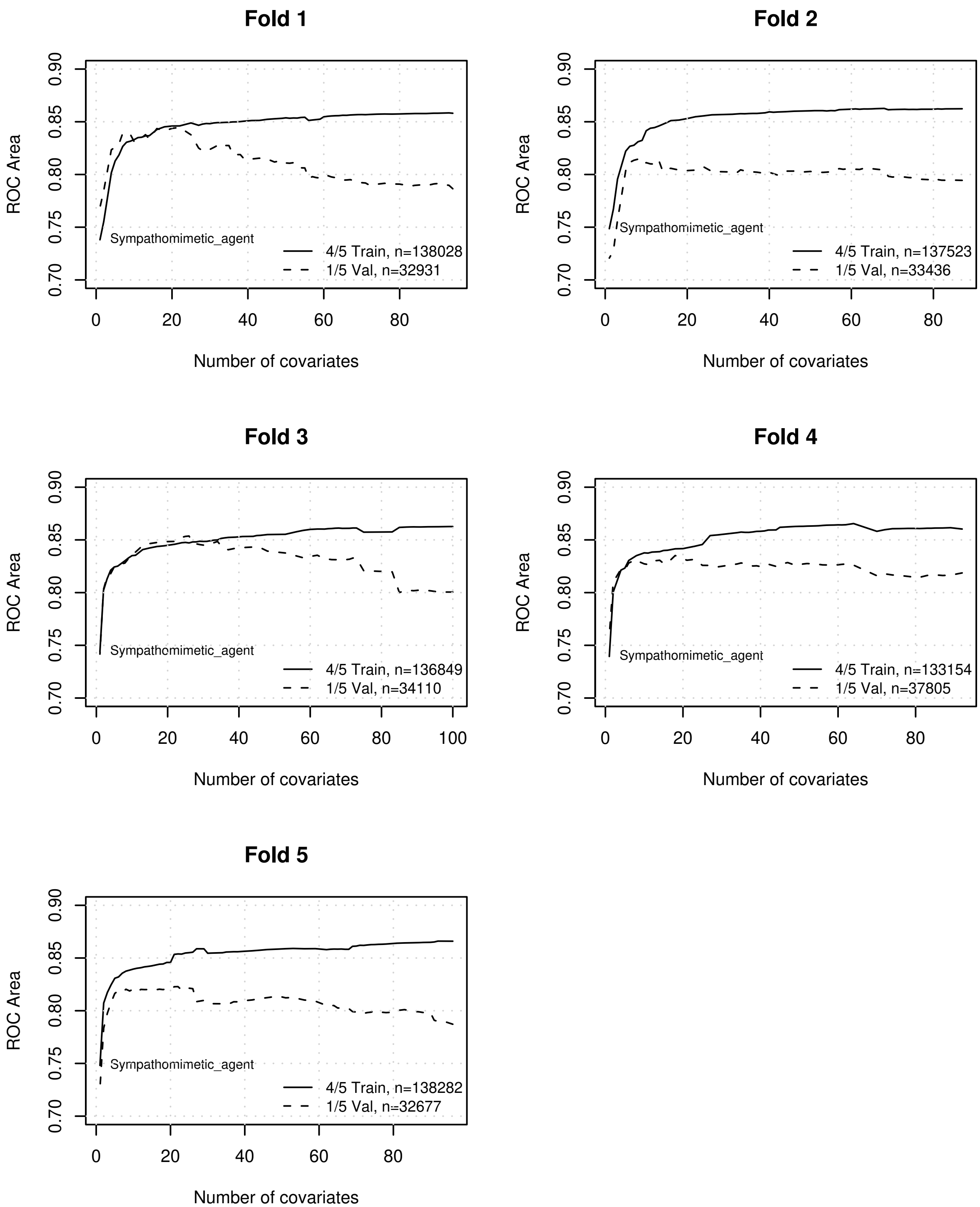


Figure F-7: `SSOM` model selection, sensitivity to number of covariates on each cross validation fold (development data)

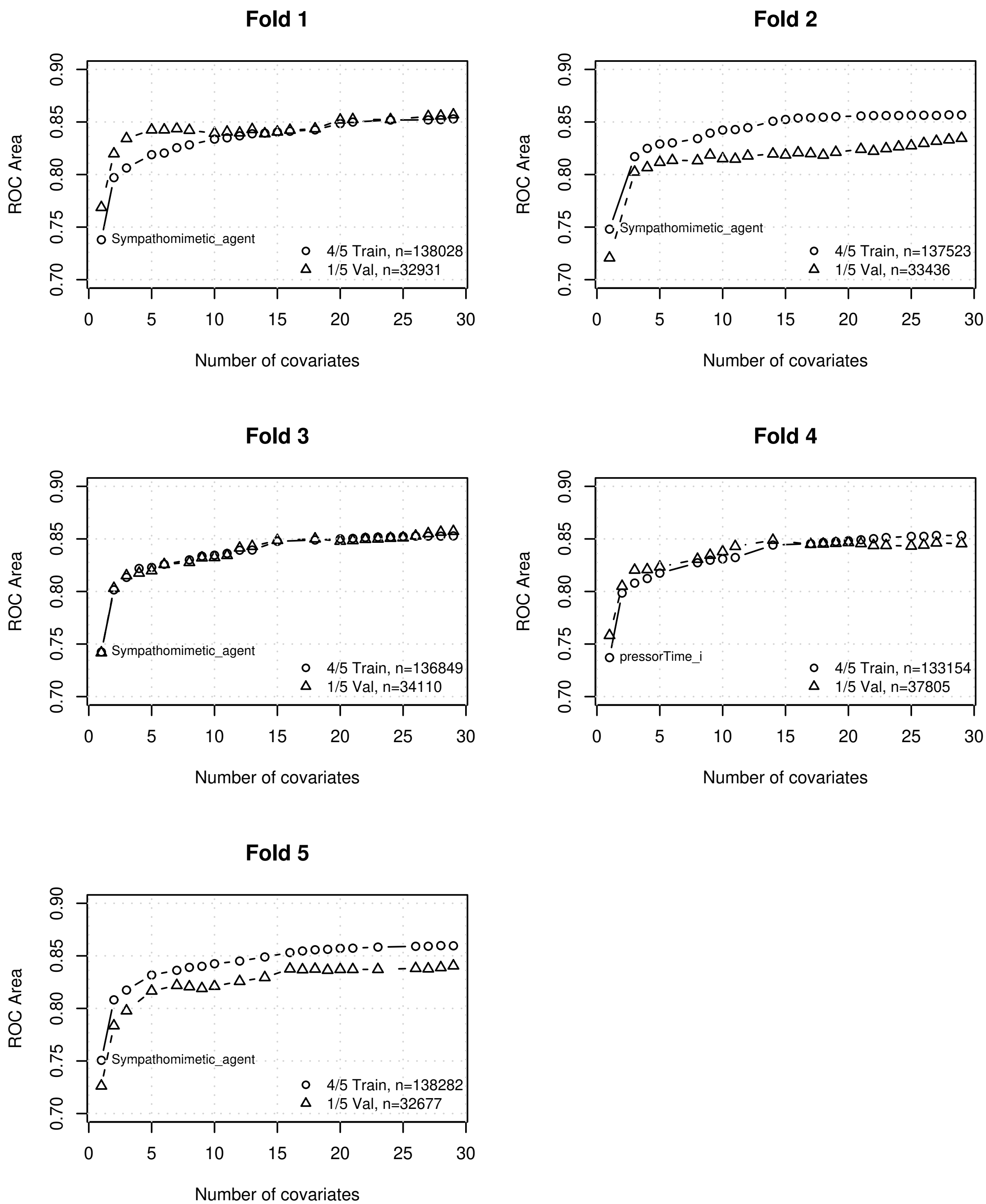


Figure F-8: `SSOM` final feature set performance on cross validation folds

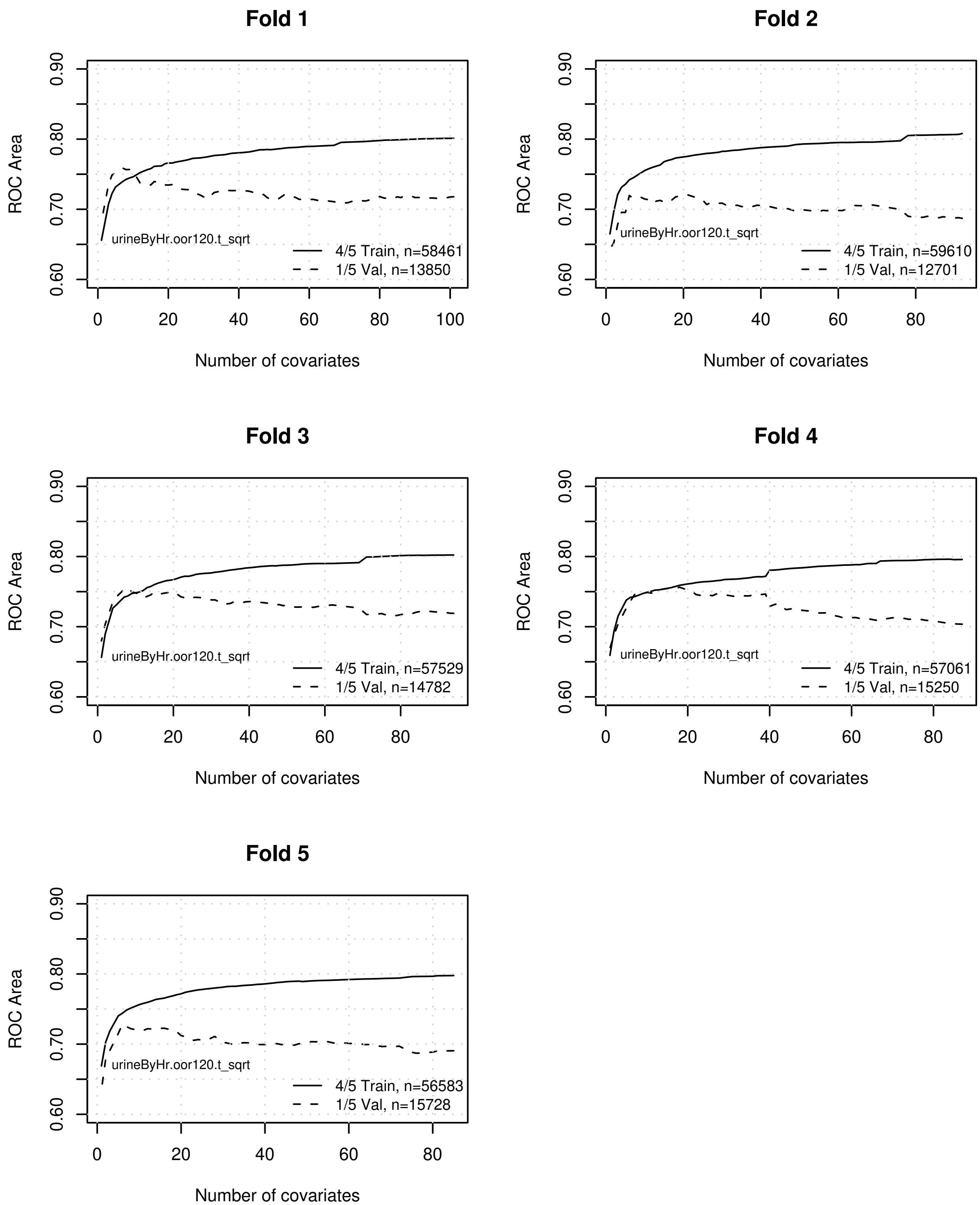


Figure F-9: `AKIM` model selection, sensitivity to number of covariates on each cross validation fold (development data)

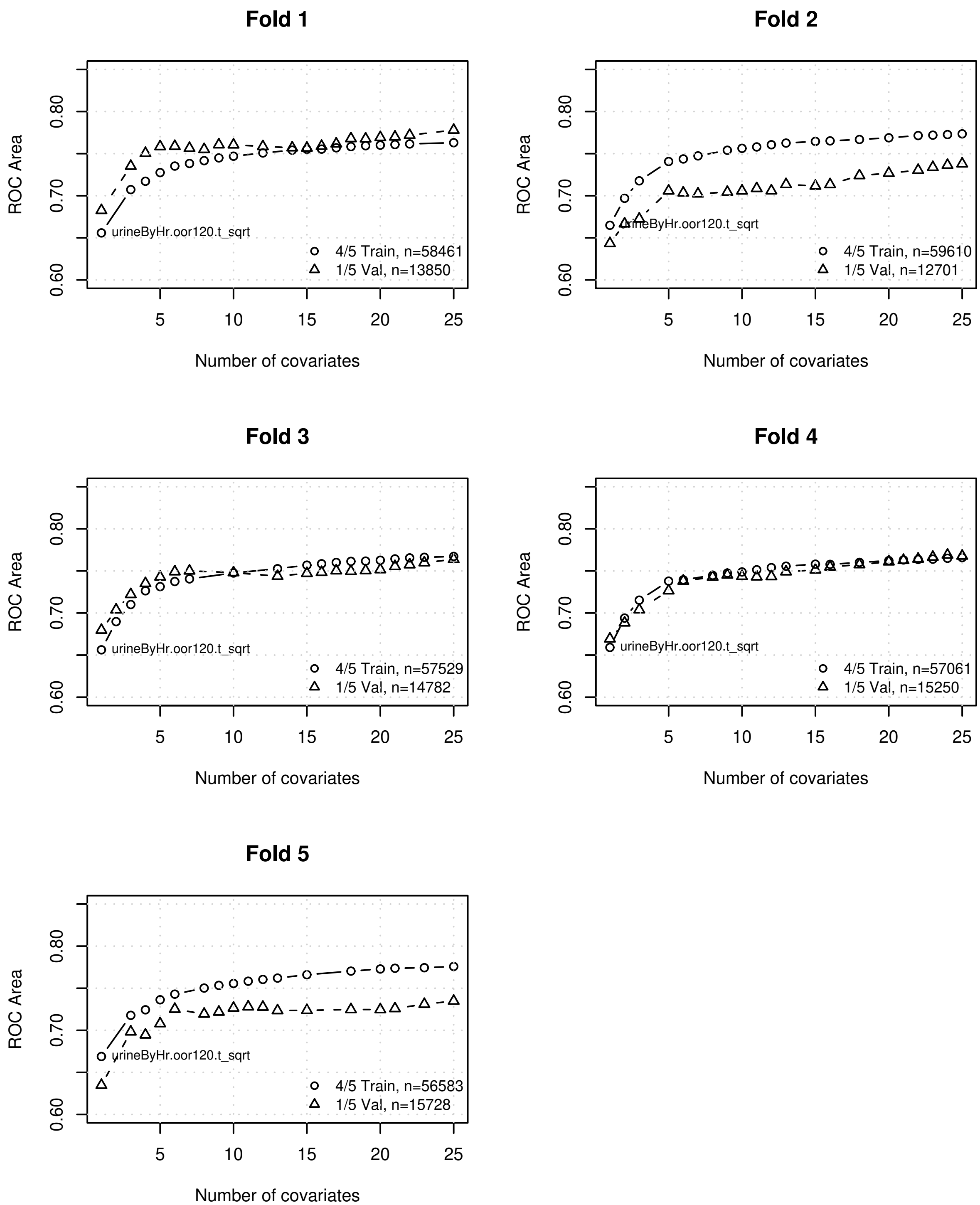


Figure F-10: `AKIM` final feature set performance on cross validation folds