Early Warning of Patient Deterioration in the Inpatient Setting

by

Gregory Alan Ciccarelli

B.S., Electrical Engineering, The Pennsylvania State University, 2009

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

Master of Science
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

February 2013

Signature of Author: ______________________________
Department of Electrical Engineering and Computer Science
January 18, 2013

Certified by: ______________________________
Thomas Heldt
Principal Research Scientist
Thesis Supervisor

Certified by: ______________________________
George C. Verghese
Henry Ellis Warren Professor
Professor of Electrical and Biomedical Engineering
Thesis Supervisor

Accepted by: ______________________________
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Committee for Graduate Students

Early Warning of Patient Deterioration in the Inpatient Setting

by Gregory Alan Ciccarelli

Submitted to the Department of Electrical Engineering and Computer Science
on January 18, 2013, in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

Early signs of patient deterioration have been documented in the medical literature. Recognition of such signs offers the possibility of treatment with sufficient lead time to prevent irreversible organ damage and death. Pediatric hospitals currently utilize simple, human evaluated rubrics called early warning scores to detect early signs of patient deterioration. These scores comprise subjective (patient behavior, clinician's impression) and objective (vital signs) components to assess patient health and are computed intermittently by the nursing staff. At Boston Children's Hospital (BCH), early warning scores are evaluated at least every four hours for each patient.

Many hospitals monitor inpatients continuously to alert caregivers to changes in physiological status. At BCH, each hospital bed is equipped with a bedside monitor that continuously collects and archives vital sign data, such as heart rate, respiration rate, and arterial oxygen saturation. Continuous access to these physiological variables allows for the definition of a continuously evaluated early warning score on a reduced rubric.

This thesis quantitatively assesses the performance of BCH's current Children's Hospital Early Warning Score (CHEWS). We also apply several standard machine learning approaches to investigate the utility of automatically collected bedside monitoring trend data for prediction of patient deterioration. Our results suggest that CHEWS offers at least a 6-hour warning with sensitivity 0.78 and specificity 0.90 but only with a prohibitively large uncertainty (48 hours) surrounding the time of transfer. Performance using only standard bedside trend data is no better than chance; improvement may require exploiting additional intra-beat features of monitored waveforms. The full CHEWS appears to capture significant clinical features that are not present in the monitoring data used in this study.

Thesis Supervisor: Thomas Heldt
Title: Principal Research Scientist

Thesis Supervisor: George C. Verghese
Title: Henry Ellis Warren Professor
Professor of Electrical and Biomedical Engineering

Acknowledgments

This thesis is the product of contributions from many individuals, each of whom has been crucial to shaping its final form. Each has earned my gratitude and deserves recognition.

Thomas Heldt, my primary research supervisor, for his patience, support, and guidance.

George Verghese, for pushing me to never compromise on clarity and his eye for detail.

BCH collaborators, especially Drs. Monica Kleinman and Paul Hickey, Christine Dube, Justine Bode, and Rachel Dabek, for their clinical perspective and responsiveness to my questions.

Steve Kogon, Dan Rabideau, Jenn Watson, and the Lincoln Scholars committee, for encouraging and enabling intellectual growth.

The Computational Physiology and Clinical Inference group, especially Sho Chaudhuri and Becky Asher, for productive discussions and proof reading.

Parents and family members, for their support, confidence, and unconditional love throughout this thesis and my life.

Mary, Queen of Saints, for interceding before God in order that I may have been granted the grace of perseverance to see this thesis through to its conclusion, and God, for granting that grace.

This work is sponsored by AFLCMC/PZE under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. This work has also been sponsored in part by the Children's Hospital Anesthesia Foundation, Boston Children's Hospital.

Contents

Abstract **3**

Acknowledgments **5**

List of Figures **11**

List of Tables **13**

1 Pediatric Early Warning Scores **15**

1.1 Project Background and Problem Statement . . . 16

1.1.1 Medical Need . . . 16

1.1.2 Early Warning Scores . . . 17

1.1.3 Multivariate Bedside Data . . . 19

1.1.4 Sensor Fusion . . . 22

1.1.5 Problem Statement . . . 23

1.2 Thesis Contributions . . . 24

1.3 Thesis Organization . . . 25

2 Vital Signs and Research Database **27**

2.1 BCH-MIT Collaboration . . . 27

2.1.1 Thesis-Specific Data . . . 28

2.2 Vital Signs: A Closer Look . . . 33

2.2.1 Heart Rate . . . 33

2.2.2 Respiration Rate . . . 34

2.2.3 Blood Oxygenation . . . 36

2.2.4 Blood Pressure . . . 38

2.2.5 Temperature . . . 39

2.3 Physiological Models . . . 39

2.4 Physiology of Cardiopulmonary Decompensation . . . 42

2.4.1 Respiratory Distress and Failure . . . 43

2.4.2 Sepsis . . . 44

2.5 General Observations on Feature Rubrics . . . 45

3 Data Exploration **49**
3.1 Transfer Population . . . 49
3.2 Transfer Reason and Call Type . . . 52
3.3 CHEWS Distribution vs. Time . . . 54
3.4 CHEWS Transition Probabilities . . . 56
3.5 CHEWS Underscoring . . . 58
3.6 Measurement Frequency . . . 61
3.7 Vital Sign Trajectories . . . 62
3.8 Bisected Changes Over Time . . . 64
3.9 Human Classification Performance . . . 70
3.10 Data Exploration Summary . . . 73

4 Classification and Prediction **75**
4.1 Feature Selection . . . 76
4.2 Decision Rules . . . 77
4.2.1 MAP Rule . . . 78
4.2.2 Support Vector Machines . . . 80
4.2.3 Decision Rule Complexity . . . 84
4.3 CHEWS and rCHEWS Classification . . . 85
4.3.1 MAP Classification . . . 86
4.3.2 SVM Classification . . . 87
4.3.3 rCHEWS Classification Method . . . 87
4.3.4 Classification Results . . . 88
4.4 The Prediction Problem . . . 90
4.4.1 Data Window, Uncertainty Window, and Observation Window . 91
4.4.2 Literature Survey . . . 94
4.4.3 The ROC with UW and WT . . . 98
4.4.4 ROC Calculation . . . 100
4.4.5 ROC Discussion . . . 104
4.4.6 Understanding Published Results with UW and WT . . . 108
4.5 CHEWS and rCHEWS Prediction . . . 113
4.5.1 SVM Training for Prediction . . . 113
4.5.2 Evaluation Method . . . 114
4.5.3 Prediction Results . . . 115
4.6 Summary . . . 117

5 Modified rCHEWS **119**
5.1 BCH Age-Normalized Trend Data . . . 119
5.2 Resampling and SVM Retraining . . . 124
5.3 The Misclassified Misfits . . . 127
5.4 Using Age . . . 128
5.5 Custom Features . . . 128

6 Conclusion and Future Work 133

A Classification Self Test 139

Bibliography 166

List of Figures

1.1 Typical bedside monitoring data. 21

2.1 BCH Children's Hospital Early Warning Score algorithm. 31
2.2 BCH normal vital sign ranges. 32
2.3 Standard ECG features. 35
2.4 ECG with the derived heart rate. 35
2.5 Oxygen hemoglobin disassociation curve. 37
2.6 Arterial blood pressure waveform. 39
2.7 Windkessel model of the heart. 40
2.8 Respiration rate from plethysmogram. 42

3.1 Study population gender comparison. 50
3.2 BCH control population by age over 12 months. 51
3.3 BCH transfer population by age over 12 months. 52
3.4 Control and transfer male to female ratios by age. 53
3.5 Transfer reason vs. call type 54
3.6 Transfer patient mean time until transfer after first threshold crossing. . 55
3.7 Control: mean score vs. time 57
3.8 Transfer: mean score vs. time 57
3.9 Control: score distributions vs. time 57
3.10 Transfer: score distributions vs. time 57
3.11 Control: selected score distributions vs time 58
3.12 Transfer: selected score distributions vs. time 58
3.13 Mean CHEWS and $+/-1$ standard deviation for control and transfer patients. 59
3.14 Transition probability: control patients. 60
3.15 Transition probability: transfer patients. 60
3.16 Control: CHEWS score evaluation frequency 62
3.17 Transfer: CHEWS score evaluation frequency 62
3.18 Decompensating transfer patient. 63
3.19 Stable control patient. 65

3.20 Unstable control patient. 66
3.21 Heart rate changes over time. 67
3.22 Respiration rate changes over time. 68
3.23 SpO_2 changes over time. 69
3.24 Self test performance by evaluator. 71
3.25 Self test performance by patient. 71
3.26 Justification key words 72

4.1 Decision rule examples. 78
4.2 SVM rule example. 80
4.3 SVM quadratic kernel example. 83
4.4 Notional simple and complex decision rules. 84
4.5 WT and UW relative to threshold crossing. 92
4.6 Observation window with non-zero warning time. 93
4.7 Different threshold-crossing time-points with the same sensitivity. 95
4.8 Different threshold-crossing time-points with the same sensitivity, WT=0. 97
4.9 Generic ROC heat map. 99
4.10 Examples of correct detection of need to transfer. 103
4.11 Relationship of small and large UW with specificity for ROCv2. 105
4.12 Algorithm sensitivity comparison under ROCv2. 107
4.13 ROCv2 heat map for OW=24 hours. 110
4.14 ROCv2 heat map for OW=48 hours. 112

5.1 BCH age-normalized AUC pt HR 121
5.2 BCH age-normalized AUC pt RR 121
5.3 BCH age-normalized respiration rate: overlap 122
5.4 BCH age-normalized heart rate and respiration rate used jointly for prediction. 124
5.5 BCH age-normalized heart rate and respiration rate with resampling used jointly for classification 125
5.6 Heart rate and respiration as vital signs, Gaussian normalized values, age is used as a feature. 129

List of Tables

1.1 Example pediatric early warning score (PEWS) rubric. 18

2.1 BCH mapping from age group name to age bracket in years. 33

3.1 CHEWS underscoring phenomenon 61

4.1 MAP and SVM classification with WT=0 hours. 89

4.2 MAP and SVM classification with WT=6 hours. 89

4.3 Selected ROC Notation 100

4.4 ROC performance from literature as well as ROCv2 results ("mimic") using BCH CHEWS scores. 111

4.5 CHEWS prediction summary 116

4.6 rCHEWS prediction summary 116

5.1 Custom Feature Classification Performance via SVM 131

Chapter 1

Pediatric Early Warning Scores

Patients admitted to the regular hospital ward or floor for observation or treatment commonly have a small number of physiological signals monitored continuously as part of their care. While the vast majority of these patients improve, a small subset might experience adverse events that necessitate transfer of the patient to a higher level of care, usually an intensive care unit (ICU). The question then naturally arises whether the transfer could have been predicted and consequently prevented, or carried out sooner. To help in the identification of patients at risk of acute physiological deterioration, clinicians have developed early warning scores that summarize, in a single number, the state of various organ systems. While useful, these scores still rely on intermittent human assessment of each patient. This thesis (i) quantitatively assesses the performance of a pediatric early warning score in use at a collaborating hospital, and (ii) investigates to what extent the continuously recorded physiological signals can be fused to aid in the automatic identification of the patient at risk of transfer to the ICU.

Section 1.1 describes the context and goals for this thesis. Section 1.2 outlines the thesis's contributions, and Section 1.3 describes the organization of the remaining chapters.

■ 1.1 Project Background and Problem Statement

To motivate this thesis and provide context for its contributions, Section 1.1.1 discusses the current medical need for early warning scores, and Section 1.1.2 reviews the current literature. Sections 1.1.3 and 1.1.4 summarize the data available for automatic transfer prediction and how that data can be processed. Section 1.1.5 defines the specific problem addressed by this thesis.

■ 1.1.1 Medical Need

Physiological decompensation is a state in which the body can no longer maintain homeostasis [1]. It can result from a variety of circumstances, such as strenuous exercise or disease progression. Studies have shown that decompensation or adverse events due to disease progression are associated with lower survival [2]. However, such decompensation might be predicted. For example, early signs of cardiac arrest [3] or the need for transfer to the ICU have been reported [4–6]. Such prediction can enable more timely and effective clinical intervention.

A study by McQuillan *et al.* observed that 39% of ward patients requiring transfer to the ICU were transferred late, and that suboptimal care definitely contributed to increased morbidity and mortality in 32.5% of the transfer patients [5]. Similar trends were identified in another study [4]. Therefore, if a patient is going to enter a decompensatory state, it would be best if the patient did so while in the ICU, where appropriate support and a higher level of care are immediately available. McQuillan *et al.* also observed that some transfers could have been prevented completely if appropriate action had been taken ahead of the transfer [5]. This is of note because among patients transferred from the wards to the ICU, the emergency department, the operating room, or the recovery room, patients transferred from the wards were most likely to die [7].

Though it is important to transfer patients in a timely manner to the ICU if necessary, it may be even better to anticipate or identify a decompensatory event with sufficient lead time so its occurrence can be averted altogether. This is because ICUs themselves can be dangerous environments, perhaps because of the complexity and invasiveness of the interventions. Between 11.9% and 19% of patients in a pediatric ICU (PICU) have been shown to develop infections, especially of the blood stream [8, 9]. A survey of 220 ICUs across twenty-nine nations found significant ICU errors at a rate of 38.8 events per 100 patient days. These errors included incorrect or inappropriate medication, equipment failures, and inappropriate monitor alarm silencing [10]. Turning off alarms stems from alarm fatigue due to the abundance and frequency of monitor alarms. Vendors have erred on the side of high sensitivity at the cost of low specificity, which is borne out by less than 1% of alarms resulting in a change in patient care [11].

A further need for early identification of impending decompensation is to prevent irreversible end-organ damage. Nguyen *et al.* have concluded that "[t]he care provided during the [emergency department] stay for critically ill patients significantly impacts the progression of organ failure and mortality. Although this period is brief compared with the total length of hospitalization, physiologic determinants of outcome may be established before ICU admission" [12]. Early goal-directed therapy (EGDT) has also been advocated, especially for sepsis management, by Rivers *et al.*. They showed a decrease of in-hospital mortality for patients with severe sepsis and septic shock when EGDT was implemented [13].

■ 1.1.2 Early Warning Scores

Because adverse events do have warning signatures, investigators have promoted clinical decision making aids, called Early Warning Scores (EWS), to ensure care keeps pace with patient condition [14]. The scores are a quantitative method for monitoring a patient's condition and appropriately escalating care if conditions worsen. They are an

Table 1.1: Pediatric early warning score (PEWS) rubric from Royal Alexandra Children's Hospital, Brighton, UK [15].

System Subscore	0	1	2	3
Behavior	Playing/ appropriate	Sleeping	Irritable	Lethargic or confused. Reduced response to pain.
Cardiovascular	Pink or capillary refill 1-2 seconds	Pale or capillary refill 3 seconds	Grey or capillary refill 4 seconds. Tachycardia of 20 above normal rate	Grey and mottled or capillary refill 5 seconds or above. Tachycardia of 30 above normal rate or bradycardia
Respiratory	Within normal parameters, no recession or tracheal tug	>10 above normal parameters, using accessory muscles, 30+% FiO2 or 4+ L/min	>20 above normal parameters, recessing tracheal tug, 40+% FiO2 or 6+ L/min	5 below normal parameters with sternal recession, tracheal tug or grunting, 50+% FiO2 or 8+ L/min
Score 2 extras for 1/4 hourly nebulisers or persistent vomiting following surgery				

example of high-level information fusion.

Pediatric early warning scores (PEWS) are a relatively recent invention and can vary in complexity [15, 16]. Like EWS rubrics, PEWS evaluate the patient in three categories: cardiovascular health, respiratory health, and neurological health. The information feeding into the categories includes vital signs such as heart rate, blood pressure, oxygenation, respiratory rate, and temperature, as well as an assessment of behavior and alertness. Fundamentally, PEWS and EWS are the same, with the only significant difference being the age-adapted ranges of normal vital signs. Deviations of the vital signs from normative values are scored based on severity, and category scores are summed to create a total score. The total score determines a particular action, such as continued four-hour assessment, increased frequency of assessment, evaluation for transfer, or immediate transfer to the PICU [17]. An example of a PEWS rubric is shown in Table 1.1.

PEWS have become widely implemented and show signs of success. A retrospective study found that 85.5% of patients transferred to the PICU showed a critical score at a median time of 11 hours and 36 minutes before transfer [18]. Duncan used a twenty-feature PEWS card specifically for identifying children in danger of cardiopulmonary arrest [16]. The rubric identified such children one hour prior to an event with a sensitivity of 78% and specificity of 95%. These lead times offer strong promise for the utility of PEWS. However, a recent PEWS review paper argues that there exists a shortage of rigorous validation for many of the proposed algorithms. Furthermore, the authors argue that clincially useful tools must be simple, with low inter- and intra-user variability [17]. Nonetheless, several studies have found that aggressive care for at-risk pediatric patients, identified through some mixture of physiological indicators, can positively impact patients by reducing respiratory and cardiac code (i.e., emergency) rates and mortality on the general ward [19–21], and ICU mortality [22].

Various hospitals have either directly adopted PEWS algorithms published in the literature or adopted them with variations. These include Children's Hospital of Denver, Children's Hospital of Orange County, and Boston Children's Hospital (BCH). The BCH early warning score (CHEWS score) is especially relevant to this thesis because of the MIT-BCH collaboration that supports this research.

■ 1.1.3 Multivariate Bedside Data

Prior to the early 1900s, bedside monitoring consisted of taking a patient's temperature and heart rate at regular intervals [23]. Gradually, the importance of charting these measurements over time as well as adding a quantifiable blood pressure measurement to the list became recognized. Today, the common vital signs are heart rate, temperature, blood pressure, respiration rate, and arterial oxygen saturation, many of which are monitored continuously in the hospital setting.

Bedside monitors aid in quantifying and tracking patient condition. They are prevalent on the general floor, in the ICU, and in the operating room. A multitude of devices that measure vital signs are plugged into the bedside monitor. In addition to temperature and blood pressure sensors, the electrocardiogram (ECG) and pulse oximeter are routinely employed. On the wards, the continuously-monitored vital signs that are most frequently available are heart rate, respiration rate, and arterial oxygen saturation. Temperature and blood pressure are assessed periodically, approximately every four hours.

The ECG is a voltage-versus-time waveform that characterizes the electrical activity of the heart, and is commonly used to calculate the heart rate. The respiration rate can also be derived from the ECG leads through measurements of the associated variation in transthoracic impedance, or through analysis of the ECG waveform itself, because the waveform is modulated by respiration [24].

The pulse oximeter primarily measures the relative amount of oxygen bound to hemoglobin, to determine arterial blood oxygenation (SpO_2). The spectroscopy underlying the pulse oximeter's function requires measuring the change in blood volume at the site of interest, for example a finger, which creates the pulse plethysmogram (PPG) waveform. This waveform can also be used to derive pulse rate and respiration rate [25].

Through the integration of the bedside monitors to a central server, the data from a dozen to two-dozen patient monitors is streamed to a central nursing station for continuous observation. The logged waveforms become part of the patient's medical record for review by clinicians.

Example vital sign data from a nine-year-old male patient and associated CHEWS scores are shown in Figure 1.1. (Chapter 2 discusses interpreting the data in this figure with respect to the prediction problem.) Data is referenced to the call time at time 0 in all subplots. Call time is the time at which the decision for transfer from the general ward to the ICU is made (magenta line). The first subplot shows the CHEWS score as documented by BCH clinicians and color coded by severity. A CHEWS score between

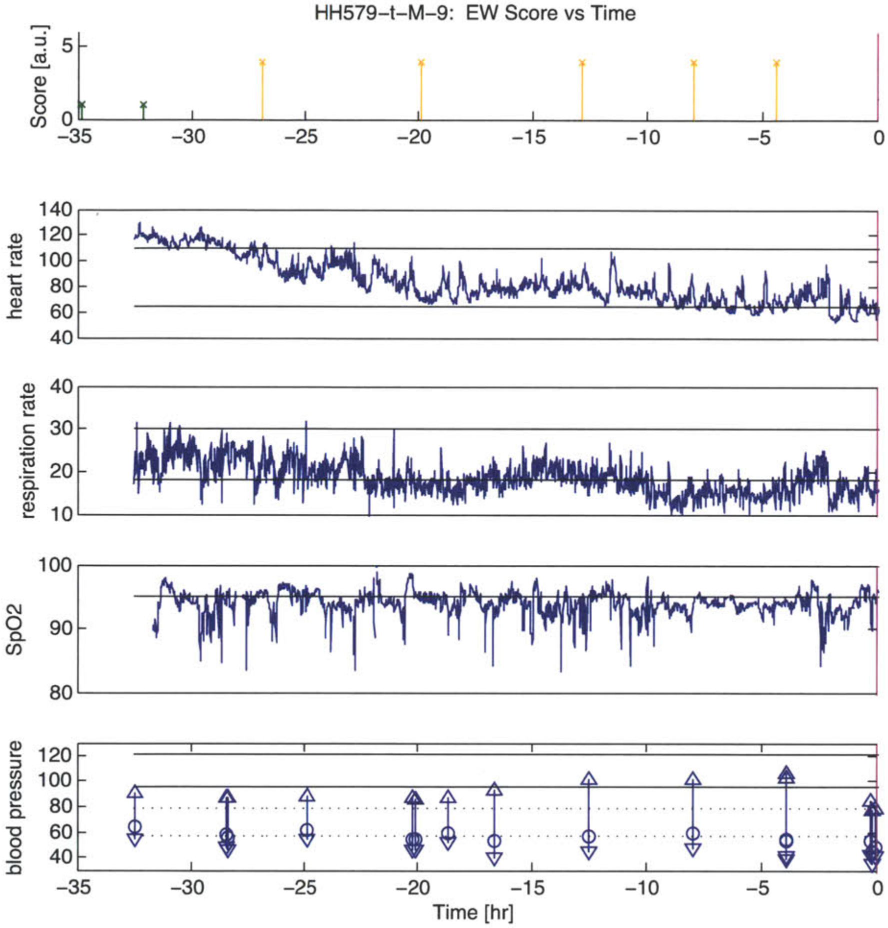


Figure 1.1: Patient HH579. CHEWS scores and vital signs of a nine-year-old male patient on the general floor. Units for the subplots are respectively: arbitrary units, beats per minute, breaths per minute, oxygen hemoglobin saturation percentage, and millimeters mercury.

zero and two (green) denotes a stable patient; a score of three (yellow) or four (orange) indicates a patient who warrants either increased monitoring or possible evaluation for transfer to the ICU. A score greater than or equal to five (red) demands immediate transfer to the ICU. The other four subplots show the trends of heart rate, respiration rate, blood oxygenation, and intermittent blood pressure measurements. Data colored blue come from a general ward monitor, and data colored green (not shown in Figure 1.1) come from an ICU monitor. Data colored blue may persist after the call time because the patient was not transferred immediately to the ICU. Each blood pressure measurement has three values: the top and bottom triangles represent the systolic and diastolic pressure, and the circle represents the mean. Thin horizontal black lines represent the upper and lower ranges of normal physiological values for the patient's age and gender, as specified by BCH. For blood pressure in particular, the solid black lines represent the normal systolic range, and the dashed black lines represent the normal diastolic range.

■ 1.1.4 Sensor Fusion

Bedside monitoring produces plentiful and diverse data, providing an ideal opportunity for sensor fusion of this data to characterize the patient. Sensor fusion is a process in which information from multiple sources is merged in order to infer characteristics of the object of interest. Sensor fusion can yield improved parameter estimation through the use of redundant measurements. Furthermore, it can offer a more complete picture as some sensors can provide information about the object of interest that others cannot provide. However, some of the advantages of sensor fusion may also be among its weak points. It is possible to corrupt "good data" with "bad data", and to formulate erroneous conclusions unless the combining framework is systematic and robust [26].

An architecture for sensor fusion involves several levels of processing raw data into actionable intelligence. Lower-level processing may analyze the raw signals separately to detect bad data, extract relevant features, or make simple logic-based decisions. Low-level features include general linear trends or abrupt departures from previous history. High-level processing considers the data jointly to extract features and to make decisions. Joint feature extraction or decision making is the first form of sensor fusion. A fusion step may use physical models that relate two or more processes to derive features, or it may look at the numerical behavior of the data such as the cross-correlation between the two processes. Both individual and joint features and decisions may then be combined at the highest level of processing to determine decisions through neural networks, Bayesian inference, or Dempster-Schafer theory, for example [26].

■ 1.1.5 Problem Statement

Unfortunately, the benefits from a data-centric environment that leverages sensor fusion have not been fully realized on the general ward. The monitors themselves may at most trigger an alarm if a particular vital sign crosses a simple threshold [27,28]. The many signal feeds at the central nursing station can be overwhelming to the one or two nurses trying to convert the stream of raw data into clinically actionable decisions. In practice, clinicians may only use five-minute windows of the data in addition to their own qualitative observations when they stop to check on the patient.

This scenario highlights several problems. First, despite continuous, real-time monitoring of the patient, the data is only used when clinical staff are physically present to assess the patient. The data that is used therefore only comprises a small snapshot of the total. For a stable patient at BCH, CHEWS assessments are done approximately every four hours. Therefore, much data may never be utilized. Such infrequent human monitoring may have been a significant cause of why an Australian study failed to find benefit in adult EWS in reducing unexpected death, cardiopulmonary arrests, and unplanned ICU admissions [29].

Second, there is little or no interaction among the alarm algorithms for different vital signs. For example, if a blood pressure reading drops to zero, an alarm might trigger even though the patient's ECG shows a normal heart beat.

Third, pediatric early warning score algorithms rely in part on subjective assessment of patient health, for example skin tone, so significant inter- and intra-clinician variability is possible. Last, some algorithms simply rely on deviations from normality, where normality is defined by an average over a group of patients. These algorithms offer little insight into a patient's specific physiological condition and could be based on derived parameters that have little if any obvious connection to a patient's health [30], making concrete intervention by the clinical staff difficult.

This thesis explores two questions. First, what is the utility of the current BCH CHEWS score? Second, to what extent can continuously acquired and streamed physiologic data from the patient's bedside be used to improve predictions of the need for escalation of care and transfer of the patient to the ICU? The second question focuses on the standard, continuously monitored vital signs of heart rate, respiratory rate, blood oxygenation, and intermittently measured blood pressure. We seek to understand what lead times might be achieved such that the medical staff can take preventative action, so the patient does not need to be transferred to the ICU, or is transferred in a timely manner.

■ 1.2 Thesis Contributions

This thesis makes three contributions. First, this thesis provides a thorough investigation of the BCH CHEWS score for monitoring patient health. Second, this thesis introduces a rigorous, clinically meaningful prediction metric that is lacking in the pediatric EWS literature. Third, this thesis uses this metric to benchmark the CHEWS predictive ability and the predictive ability of the stand-alone bedside monitoring data.

■ 1.3 Thesis Organization

Chapter 2 continues discussing the basics of the relevant physiological variables available for analysis, and introduces the data set under investigation. Chapter 3 describes data mining results and lays out performance benchmarks for subsequent algorithms. Chapter 4 introduces a prediction metric and compares current BCH CHEWS performance with an automated version of the BCH CHEWS algorithm on a reduced data set. Chapter 5 proposes and evaluates several modifications to the BCH CHEWS algorithm that exploit bedside monitoring data. Chapter 6 closes with a summary of this work and discusses future directions for pediatric physiological monitoring research.

Chapter 2

Vital Signs and Research Database

This thesis was done in collaboration with Boston Children's Hospital (BCH), which provided the vital signs from a subset of all patients treated over the last three years. This chapter will describe physiological models that underly these vital signs, the physiology captured by them, and their pathophysiological changes present under cardiopulmonary decompensation.

Section 2.1 begins with an overview of the BCH-MIT collaboration and continues with a description of the thesis data set. Section 2.2 describes in detail the vital signs available for decision making. Section 2.3 introduces several physiological models and parameters which are referenced in later chapters. Section 2.4 describes the mechanics of cardiopulmonary decompensation from respiratory distress and sepsis, which are two common causes for transfer to the ICU. Section 2.5 concludes this chapter with general observations on current early warning score (EWS) rubrics.

■ 2.1 BCH-MIT Collaboration

In January 2010, a project began that laid the groundwork for the use of continuously monitored general ward data for improving early warning systems. It was a collaboration between the Department of Anesthesia, Critical Care, and Pain Medicine at Boston Children's Hospital (BCH) and the Computational Physiology and Clinical Inference (CPCI) group at MIT's Research Laboratory of Electronics. BCH is a tertiary care facility that has over 300 patient beds and specializes in treating children and even some adults if their primary condition is from childhood or development.

BCH offers an ideal opportunity for exploring the utility of long-term, frequently sampled vital signs because all of the patient beds are equipped with bedside monitors (Philips Healthcare), and the data from all bedside monitors are archived for retrospective analysis. These monitors sample and digitally archive various patient vital signs as part of standard care. Specifically, the monitors usually collect three types of data: waveform data, trend data, and alarm data.

Waveform data include electrocardiogram (ECG) and pulse plethysmogram (PPG) signals sampled at 125 Hz. The waveform data are processed into trend data by Philips's algorithms. Trend data include the heart rate (HR) and respiratory rate (RR). Both are derived from the ECG and are output once per minute and possibly averaged over a longer duration. The blood oxygen saturation (SpO_2) and pulse rate are derived from the PPG. In the regular hospital rooms, blood pressure is measured intermittently, usually on the order of once every several hours, by an arm cuff using the oscillometric method. Alarm data consist of the alarms generated by Philips' algorithms. These alarms are typically based on simple threshold crossings of the waveform or trend data. For example, if heart rate crosses a pre-set upper bound, an alarm for tachycardia may sound.

■ 2.1.1 Thesis-Specific Data

The Philips data is logged and time-synchronized in a proprietary Philips format called the Research Data Export (RDE) format. To convert the RDE data into a format for algorithm development, Philips supplies a data viewer with data export capabilities. Furthermore, a converter was developed at MIT that reformats RDE data into waveform database (WFDB) format [24]. WFDB is an open-source format used for over twenty years by universities around the world as a way of interacting with a large database of physiological signals, PhysioNet, which is hosted by MIT [24]. Using an additional converter from PhysioNet, the WFDB data was converted to a Matlab file format to allow algorithm development on the Matlab computing software (Version 2011b) [31].

In addition to de-identified patient RDE data, clinical researchers also provide relevant patient meta-data such as the patient's age, gender, height, weight, clinical notes, call time, and Children's Hospital Early Warning Score (CHEWS). CHEWS is the BCH early warning score. The call time is the time at which the decision for transfer from the general ward to the ICU is made. The call time is an essential piece of clinical data because it acts as the fiduciary marker against which any predictive algorithm will need to be evaluated.

Because this thesis concerns data from real patients, a plan for data use, handling, and storage needed to be approved by the Institutional Review Boards at BCH and MIT to ensure that patient safety and patient health information were properly protected. De-identification of the patient data was accomplished at BCH. Data storage of de-identified data for patients is on MIT campus computers, though original copies of patient records also remain stored on BCH servers. Data for this project dates back to August 2010, when previous CPCI group members worked through initial data logistics and format conversion.

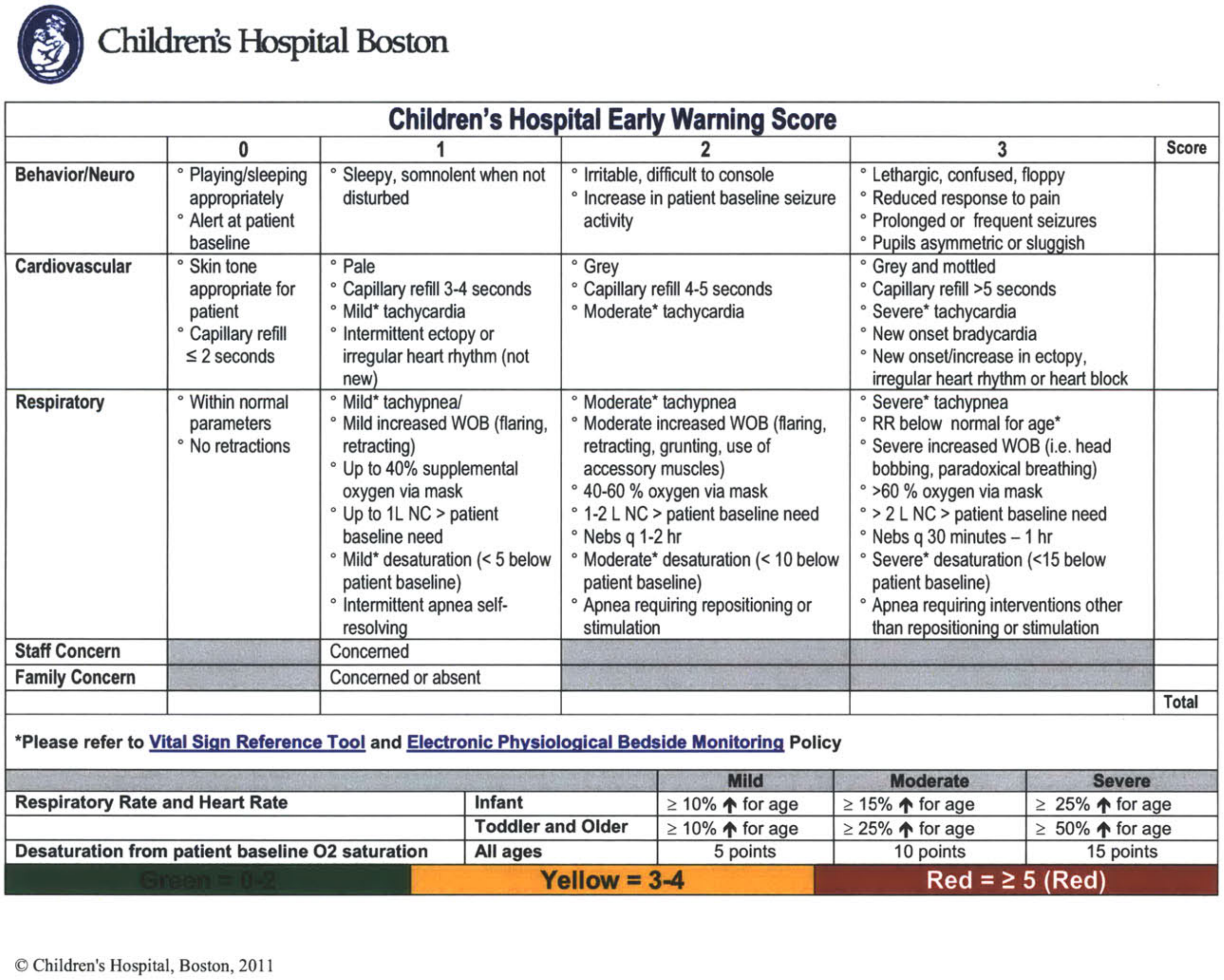
The required thesis data concerns two groups of patients: those patients on the general ward who are ultimately transferred to the ICU (the 'transfer' patients), and those patients who are not transferred (the 'control' patients). Each month, about 30-40 transfers from the general floor to the ICU occur at BCH, so potentially this many patients could be added to the project data base each month.

Unfortunately, this potential pool of hundreds of patients per year is not realized. The potential pool is shrunk by several factors. The first and most common reason is the quality of the recorded data itself. The core trend data set (HR, RR, SpO_2, and blood pressure) is already an impoverished data set of physiological indicators, so if one or more of these channels is missing, the data record is critically reduced. Also, some patients have less than ten hours of data. We excluded them because we hoped to predict six hours ahead of the transfer which would mean we would have less than four hours of data with which to do the prediction. Poor data quality accounts for over 50% of the patient data sets that are unusable. The second reason is a lack of a call date and time. Because the call time is the fiduciary marker, algorithm performance cannot be assessed without it. Call time is missing in approximately 17% of the patients, irrespective of whether the data quality is acceptable. Finally, we limited our study to patients $\leq$18 years of age, so older transfer patients were excluded. However, older patients represented only a small fraction of all transfers.

With data collected over October 2010 to March 2012, approximately 50 transfer patients and 50 control patients have good trend data and the required meta data for analysis. However, certain investigations only require the meta data itself, and that allows analysis of a larger set of over 200 transfer patients and 200 control patients.

BCH also provided their CHEWS scoring algorithm pictured in Figure 2.1. The CHEWS algorithm was an essential piece of knowledge because it provided a starting point for automated algorithms and an opportunity to benchmark algorithms that only operate on a subset of the vital-sign data. A significant challenge with pediatric early warning score systems is the age dependence of normal vital signs. Application of the CHEWS rubric as well as other published pediatric scores hinges upon an auxiliary table that lists age-appropriate normal values. BCH provided their table, which is reproduced in Figure 2.2. Table 2.1 lists the age ranges associated with each age category.

The normal vital sign ranges were determined from a literature survey by BCH. The ranges are similar to those used in other pediatric early warning scores. The



Children's Hospital Boston

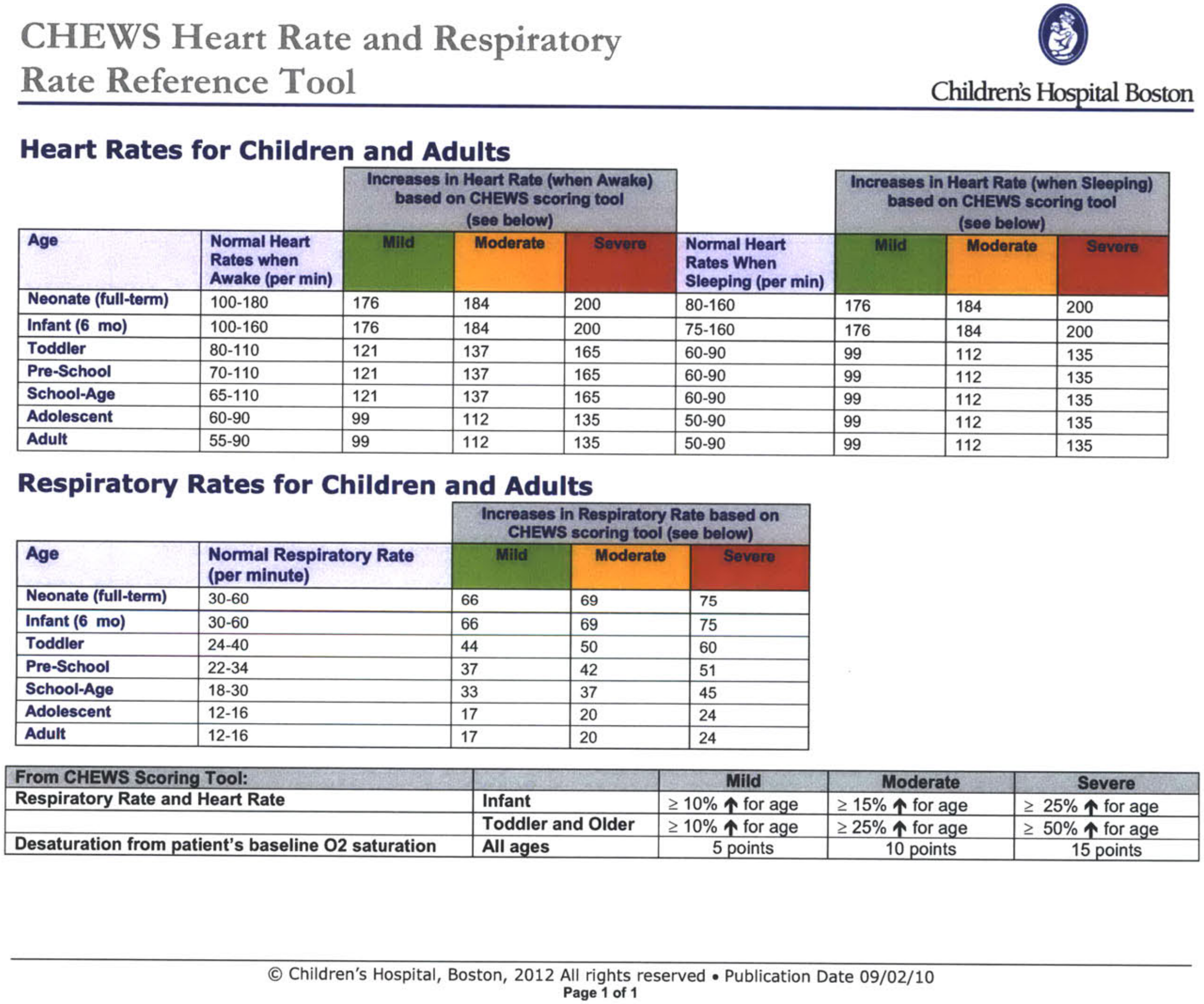
Children's Hospital Early Warning Score					
	0	1	2	3	Score
Behavior/Neuro	° Playing/sleeping appropriately ° Alert at patient baseline	° Sleepy, somnolent when not disturbed	° Irritable, difficult to console ° Increase in patient baseline seizure activity	° Lethargic, confused, floppy ° Reduced response to pain ° Prolonged or frequent seizures ° Pupils asymmetric or sluggish	
Cardiovascular	° Skin tone appropriate for patient ° Capillary refill ≤ 2 seconds	° Pale ° Capillary refill 3-4 seconds ° Mild* tachycardia ° Intermittent ectopy or irregular heart rhythm (not new)	° Grey ° Capillary refill 4-5 seconds ° Moderate* tachycardia	° Grey and mottled ° Capillary refill >5 seconds ° Severe* tachycardia ° New onset bradycardia ° New onset/increase in ectopy, irregular heart rhythm or heart block	
Respiratory	° Within normal parameters ° No retractions	° Mild* tachypnea/ ° Mild increased WOB (flaring, retracting) ° Up to 40% supplemental oxygen via mask ° Up to 1L NC > patient baseline need ° Mild* desaturation (< 5 below patient baseline) ° Intermittent apnea self-resolving	° Moderate* tachypnea ° Moderate increased WOB (flaring, retracting, grunting, use of accessory muscles) ° 40-60 % oxygen via mask ° 1-2 L NC > patient baseline need ° Nebs q 1-2 hr ° Moderate* desaturation (< 10 below patient baseline) ° Apnea requiring repositioning or stimulation	° Severe* tachypnea ° RR below normal for age* ° Severe increased WOB (i.e. head bobbing, paradoxical breathing) ° >60 % oxygen via mask ° > 2 L NC > patient baseline need ° Nebs q 30 minutes – 1 hr ° Severe* desaturation (<15 below patient baseline) ° Apnea requiring interventions other than repositioning or stimulation	
Staff Concern		Concerned			
Family Concern		Concerned or absent			
					Total

***Please refer to Vital Sign Reference Tool and Electronic Physiological Bedside Monitoring Policy**

		Mild	Moderate	Severe
Respiratory Rate and Heart Rate	**Infant**	≥ 10% ↑ for age	≥ 15% ↑ for age	≥ 25% ↑ for age
	Toddler and Older	≥ 10% ↑ for age	≥ 25% ↑ for age	≥ 50% ↑ for age
Desaturation from patient baseline O2 saturation	**All ages**	5 points	10 points	15 points
Green = 0-2		Yellow = 3-4		Red = ≥ 5 (Red)

© Children's Hospital, Boston, 2011

Figure 2.1: The BCH Children's Hospital Early Warning Score algorithm, reproduced with permission.



CHEWS Heart Rate and Respiratory Rate Reference Tool

Children's Hospital Boston

Heart Rates for Children and Adults

		Increases in Heart Rate (when Awake) based on CHEWS scoring tool (see below)				Increases in Heart Rate (when Sleeping) based on CHEWS scoring tool (see below)		
Age	**Normal Heart Rates when Awake (per min)**	Mild	Moderate	Severe	**Normal Heart Rates When Sleeping (per min)**	Mild	Moderate	Severe
Neonate (full-term)	100-180	176	184	200	80-160	176	184	200
Infant (6 mo)	100-160	176	184	200	75-160	176	184	200
Toddler	80-110	121	137	165	60-90	99	112	135
Pre-School	70-110	121	137	165	60-90	99	112	135
School-Age	65-110	121	137	165	60-90	99	112	135
Adolescent	60-90	99	112	135	50-90	99	112	135
Adult	55-90	99	112	135	50-90	99	112	135

Respiratory Rates for Children and Adults

		Increases in Respiratory Rate based on CHEWS scoring tool (see below)		
Age	**Normal Respiratory Rate (per minute)**	Mild	Moderate	Severe
Neonate (full-term)	30-60	66	69	75
Infant (6 mo)	30-60	66	69	75
Toddler	24-40	44	50	60
Pre-School	22-34	37	42	51
School-Age	18-30	33	37	45
Adolescent	12-16	17	20	24
Adult	12-16	17	20	24

From CHEWS Scoring Tool:		**Mild**	**Moderate**	**Severe**
Respiratory Rate and Heart Rate	**Infant**	≥ 10% ↑ for age	≥ 15% ↑ for age	≥ 25% ↑ for age
	Toddler and Older	≥ 10% ↑ for age	≥ 25% ↑ for age	≥ 50% ↑ for age
Desaturation from patient's baseline O2 saturation	**All ages**	5 points	10 points	15 points

Page 1 of 1

Figure 2.2: The normal vital sign ranges associated with the BCH CHEWS, reproduced with permission.

Table 2.1: BCH mapping from age group name to age bracket in years.

Age Category	Lower Bound [yrs]	Upper Bound [yrs]
Neonate	0	0.82
Infant	0.82	2
Toddler	2	4
Pre-School	4	6
School-Age	6	12
Adolescent	12	18

interaction of the age-based vital signs and the CHEWS score can lead to substantial scoring swings if ranges are strictly followed. For example, if a toddler's heart rate is 79 bpm (below normal by one bpm) he automatically rates a CHEWS of 3 in the cardiovascular category. However, if his birthday the next day places him in the preschool category, suddenly his CHEWS score is 0; he is perfectly healthy. The question naturally arises if there are not data driven ranges that could better classify patients.

■ 2.2 Vital Signs: A Closer Look

This section provides a closer exposition of common vital signs used in clinical monitoring. In particular, we present the underlying measurement modalities for acquisition of physiologic waveforms from which the vital sign trend data are derived. We also provide some physiological background for why monitoring HR, RR, and SpO_2 might allow us to determine which patients are at risk of decompensation.

■ 2.2.1 Heart Rate

The heart rate, HR, can be derived from the ECG waveform. The ECG is a time series of the heart's electrical activity. A single heartbeat contains a sequence of electrical signatures that are labeled chronologically as P, Q, R, S, and T as shown in Figure 2.3. The P wave is the depolarization of the atria. The QRS complex is the depolarization of the ventricles. The atria repolarize during this time, but the signature is buried by the large-amplitude ventricular depolarization. The T wave is the repolarization of the ventricles.

Because the R peak is prominent, it is commonly used as the temporal marker for calculating heart rate. The time between two R-R peaks is the R-R interval. The reciprocal of the R-R interval is the instantaneous HR. The HR signal is held between R-R intervals, as shown in Figure 2.4.

Heart rate is thus the beating frequency of the heart; it is one of several effectors that can change to maintain a constant blood pressure. A constant blood pressure level is necessary for proper perfusion of the body. The autonomic feedback control loop that maintains constant blood pressure is called the baroreflex [32]. If blood pressure falls, the baroreflex triggers an increase in heart rate, and total peripheral resistance, among other responses, and if blood pressure rises, the baroreflex triggers a decrease in these. Therefore, heart rate deviations from normal may indicate a compensatory response because of challenges to blood pressure. For example, if stroke volume is reduced, heart rate must increase to compensate for what otherwise would be a decrease in cardiac output and a concomitant decrease in blood pressure in the absence of changes in peripheral resistance [32].

There is significant research that links reduced variability in instantaneous HR with decreased autonomic function and poor patient outcome [35]. The variability may be measured at the beat-to-beat level via an analysis of R-R intervals [35] but also on the minute level [36].

■ 2.2.2 Respiration Rate

Respiration rate, RR, is the frequency of the inspiratory/expiratory cycle. In our data, a high-frequency current is injected across the ECG leads in order to measure the impedance change of the chest with time, as chest volume changes cyclically. The

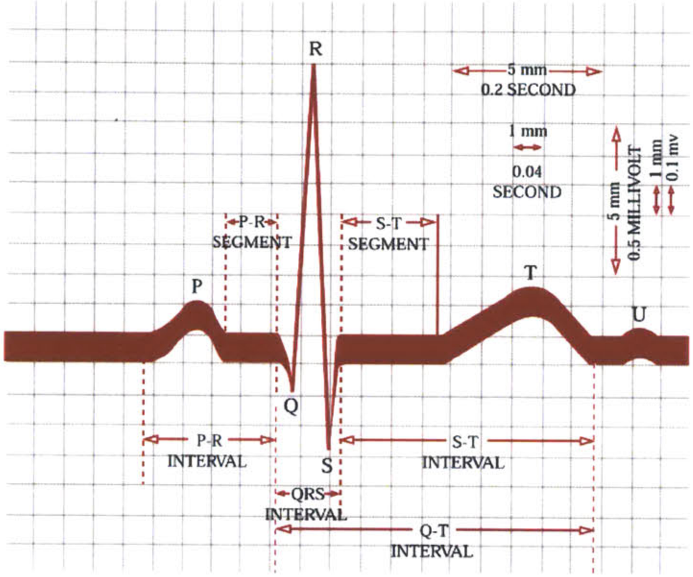


Figure 2.3: The standard features of an ECG trace with normal values [33].

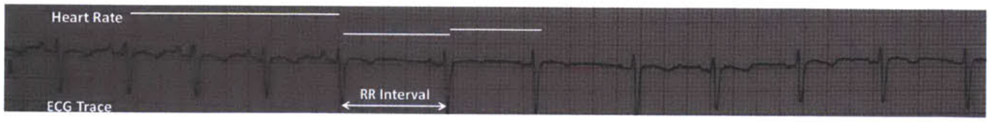


Figure 2.4: An ECG with heart rate derived as the reciprocal of the interval between R peaks [34].

injected signal frequency is outside the ECG frequency band. From the respiratory waveform, a Philips monitor derives the respiratory rate and displays the respiratory rate as a vital-sign trend, possibly averaged over several breaths.

Respiration rate is a controlled variable that is primarily sensitive to the partial pressure of arterial carbon dioxide, $PaCO_2$. Only if the partial pressure of arterial oxygen, PaO_2, drops significantly will oxygen chemoreceptors drive breathing. The alveolar ventilation equation quantifies how RR and $PaCO_2$ are inversely related, and $PaCO_2$ is related to blood pH through the Henderson-Hasselbalch equation [32]. A serious respiratory rate indicator is if the respiratory rate falls below normal. While that might be a pH-compensatory response, it might also mean that the patient has become tired and can no longer maintain the breathing rate necessary for his oxygen demands.

2.2.3 Blood Oxygenation

Blood oxygenation is the average percentage of oxygen bound to hemoglobin relative to its maximum (of four oxygen atoms per hemoglobin molecule). It is measured non-invasively through a pulse oximeter instead of a direct blood gas measurement, so the value is labeled SpO_2 instead of S_aO_2, which has been the traditional designation of blood oxygenation by direct arterial sampling. However, SpO_2 generally is a valid surrogate for S_aO_2 for specific applications, and it is the most widely used physiological measurement in clinical practice.

While SpO_2 is frequently monitored because it is so accessible, it presents a number of practical difficulties for predictive use. Because of the sigmoidal shape of the relationship between oxygen saturation and arterial partial pressure of oxygen, PaO_2 (Figure 2.5), the PaO_2 can actually be substantially reduced before there is a significant drop in SpO_2.

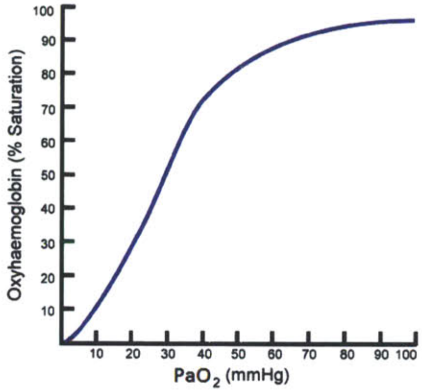


Figure 2.5: Sigmoidal oxygen hemoglobin disassociation curve causes relative independence of SpO_2 at moderate to high levels of arterial partial pressure of oxygen [37].

Even more unhelpful from the diagnostic perspective is how SpO_2 can be a misleading indicator of respiratory health if the fraction of inspired oxygen is unknown. For example, a patient might have an oxygen saturation of greater than 98% only because he is breathing 100% oxygen. This patient's respiratory system would be significantly compromised compared to a patient with the same oxygenation levels, but breathing room air.

CHEWS scores as well as other rubrics take into account both the absolute SpO_2 value and the amount of inspired oxygen support. Unfortunately, the latter information is not available from the bedside monitors. A normal SpO_2 value may only exist because of oxygen therapy whose presence is unknown to us. Therefore, the SpO_2 trend data may overestimate a patient's health. On the other hand, acute or chronic declines or sustained depressions of SpO_2 or intermittent desaturations are strong indicators of respiratory distress.

■ 2.2.4 Blood Pressure

Blood pressure (BP) is the force per area exerted by blood on the vessel wall. It changes with location in the body and as a function of time. In our data set, arterial BP is collected every four hours via an automated arm cuff that uses the oscillometric method to automatically detect systolic, mean, and diastolic pressures. More generally, arterial blood pressure is a waveform that varies characteristically over the course of a cardiac cycle (Figure 2.6). The systolic pressure, P_s, is the peak pressure obtained during the cardiac cycle. The diastolic value, P_d, is the minimum pressure during the cardiac cycle. Their difference, termed pulse pressure, is roughly proportional to stroke volume and therefore a surrogate for it. Systolic and diastolic values can be used to approximate the mean blood pressure using the $1/3\ P_s + 2/3\ P_d$ rule.

Blood pressure is a controlled variable. Therefore, the body will use effectors such as the heart rate, venous tone, total peripheral resistance, cardiac contractility, and fluid retention to maintain sufficient blood pressure to perfuse all organs. A low blood pressure has more severe immediate consequences than a high blood pressure because blood pressure is the driving force for organ perfusion. If blood pressure is high, local arteriolar resistance may be increased to reduce local blood flow. However, if blood pressure is too low, compensatory mechanisms might become exhausted. If perfusion is inadequate, the organ can suffer acute and sometimes irreversible damage [32]. Consequently, an acute decrease in mean BP is dangerous in itself. It also is an indicator because that the body is no longer able to hold it at a normal level [32].

One challenge associated with the arterial pressure measurement in our work is how to interpret two, near-simultaneous readings that are significantly different. Additionally, blood pressure in our study is taken only approximately every four hours, and sometimes even less frequently, thus limiting our ability to leverage this important vital sign for early detection of acute physiological decompensation.

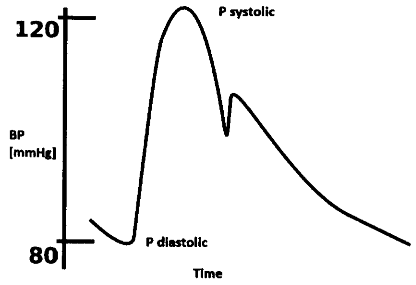


Figure 2.6: Arterial blood pressure waveform with typical adult values for systolic and diastolic pressures [38]

2.2.5 Temperature

Temperature is not available electronically from BCH, and it is included in only some published early warning score rubrics. In children, it has been found that temperature independently increases heart rate by 10 beats per minute (bpm) for each increase of 1 degree Celsius [39]. An elevated heart rate may therefore be a surrogate marker, though a non-specific one, for an elevated temperature. An elevated temperature is a key indicator for systemic inflammatory response syndrome (a precursor to sepsis) [40].

2.3 Physiological Models

In addition to leveraging trend data features, we hope to exploit known relationships among organ systems to aid meaningful data fusion. One method includes using established physiological models. As an example, the Windkessel model is a simple model for the systemic circulation. It is shown in the form of an electrical circuit analog in Figure 2.7.

The heart is modeled as a current source that generates impulses at the frequency of cardiac contraction. The impulse area is the stroke volume (SV), which is the amount of blood ejected from the left ventricle per beat. The average volume of blood pumped

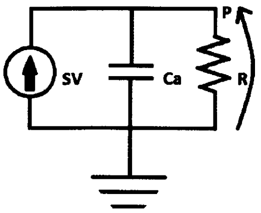


Figure 2.7: Windkessel model of heart (current source), arterial compliance, C_a, and resistive peripheral vasculature, R. Cardiac stroke volume is represented by SV.

by the heart per unit time is the cardiac output (CO) and is equal to the stroke volume times the heart rate (HR):

$$CO = SV \cdot HR \tag{2.1}$$

The blood enters the systemic circulation, which can be modeled as a capacitor or compliance in parallel with a resistor. The compliance represents the storage ability of the arteries. The resistor represents the resistance of the arterioles and capillaries. Stroke volume in this impulsive model is equal to arterial capacitance (C_a) times the pulse pressure (PP), where the pulse pressure is the systolic pressure, P_s, minus the preceding diastolic pressure, P_d:

$$SV = C_a \cdot PP = C_a \cdot (P_s - P_d)\,. \tag{2.2}$$

The physiological analog of Ohm's law states that pressure is equal to blood flow times resistance. Assuming steady state, which ensures no average flow through the compliance, we can now write

$$P = CO \cdot R. \tag{2.3}$$

Combining the above relationships yields several useful results. For example, though

C_a is unknown, a quantity proportional to SV and therefore CO can be estimated, which in turn can be used to estimate a quantity proportional to total peripheral resistance, TPR or R [33]:

$$CO = C_A \cdot PP \cdot HR \propto PP \cdot HR \tag{2.4}$$

and

$$R = \frac{\text{MABP}}{CO} \propto \frac{\text{MABP}}{PP \cdot \text{HR}} \tag{2.5}$$

The mean arterial blood pressure (MABP) is computed approximately from a blood pressure cuff measurement as

$$MABP = \frac{1}{3}P_s + \frac{2}{3}P_d. \tag{2.6}$$

Cardiac output (CO) reflects in part how well the heart is working as a pump, and TPR reflects the state of the patient's vasculature. For example, constricted arterioles substantially increase TPR because arterial resistance scales inversely with the fourth power of vessel radius [32].

One manifestation of the coupling between the respiratory system and the cardiovascular system is the modulation of the pulse pressure waveform at the respiration rate. During inspiration, the pulse pressure decreases, and during expiration, the pulse pressure increases. (When the increase is unusually high, this phenomenon is called pulsus paradoxus [41].) If the coupling is absent or changes substantially over a patient's stay, then presumably a pathological stimulus has altered the cardiovascular system's response to breathing. Furthermore, the relative change in amplitude of the modulation may suggest possible clinical treatments, because it has been shown that large amplitude modulation correlates with hypovolemia in ventilated patients [42]. Unfortunately,

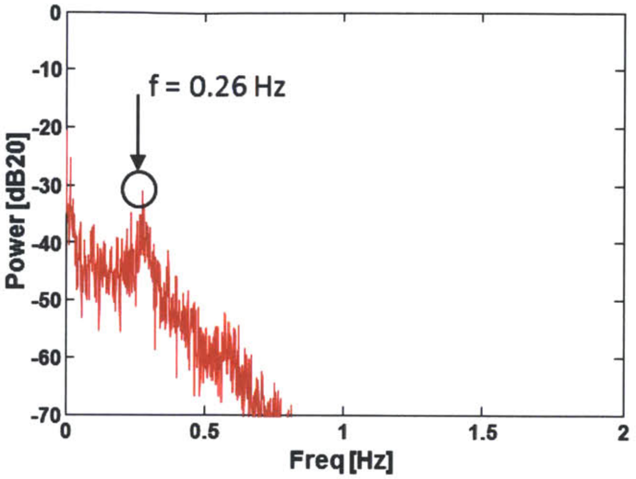


Figure 2.8: Power spectral density of pulse plethysmogram after envelope detection over one hour of data, estimated with Welch periodogram. Peak occurs at 16.5 events/minute. Philips respiratory rate trend data for this time is 16.3 breaths/min.

a continuous arterial pulse pressure reading is not available, as this requires an invasive measurement of arterial blood pressure. However, the pulse plethysmogram provides alternative access to the continuous pulse amplitude information. Using the PPG waveform, envelope detection, artifact removal, and basic spectral analysis, a distinct peak is present in the example shown in Figure 2.8. This peak agrees well with the respiratory rate from the respiratory rate trend data for this patient during this time period.

■ 2.4 Physiology of Cardiopulmonary Decompensation

The two primary motivations for closely monitoring vital signs are to quickly identify signs of cardiopulmonary decompensation and to evaluate response to treatment [43]. Early detection and treatment is crucial. While full recovery happens in 80% of patients with respiratory failure, if the condition deteriorates to cardiac failure, recovery probability is drastically reduced to 9% [44]. The sharp change in prognosis highlights the presence of a physiological tipping point beyond which recovery is improbable. Unlike in adults where cardiac arrest is primarily caused by ischemia to the heart, in children cardiac arrest is generally secondary to respiratory failure and/or severe, adverse metabolic changes such as those associated with sepsis [43]. Because respiratory distress and sepsis are two primary reasons for transfer, a basic overview of their physiology and trajectory to cardiopulmonary decompensation will be reviewed.

■ 2.4.1 Respiratory Distress and Failure

Respiratory distress is any condition that entails an increased work of breathing, even though oxygenation requirements may still be met [44]. By contrast, respiratory failure is insufficient ventilation and delivery of oxygen to meet the body's needs. Respiratory arrest is the absence of breathing [44]. While respiratory distress may not always proceed to respiratory failure, both are precursors for cardiac arrest in children, and therefore demand prompt treatment [44]. Additionally, respiratory distress is estimated to contribute to approximately 50% of pediatric ICU admissions [45].

The anatomy and physiology of children makes children especially prone to respiratory problems. Very young children have a disproportionately large tongue, smaller airways, and a more cartilaginous chest compared to children above eight years old (at age eight, the pediatric respiratory system is similar to an adult system, though it is still smaller in scale). Young children also have fewer alveoli and surface area for gas exchange than adults. Additionally, they have an oxygen demand per unit mass greater than adults which leads to hypoxia in about half the time as adults upon cessation of breathing. Upon cessation of breathing, a drop in oxygen saturation from 100% to 95% in infants takes less than two minutes, for toddlers it takes 2.5 minutes, and for children greater than three it takes 4 minutes [46] (Recall from Figure 2.5 that a 5% drop in saturation is associated with a very significant decrease in $PaO2$). Finally, the compensatory mechanism of rapid breathing may be counter-productive, because if breathing is too rapid there is insufficient time for gas exchange to occur [44].

Early signs of respiratory distress or failure exist, but they can be non-specific or hidden. Use of accessory muscles is a sign of increased work of breathing. Unfortunately, the muscles are not optimally positioned for benefit in the young, and they tire easily [46]. Therefore, respiration rate may exhibit an oscillatory pattern that foreshadows respiratory failure [44]. While increased respiratory rate may be a sign of respiratory distress, it could also be compensation for metabolic acidosis, or an indicator for increased temperature. Every one-degree increase in temperature can lead to an increase in five breaths per minute in respiratory rate [44]. While a SpO_2 greater than 93% may indicate adequate oxygenation, it may obscure the additional underlying effort put forth to maintain that level [44].

■ 2.4.2 Sepsis

Sepsis is another pathology that can ultimately lead to cardiopulmonary failure. Sepsis is actually a spectrum of conditions. The pre-sepsis condition is called severe inflammatory response syndrome (SIRS). When SIRS is diagnosed in conjunction with an infection, the condition is called sepsis. Sepsis with non-cardiac organ failure is severe sepsis, and sepsis with cardiac failure is septic shock. Severe sepsis affects 42,000 children each year in the US and results in about 4,000 deaths annually. It is especially prevalent in children less than one year of age [47].

Despite intensive study, sepsis is still poorly understood. While its detrimental effects were initially thought to be due to overcompensation of the immune system, more recent research suggests there is a strong component of immune suppression [48]. Therapy generally uses antibiotics and aims for cardiopulmonary stability by maintaining adequate oxygen and fluid levels to avoid respiratory distress and hypotension [47]. A number of adult studies have shown the efficacy of early goal-directed therapy that aggressively treats patients to keep vital signs stable [13,49,50]. Therefore, early detection of sepsis is crucial for favorable outcomes.

Unfortunately, sepsis detection is difficult because it presents with non-specific symptoms. These include hypoxia, tachycardia, tachypnea, fever ($> 38.5^o$ C) or hypothermia ($< 36^o$ C). Detection is especially challenging in children because adults show a progressive decline in health while children appear fine until a sudden, severe decompensation [47].

■ 2.5 General Observations on Feature Rubrics

Many researchers have proposed rubrics to quantify a patient's health by a numerical score, as an aid to current treatment and/or prediction of the patient's course of health [16, 18, 51, 52]. Some rubrics predict mortality upon transfer to the ICU, probability of transfer from the floor to the ICU, probability of transfer from the emergency department directly to the ICU, or probability of cardiac or respiratory arrest.

In all studies reviewed for this thesis, only the one by Sharek [20] considered change in condition as an indicator of importance. Specifically, Sharek conducted a prospective study at Lucile Packard Children's Hospital in California, in which a rapid-response team was activated if any of the following criteria were met:

1. a staff member was concerned about the child;
2. acute change in heart rate;
3. acute change in respiration rate;
4. acute change in oxygen saturation;
5. acute change in blood pressure;
6. acute change in level of consciousness.

Unfortunately, “acute” was not defined, which shows that even in attempts at quantitative, repeatable rubrics, subjectivity can still prevail in the analysis. Interestingly, this study showed statistically significant improvement in patient outcome as measured by a reduction in the number of code events outside the PICU and a reduction in hospital-wide mortality.

Tibballs [19] has an annotation at the bottom of his rubric in which worsening vital-sign trends should be observed and reported, but stops short of saying that such deterioration is sufficient grounds for activation of the medical emergency team. The other studies besides Sharek, including the Tibballs activation rubric, all focus only on absolute values of vital signs. Generally those values are compared against age-dependent norms for heart rate, respiration rate, oxygenation, and blood pressure, in particular systolic blood pressure.

Normal vital sign ranges as well as the definitions of age brackets show moderate variability among the rubrics, yet those norms significantly affect the contribution of the vital signs to the complete score. Furthermore, the norms themselves, as well as the the weight assigned to the score for deviations from normal, appear with little or no quantitative justification. Brilli provides a sensitivity analysis for pairs of calling criteria [21], but the best methodology in this regard is by Pollack [53], with the Pediatric Risk of Mortality (PRISM) III score. The PRISM III score predicts pediatric mortality in the ICU. The investigators performed a series of Monte Carlo simulations and logistic regression calculations to determine the best normal ranges.

Existing algorithms are almost always memoryless, yet this is contrary to the avowed practice of several BCH personnel and to physiological intuition. Clinicians baseline the patient upon admission; even if the patient has an abnormal vital sign, clinicians do not give this as high a concern as if the patient's vital sign changed from baseline. Measurements are done infrequently, approximately every 4-5 hours at BCH. The assigned scores are based upon spot measurements of the vital signs, with no explicit dependence upon what happened in the intervening time, even though the relevant data may be archived by monitoring equipment. One study accounts for memory by incorporating the maximum deviation from normal over a past window when computing a score; however, it does not look for trends or the duration of that maximum value [53].

Because the truth for predicting patient transfer when the patient has not coded is based upon the judgment made by human personnel, any algorithm that seeks to mimic human judgment as a precursor for improving upon human judgment should also take into account trends in vital signs as well as acute changes in vital signs.

Armed with an understanding of the vital signs captured by the bedside monitoring data as well as the physiology of cardiopulmonary decompensation, we can begin mining the collected data for significant features in Chapter 3.

Chapter 3

Data Exploration

Because this research project provides an investigation into the utility of automated data collection, a significant effort went into exploring the acquired data for obvious and interesting patterns. This chapter will summarize several explorations in the analysis and trends of the data.

Sections 3.1 and 3.2 characterize the transfer patient population's meta-data. Sections 3.3 and 3.4 analyze the progression of CHEWS scores over time. Section 3.5 investigates a bias towards assigning lower CHEWS scores than are merited and Section 3.6 analyzes the frequency of CHEWS evaluation. Section 3.7 focuses on the evolution of vital-sign trends from up to 48 hours before transfer till the actual transfer occurs. In Section 3.8, we explore how the vital signs compare in a particular window of time with respect to a final window just prior to transfer. Section 3.9 describes a small experiment in which humans were asked to classify patients as control or transfer. Section 3.10 closes this exploratory chapter with a summary of preliminary findings.

■ 3.1 Transfer Population

As part of the characterization of the patient cohort, we studied the age and gender distributions of transfer patients. Figure 3.1 shows histograms based entirely on meta data of transfer and control patients. The general control population and transfer population are approximately 50/50 male and female. However, controls were selected

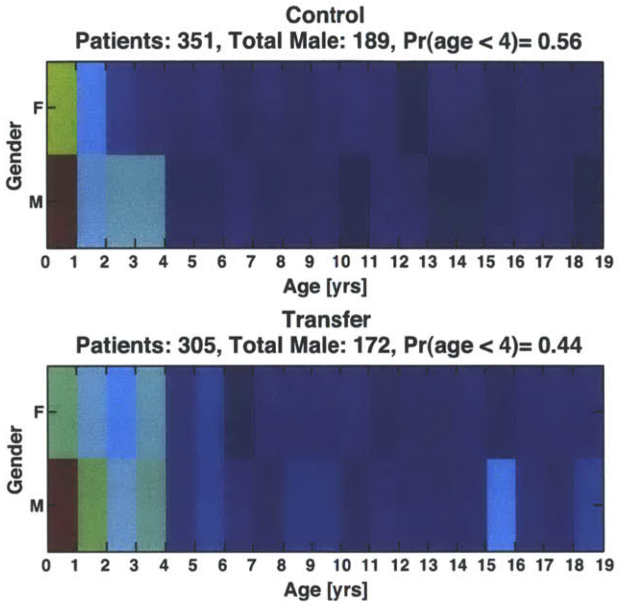


Figure 3.1: Study population gender comparison. The color is proportional to the number of patients in each bin with red corresponding to more patients and blue to fewer patients. Each bin counts patients that satisfy: lower age limit $\leq$ patient's age $<$ upper age limit.

to match specific transfer patients with regard to several variables, including gender and age, so similarity between the two populations is expected.

Nearly half of all transfer patients are less than 4 years old. Furthermore, males less than 1 year old are represented twice as much as females less than one year old in the transfer population. Remarkable as these observations are, the general population of BCH patients may exhibit the same asymmetries. Consequently, these results only describe the data set and may not allow any conclusion about a prior probability of transfer based on age or gender.

To assess prior probability of transfer in general, and when conditioned on gender, we received a census of all control patients admitted to BCH for the 2011 calendar

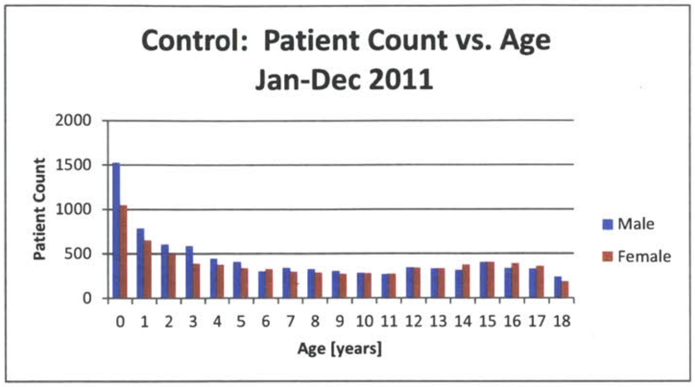


Figure 3.2: BCH control population by age over 12 months.

year (patients admitted to the psychiatric ward and to the ICU were excluded). A histogram broken down by age for control patients is shown in Figure 3.2. The transfer plot is shown in Figure 3.3. We limited the maximum age to ≤18 years of age in order to compare against all the transfers considered in our study. Unfortunately, our study data is continuous starting in April 2011, so to compare a 12 month calendar year, we perform our age histogram analysis from April 2011 through March 2012. From the census data, we conclude that the prior probability of transfer to the ICU is only 1.6%. Consequently, any early warning system must have excellent specificity in order to minimize total false alarms.

We observe that the estimated fraction of transfer patients less than four years of age is 0.44, while in the general control population in Figure 3.2, the fraction of patients less than four years of age is 0.38. Consequently, younger patients may be more susceptible to transfer than older patients. We also see that the high probability of transfer for males compared to females less than one year of age is not strictly a

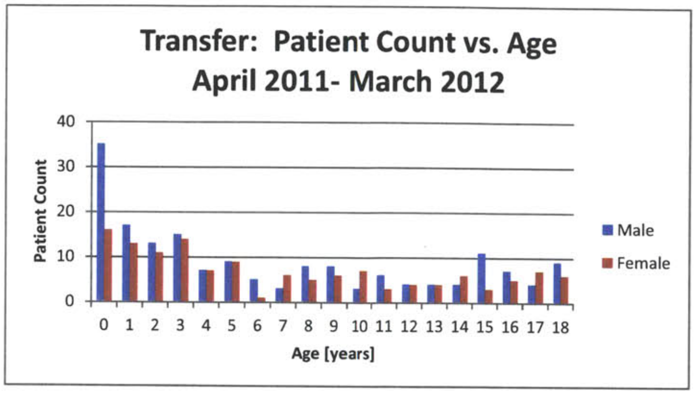


Figure 3.3: BCH transfer population by age over 12 months.

reflection of more males than females in the general population. Figure 3.4 compares the male female ratio for the transfer population to the malefemale ratio in the control population. Because the ratios are dissimilar in the youngest age bracket, it appears that males may be at a higher risk of transfer than females.

■ 3.2 Transfer Reason and Call Type

Every transfer patient at BCH has a call type and transfer reason. The call type is the justification for activating the transfer team. The transfer reason is the clinical justification for moving the patient to a higher level of care. We analyzed the joint transfer reason and call type matrix to identify potentially common pathological causes for transfer. If such subgroups existed, algorithms might leverage pathology specific features for improved early warning. The joint histogram of transfer reason and call type in Figure 3.5 shows that the most prevalent transfer reason is respiratory distress and

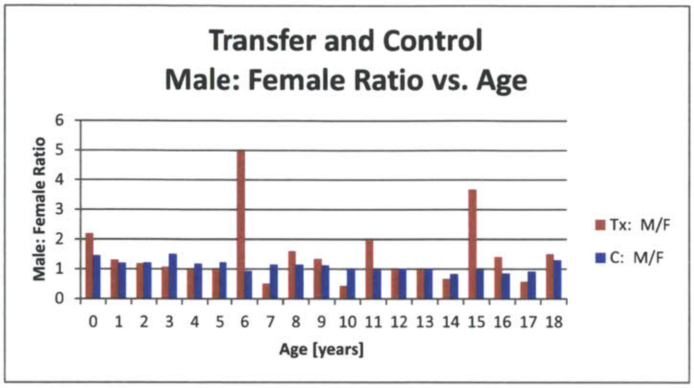


Figure 3.4: Control and transfer male to female ratios by age.

the most prevalent call type is for evaluation. The need for more extensive monitoring is the second most common transfer reason. By contrast, our initial expectation was that sepsis would be the number-two transfer reason after respiratory distress. However, some patients transferred to allow more vigilant monitoring may actually have had sepsis, though they may not have been diagnosed until after transfer. Re-visiting the transfer reason would be necessary to accurately assess whether or not sepsis is a major transfer reason. Nonetheless, this analysis suggests that the fraction of specific pathological conditions, such as congestive heart failure (CHF) or hyperkalemia, is negligible compared to the non-specific but widely prevalent "respiratory distress" transfer reason.

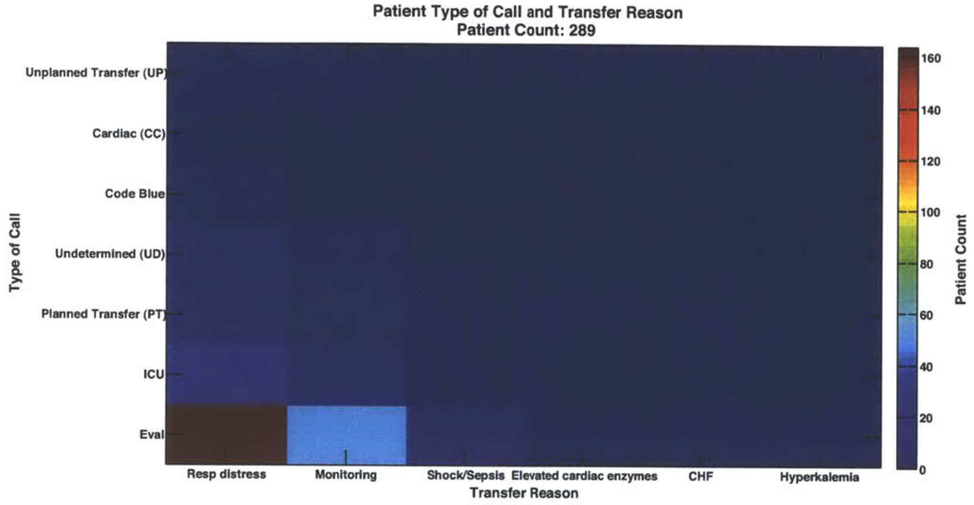


Figure 3.5: Transfer reason vs. call type

■ 3.3 CHEWS Distribution vs. Time

Figures 3.7 and 3.8 show the mean value (solid line) and quartiles (dotted lines) for the CHEWS scores over time. For transfer patients, time 0 was the call time; for controls, time 0 was the time of the last documented CHEWS score. All scores that fell within the respective four-hour window were averaged. The number of scores in each four-hour window is shown in the bottom subplot as a histogram.

Figures 3.7 and 3.8 investigated at what point, if any, the *average* transfer CHEWS score begins to deviate from the *average* control score. The point of deviation would be a first estimate at predicting whether the patient's condition had deteriorated beyond a control patient. As expected, the transfer patients' CHEWS scores ramp up very quickly in the final four hours before transfer while the control scores stay relatively flat. Furthermore, the control scores do not rise far beyond a CHEWS score of 2 while the average transfer score climbs nearly to 4 in the final 10 hours before transfer. This suggests that the decision score of 4 may be too conservative. The decision threshold

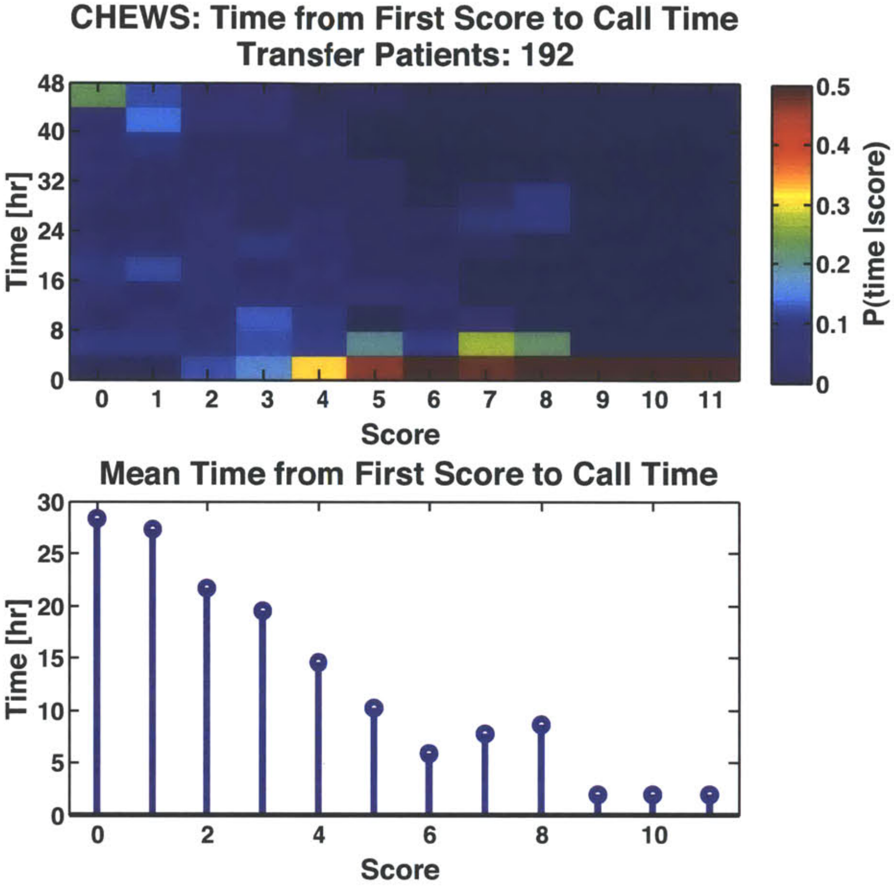


Figure 3.6: Transfer patient mean time until transfer after first threshold crossing.

to take action could perhaps be lowered to 3 or some fraction greater than 2 because transfer patients have begun to diverge from controls before a score of 4. If the threshold were lowered to 3 (i.e., declare a transfer if the score is greater than or equal to 3), then sensitivity is 0.83, specificity is 0.78, and the mean time from the first recorded score greater than or equal to 3 is 19.5 hours. Figure 3.6 shows the approximate mean time until transfer from the first instance of a score greater than or equal to the threshold.

Unfortunately, the clean deviation of the mean overstates the discriminating ability of the score. Figures 3.9 and 3.10 show a two dimensional version of Figures 3.7 and 3.8. The 2D versions show the empirical probability distribution for each four-hour window. While there is still a clear migration of the most probable CHEWS scores to higher values for transfer patients, significant variability in scoring is still obviously present. The standard deviation of the mean score for both control and transfer patients is shown in Figure 3.13.

Figures 3.11 and 3.12 are identical to Figures 3.9 and 3.10 but show the distributions from a traditional hypothesis testing view point. The histogram in Figure 3.12 shows the distribution of CHEWS scores from transfer patients for the two days prior to transfer. For as much as 24 or more hours before transfer, the scores are clustered around the 0 and 1 mark with heavy tails out to 4. However, as time approaches the call time, t=0, the center of mass gradually shifts outward and the histogram begins to flatten. By contrast, the mean of the distribution for control patients, stays around the 1-2 range.

In summary, while a divergence of CHEWS scores between controls and transfers is present at an aggregate level, significant variability exists between the two groups. The overlap between the populations limits the success of setting a simple threshold.

■ 3.4 CHEWS Transition Probabilities

A transition probability is the probability of changing to a particular state, given a current state. In this section, we evaluated the CHEWS score transition probabilities. The motivation was to explore trends in scoring and transfer values. In particular, we were interested to find out whether there is a particular CHEWS score following which there is a high probability for a patient's score to escalate to higher values.

Figures 3.14 and 3.15 show the transition probabilities for control and transfer patients. The probability $Pr(to = Y|from = X)$ is the probability of being rated a

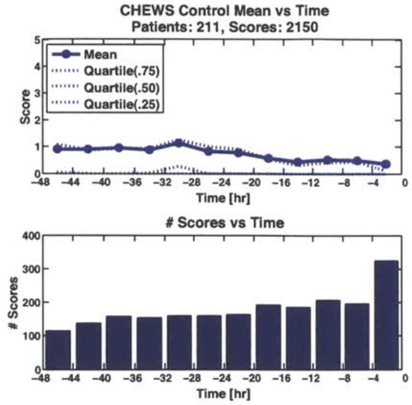


Figure 3.7: Control: mean score vs. time

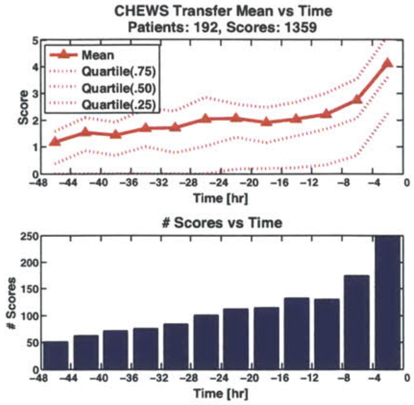


Figure 3.8: Transfer: mean score vs. time

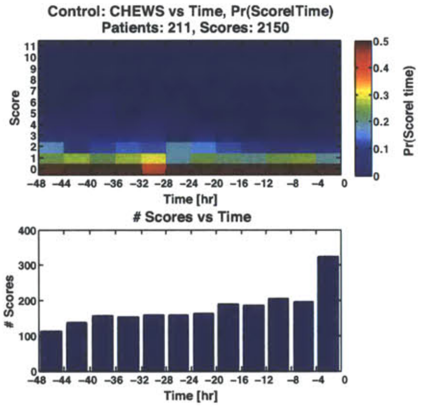


Figure 3.9: Control: score distributions vs. time

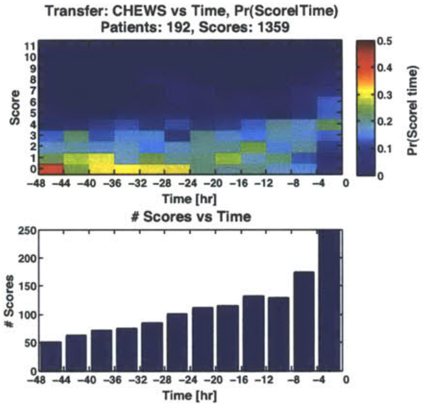


Figure 3.10: Transfer: score distributions vs. time

score of Y given the score X at the previous evaluation point. Because the numbers in the upper subplot define a conditional probability mass function, where the conditional variable is the "from" score, each column sums to one. The maximum possible CHEWS score is 11, and scores range from 0 to 11. The final row and column are "Tx" for transfer. Probability mass in that cell means that after being assigned a "from" CHEWS score, no additional scores were taken before the patient was transferred. The lower

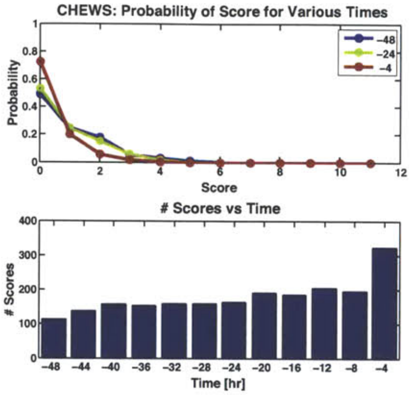


Figure 3.11: Control: score distributions at three time intervals [-48, -44], [-24, -20], and [-4, 0] hours

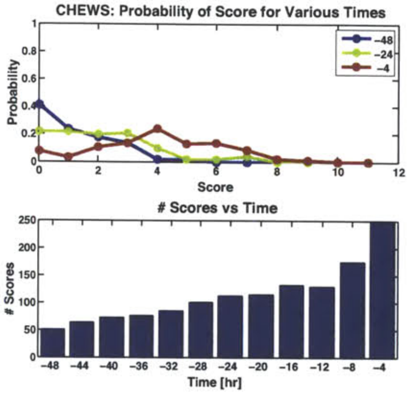


Figure 3.12: Transfer: score distributions at three time intervals [-48, -44], [-24, -20], and [-4, 0] hours

subplot shows the number of scores used to estimate each column's distribution.

These results were generated using the meta data collected retrospectively from nurse notes and are completely independent of the Philips RDE data. The bright diagonal line for scores 0 to 4 means that the most likely score transition is a self transition; the patient score does not change. Also, the decreasing number of events from 0-0 to 4-4 along the diagonal show that the 0-0 state is more common than the 4-4 state. All else being equal, once a patient is in the four state, the patient is likely to need transfer.

In conclusion, the transition matrix showed a tight clustering of transitions in low scoring states for controls, while transfers tended to maintain their current state or deteriorate. Controls with higher scores tended to improve while transfers deteriorated.

■ 3.5 CHEWS Underscoring

We conducted a CHEWS underscoring analysis in which we benchmarked how well the CHEWS rubric definitions were followed. At each point at which a CHEWS score was

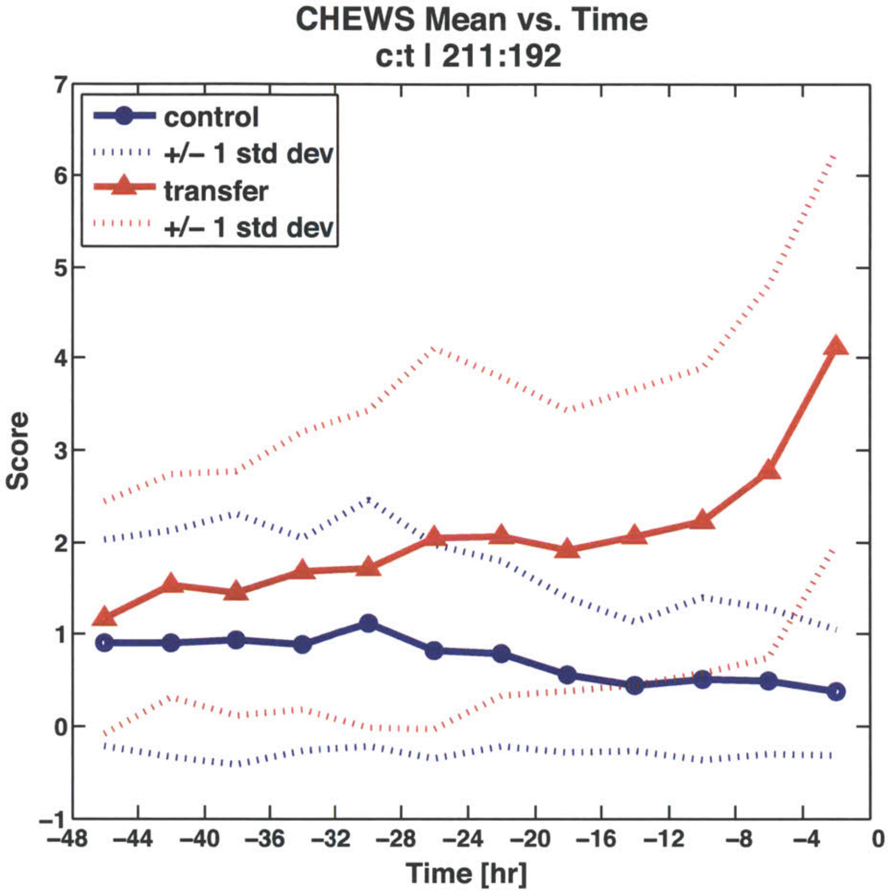


Figure 3.13: Mean CHEWS and +/ − 1 standard deviation for control and transfer patients.

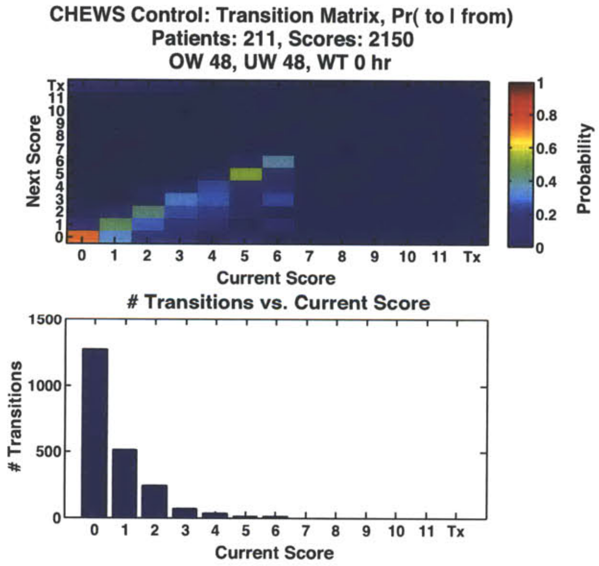


Figure 3.14: Transition probability: control patients.

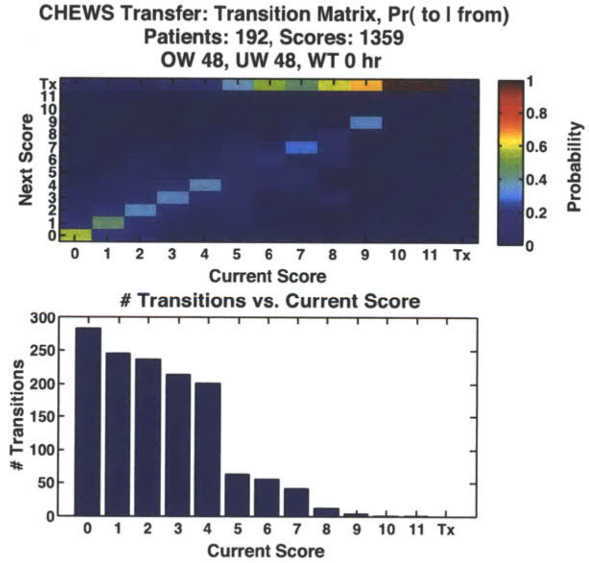
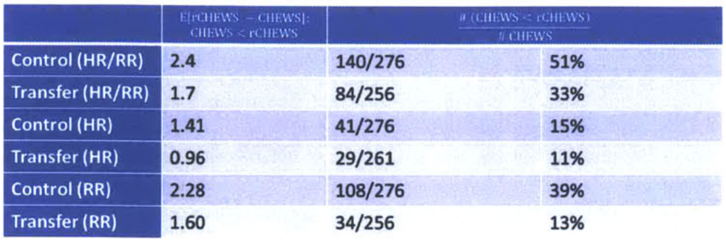


Figure 3.15: Transition probability: transfer patients.



	$E[rCHEWS - CHEWS]$: $CHEWS < rCHEWS$	$\frac{\#\,(CHEWS < rCHEWS)}{\#\,CHEWS}$	
Control (HR/RR)	2.4	140/276	51%
Transfer (HR/RR)	1.7	84/256	33%
Control (HR)	1.41	41/276	15%
Transfer (HR)	0.96	29/261	11%
Control (RR)	2.28	108/276	39%
Transfer (RR)	1.60	34/256	13%

Table 3.1: CHEWS underscoring phenomenon

taken, we retrospectively used the monitoring heart rate and respiratory rate data to compute a reduced CHEWS (rCHEW) score. The rCHEWS used the BCH normal vital sign limits and BCH rubric. We saw from Section 3.3 that CHEWS does have some discriminating ability, and we sought to quantify the contribution of a subset of the rubric data to the score.

Because the reduced CHEWS is a strict subset of the full CHEWS score, the CHEWS score should always be at least as great as the reduced CHEWS score. If the CHEWS score is less than the rCHEWS, we say that the patient has been underscored. Table 3.1 shows the results. While the magnitude of underscoring is about the same for controls and transfers, controls are more frequently underscored than transfers. The respiration rate contributes more to this bias than heart rate. This study suggests that clinicians are selectively discounting high CHEWS merited by respiration rate for control patients. Clinicians are focusing on other indicators that supersede the quantitative scoring measures. This result does not bode well for the performance of an automated system based only on moment-to-moment rCHEWS values.

■ 3.6 Measurement Frequency

Figures 3.16 and 3.17 capture the frequency of CHEWS score evaluations. The potential prediction feature at stake is whether the frequency of measurements signals an

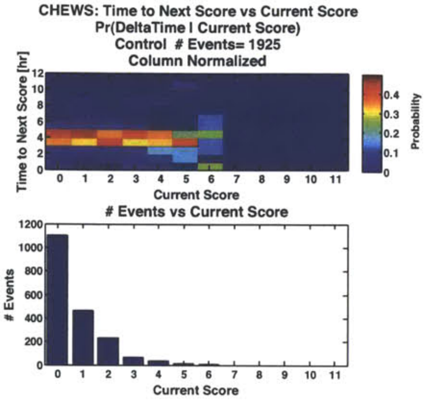


Figure 3.16: Control: CHEWS score evaluation frequency

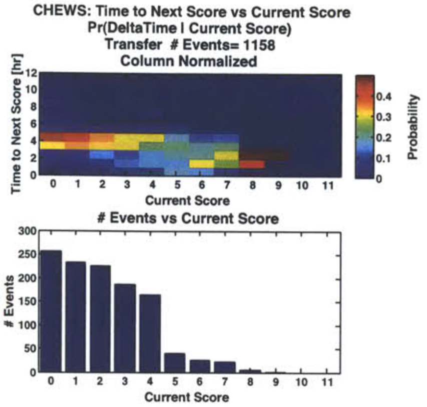


Figure 3.17: Transfer: CHEWS score evaluation frequency

underlying staff concern. In the top panels, the abscissa shows the last given CHEWS score and the ordinate shows the time until the next CHEWS score evaluation. These scores were only considered for the period up to 48 hours prior to transfer.

For control and transfer patients, there does not appear to be any difference in monitoring frequency when scores are less than 5. CHEWS scores are consistently calculated every 4-5 hours. However, for transfer patients with scores greater than or equal to 5 the interval between evaluations lessens.

3.7 Vital Sign Trajectories

The vital signs described in Chapter 2 are recorded for early detection of impending cardiopulmonary decompensation. One feature this thesis seeks to leverage in classifying patients and predicting transfer is whether or not significant trends are present in the measured vital signs prior to transfer to the ICU.

For the case shown in Figure 3.18, heart rate shows a gradual, though distinct, decline over the thirty-hour period prior to transfer. This decline would probably not be noticed over a five-minute observation window. Furthermore, spot checks at any given

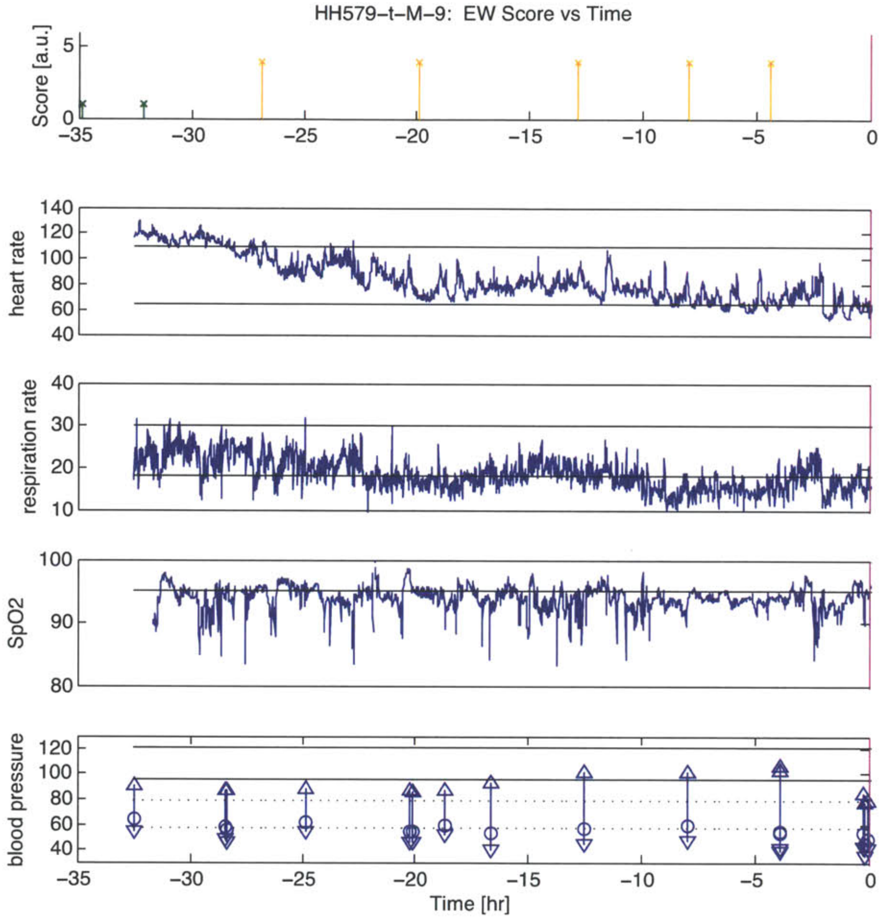


Figure 3.18: Vital signs and CHEWS score of a nine-year-old male patient on the general floor. Units for the bottom four subplots are respectively: beats per minute, breaths per minute, oxygen hemoglobin saturation percentage, and millimeters mercury.

moment will actually show that the heart rate is within normal bounds. However, the declining heart rate could be a serious sign of patient decompensation. The subplots also reveal a low respiratory rate for the patient age, which could have contributed to the poor oxygenation. At time $t = 0$, the patient was transferred to the ICU.

In contrast, control patients are those who are not transferred to the ICU. Their vital signs tend to exhibit an essentially steady trajectory, as seen for example, in Figure 3.19. They can also show some variation outside the age-adjusted bounds of normality.

Figure 3.20 also shows a control patient, yet the respiration rate frequently dips below normal for the age of the patient and the SpO_2 shows numerous dips below 95%. She also swings from slightly tachycardic to slightly bradycardic, and she even has a CHEWS of 4 at one point. Despite these indicators, the patient is not transferred to the ICU. These plots serve to illustrate that the classification of transfer and control patients is not a trivial task that can be adjudicated on the basis of simple threshold crossings.

■ 3.8 Bisected Changes Over Time

When the heart rate and respiration rate for transfer patients are each plotted against themselves over the course of a bisected window with regions [-24, -20] and [-4, 0] hours, it is startling that essentially all patients lie on or close to the forty-five degree line. This is seen in Figures 3.21 and 3.22, in which we have normalized heart rate and respiratory rate to the BCH age-appropriate normal value. This means that for these patients, their state (at least as evidence in heart rate and respiration) is not evolving much with time: if they have a high heart rate or respiratory rate for the first time block, then they will also have a similar high heart rate or respiratory rate right up until transfer.

A second unexpected observation from these plots is the correlation of high nor-

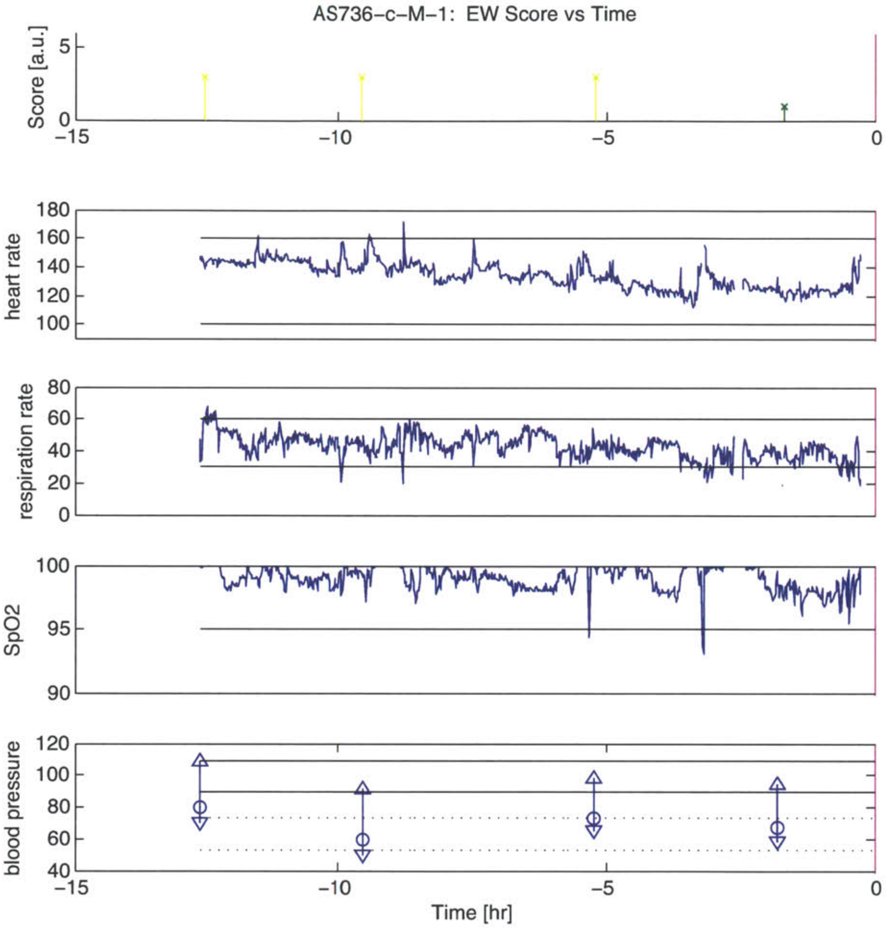


Figure 3.19: Control patient. Vital signs and CHEWS score of a one-year-old male patient on the general floor. Note the stability of the vital signs within normal ranges (horizontal lines). Units for the bottom four subplots are respectively: beats per minute, breaths per minute, oxygen hemoglobin saturation percentage, and millimeters mercury.

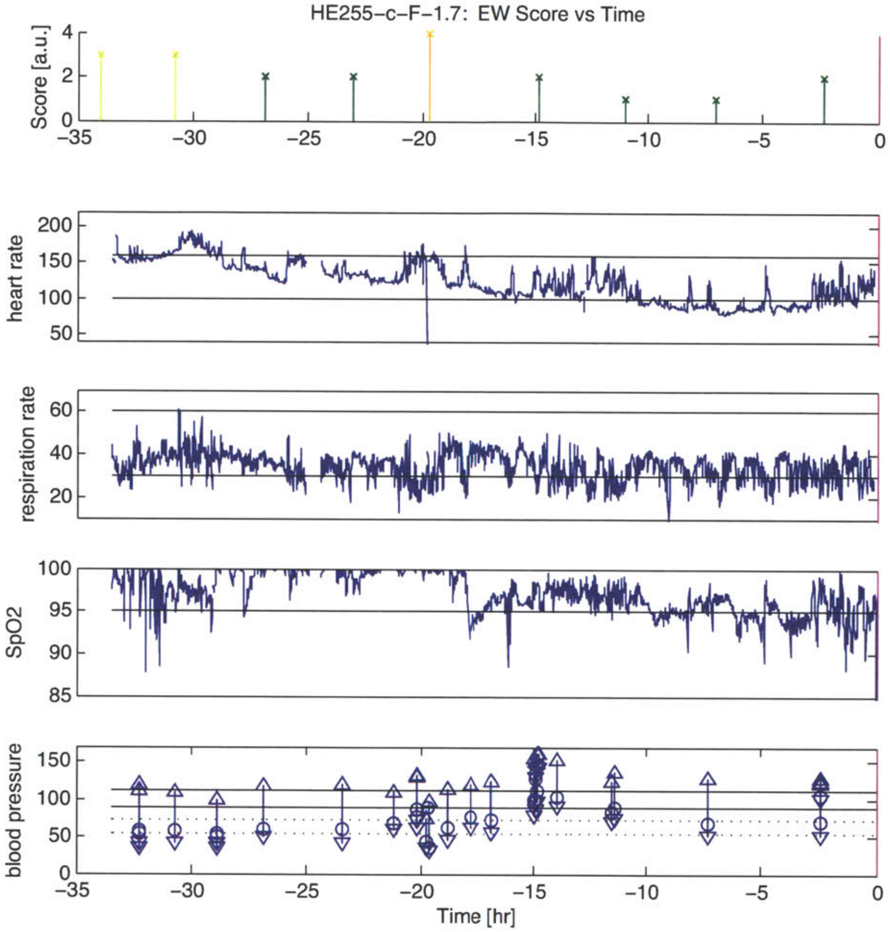


Figure 3.20: Control patient. Vital signs and CHEWS score of a 20 month-old female patient on the general floor. Note the significant vital sign deviations from normal ranges (horizontal lines). Units for the bottom four subplots are respectively: beats per minute, breaths per minute, oxygen hemoglobin saturation percentage, and millimeters mercury.

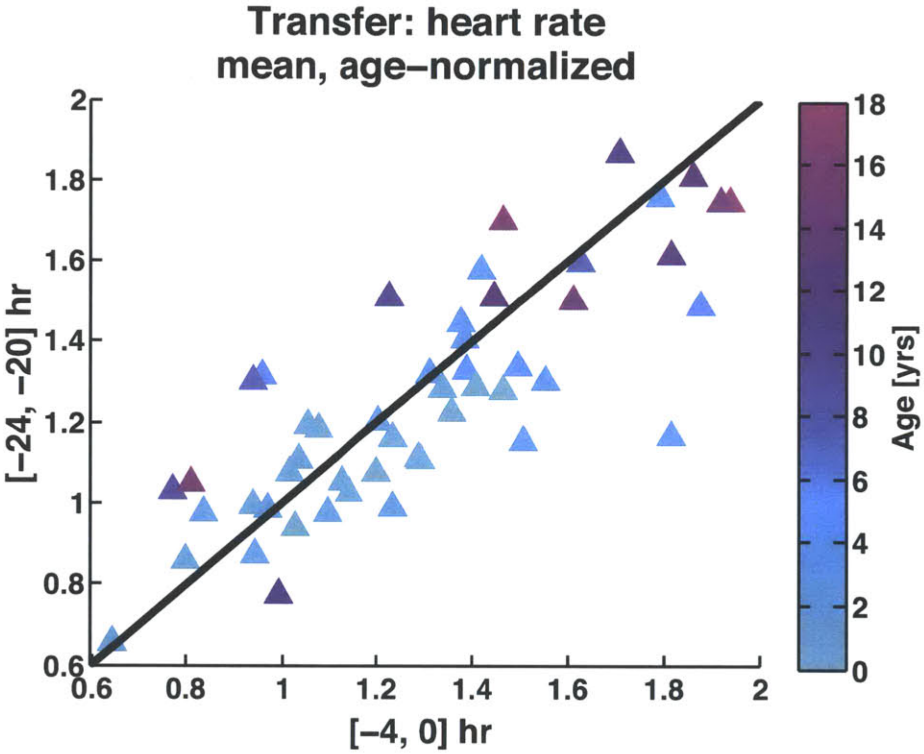


Figure 3.21: A comparison of the age-normalized, mean heart rate for transfer patients. The ordinate and abscissa each display the mean heart rate taken over the designated windows of time relative to call time (t=0) normalized to 1.0. Each triangle represents a transfer patient colored by age. The coloring shows that the cluster of patients with HR well above normal are also some of the oldest patients. The diagonal line is a visual reference that draws out the essentially static character of the mean heart rate over nearly a day.

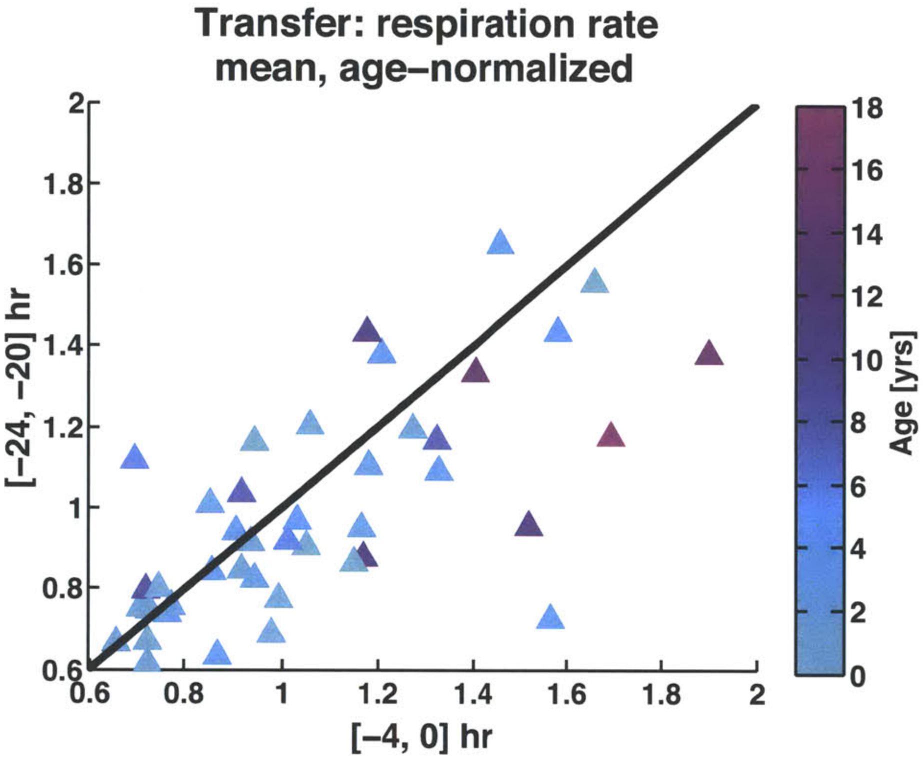


Figure 3.22: A comparison of the age-normalized, mean respiration rate for transfer patients. The ordinate and abscissa each display the mean RR taken over the designated windows of time relative to call time (t=0) normalized to 1.0. Each triangle represents a transfer patient colored by age. The coloring shows that the cluster of patients with RR well above normal are also some of the oldest patients. The diagonal line is a visual reference that draws out the essentially static character of the mean RR over nearly a day.

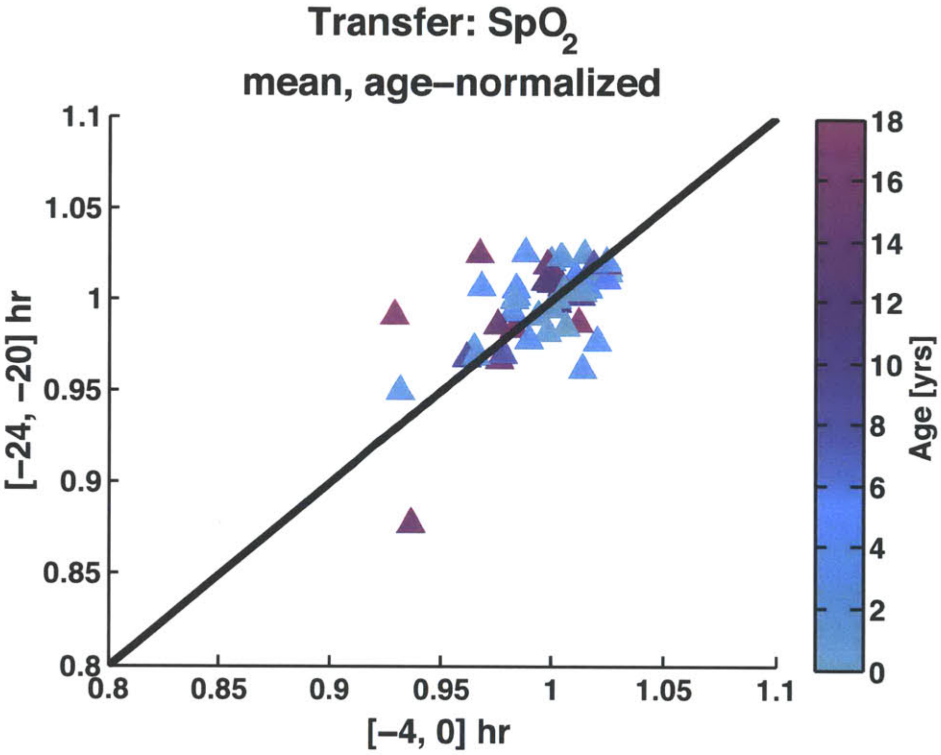


Figure 3.23: A comparison of the normalized, mean SpO_2 for transfer patients. The ordinate and abscissa each display the mean SpO_2 taken over the designated windows of time relative to call time (t=0) normalized to 1.0. Each triangle represents a transfer patient colored by age. With only three outliers, the mean SpO_2 value is almost always kept within normal limits and does not appear to be a valuable early warning sign.

malized exceedence with age. Figures 3.21 and 3.22 reveal that patients who are older are more likely to have sustained high heart rates and/or high respiration rates. Two hypotheses exist for this occurrence. First, younger children, and especially newborns, have less reserve capacity for their heart rate and respiration rate since the normal values are already high. Second, there may be a nursing permissiveness in which high heart rates and respiration rates are allowed to persist for older patients, but not for younger ones.

Other vitals signs, such as (proportional) total peripheral resistance, (proportional) cardiac output, and SpO_2, were also considered. Values proportional to TPR and cardiac output were computed using the models from Chapter 2, and their values were normalized to the first point of each patient's time series. Estimated cardiac output also showed a linear relationship but it was less strong than HR and RR. TPR showed a strong linear relationship. Normalized SpO_2 is shown in Figure 3.23, where 1.0 corresponds to 97.5% regardless of age. Essentially all transfer patients have healthy mean levels of oxygenation right up until transfer, so this aspect of SpO_2 may not be a good early warning sign.

■ 3.9 Human Classification Performance

The performance of CHEWS as a discriminator is qualitatively evident in the divergence of the transfer and control CHEWS mean values as the transfer patients approach their call time (Figures 3.7 and 3.8). However, the CHEWS score includes more information than is available through the bedside monitoring data and is subject to clinician bias, which is seen in the CHEWS underscoring phenomenon. Therefore, a retrospective "human expert test" was performed on vital-sign records from our database. A total of six clinicians, researchers, and nurses were asked to classify patients into transfer and control solely on the basis of the vital-sign records of HR, RR, SpO_2, and intermittent mean arterial blood pressure. This evaluation provided insights into how well human experts can classify patients and revealed features that clinicians considered discriminating.

The author randomly selected ten control and ten transfer vital-sign records from the cohort, and included up to 48 hours worth of data if that much data existed for a particular patient. For transfer patients, the record extended backwards from call time, and for controls, the record extended back from the end of the record. The heart rate, respiration rate, and SpO_2 were shown at a one-minute sampling resolution. The blood pressure was provided when available. Meta data included patient age, gender, and the BCH CHEWS scoring rubric, though the actual CHEWS scores were withheld. Reference ranges for age-adjusted vital-sign values were also provided. The normal values were taken from the BCH rubric. The six evaluators included four BCH clinicians and two MIT personnel with clinical research experience. The evaluators did not know how many control and transfer records were included in the test set of twenty records.

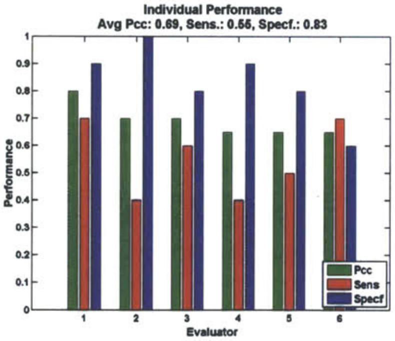


Figure 3.24: Self test performance by evaluator, ordered by decreasing probability of correct classification.

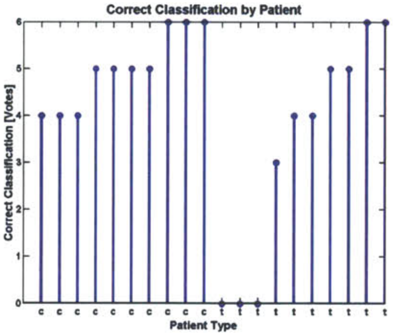


Figure 3.25: Self test performance by patient.

Performance on an individual basis is shown in Figure 3.24. Probability of correct classification (Pcc), sensitivity, and specificity are shown for individual evaluators. The

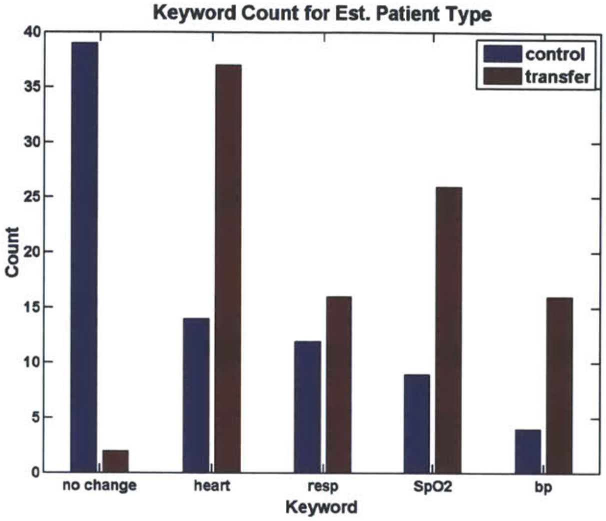


Figure 3.26: Justification key words

average Pcc of 0.69, sensitivity of 0.55, and specificity of 0.83 suggest humans can identify controls well but frequently miss transfer patients. Figure 3.25 shows performance as a function of the patient type (control = c, and transfer = t). The abscissa labels are the true patient type, and the ordinate shows the number of votes for the true patient type. This chart shows significant agreement among all evaluators even when the evaluators are wrong. By using a majority rule fusion of votes, the Pcc, sensitivity, and specificity can be boosted to 0.80, 0.60, and 1.0 respectively. The fused performance is on par with the best individual evaluator.

All evaluators were asked to write a one-sentence text justification of their decision for each record. Using a basic form of natural language parsing, justification features of "no change, heart rate, respiration rate, SpO_2, and blood pressure" were extracted from the responses. The number of times a feature occurred for a declared patient type (regardless of whether the decision was correct) were tallied and are shown in Figure 3.26. The figure clearly shows that when evaluators are declaring patients as controls, the dominant feature is "no change" in the patient's status. This feature, taken with the good specificity of the study, suggests steady state stability is a clear marker of a control patient. If the patient is declared a transfer patient, specific vital signs are called out. In order of frequency, these are heart rate (e.g. tachycardia), SpO_2 (e.g. SpO_2 desaturations), and blood pressure (e.g. hypotensive). Respiration rate is only occasionally noted and appears equally for both controls and transfers.

■ 3.10 Data Exploration Summary

Data exploration has suggested several findings that will shape the questions posed by and feature extraction approaches of subsequent chapters. Control patients tend to have stable vital-sign trend trajectories while transfer patients tend to exhibit larger variations. However, large trends are present for both patient groups, so simple threshold crossings may not distinguish between the two. Furthermore, a comparison of transfer patient vital signs a day before transfer to four hours before transfer shows a strong concordance between the two average values for heart rate and respiration rate. This suggests that trends may not be a dominant feature of transfer patient vital-sign trajectories. Analysis of the *average* CHEWS score shows a divergence between the transfer and control populations even as much as 48 hours before transfer. However, a detailed investigation reveals substantial variability of scores, especially within the transfer group, so the CHEWS score's predictive utility may be limited. Finally, the human classification test showed that humans overall are specific (0.83) but not especially sensitive (0.55), so the benchmark probability of correct classification is only 0.69.

Chapter 4

Classification and Prediction

The previous chapter ended with an informal human classification performance test. In this chapter, we first explore automated classification and then automated prediction. We compare the performance between the BCH CHEWS score and a reduced CHEWS (rCHEWS). The rCHEWS uses the bedside monitoring data to which a machine has access; the data is a strict subset of the information that goes into the CHEWS.

In our classification problem, an observed data set must be categorized as either belonging to a control patient or a transfer patient. We will discuss this classification problem first in order to tie back to the human performance in Chapter 3 and to discuss common obstacles and techniques that apply to both classification and prediction. In our prediction problem, we must decide if a patient is going to be transferred and localize the transfer event in time. After presenting a clinically meaningful evaluation metric, we will discuss automatic prediction performance.

In this thesis, the classification problem and the prediction problem both require making a binary decision: given the observed data set, we must decide if the patient is a control patient or a transfer patient. While the fundamental problem is the same in both situations, we will focus on classification first to provide a concrete context for terms that apply to both situations.

The information in the first two sections applies to both classification and prediction. Section 4.1 will define feature vectors and explain our approach to feature selection. Section 4.2 will introduce the concept of decision rules and explain two particular rules that will be used in this chapter. In Section 4.3, we will apply the decision rules introduced in Section 4.2 to the problem of classifying patients based upon their CHEWS or rCHEWS scores. Sections 4.4 and 4.5 advance the chapter from discussing classification to prediction. The former section describes our approach to evaluating a prediction metric, and the latter applies the metric to CHEWS and rCHEWS features. Section 4.6 summarizes the results of the chapter.

■ 4.1 Feature Selection

In both classification and prediction, we will frequently refer to feature vectors. A feature vector is a sequence of numbers that are some abstraction of the observed data. As a specific example, the feature vector may be a 3-by-1 vector whose three elements are the three most recent CHEWS scores. As another example, the feature vector may be a 10-by-1 vector whose elements are the instantaneous heart rate at ten different points in time. A feature vector could also have components of different types, e.g., both CHEWS scores and heart rates.

A fundamental issue is how many and what kinds of features should be chosen. In the context of our work, a feature vector can be considered to exist in a two-axis space. The first axis is the temporal axis, and the second axis is the measured physiological data. These two axes refer to how finely the data is sampled to create the feature vector. At the coarsest level, one can begin on the temporal axis by using only the most recent early warning score, and on the measured physiological parameter axis by using the CHEWS score as the early warning score. Features can be added along these two dimensions. For example, on the temporal axis, not just the most recent but several recent CHEWS scores could be used. On the feature axis, one could unmask oneself to the respiratory health and heart rate health subscores. An even deeper level on the feature axis would be how the individual respiratory rate and SpO_2 category scores are combined. A still deeper level would be changing the quantization limits for each of the heart rate, respiration rate, and SpO_2 category scores. At each level, the feature vector increases in dimensionality, which conceivably allows better performance on a training set but potentially worse performance during an actual performance test. As dimensionality increases, a user can begin to lose touch with the underlying classification process. The principle of parsimony applies: the simpler model should be preferred among two models that have similar performance.

At first glance, the CHEWS score itself only has a temporal dimension. However, while the score cannot be changed, it can be conditioned on the patient population characteristics such as age, gender, and/or transfer reason. An automated score can go all the way from replicating aspects of the BCH CHEWS score using the trend data to incorporating features from the waveform data.

Our strategy will be to begin at the coarsest level (lowest dimensionality), and increase the dimensionality as needed.

■ 4.2 Decision Rules

A decision rule is a mapping of each feature vector to a class. Figure 4.1 shows two classes (black and white circles) in a two dimensional feature space. Several possible decision rules are represented by the solid, dotted, and dashed lines. If the solid line were the final decision rule, then all feature vectors that were above and to the right of the solid line would be labeled as one class, and all other feature vectors would be labeled as the second class. The following two subsections will provide details on two kinds of decision rules used in this thesis and a discussion of decision rule complexity. However, they are not necessary to understand the basic thrust of the results in subsequent sections.

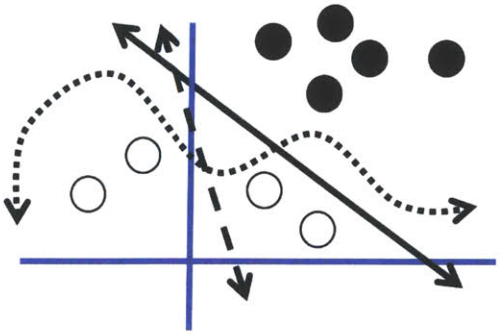


Figure 4.1: Several possible decision rules (solid, dashed, dotted lines) for partitioning the two-dimensional feature space. The feature vectors are circles, and the two classes are distinguished by the color of the circles. Using the solid black line as a decision rule, all feature vectors to its upper right would be one class while all other feature vectors would be the second class. The solid line allows perfect class separation unlike the straight dashed line, which groups some of the white circles with the black circles.

■ 4.2.1 MAP Rule

Basic probability theory provides an intuitive classification strategy that minimizes the probability of an erroneous classification given a specific feature vector $\mathbf{x}_0$. We start with a discrete probability mass function (PMF) $p(\mathbf{x}|y)$ where y is the patient type ($y \in \{\text{control}, \text{transfer}\}$), and $\mathbf{x}$ is the feature vector. This PMF can be estimated during a training phase, given example feature vectors from each patient type.

In the classification step, given a single feature vector $\mathbf{x}_0$, from a patient of unknown type, the patient can be classified by choosing the patient type $\widehat{y}$ that maximizes the *a posteriori* PMF. This decision rule is known as the *Maximum a Posteriori* (MAP) rule [54]:

$$\widehat{y} = \underset{y}{\operatorname{argmax}}\, p\left(y|\mathbf{x}_0\right) \tag{4.1}$$

The MAP rule maximizes the probability of a correct decision [54]. Using Bayes's rule, Equation 4.1 can be rewritten in terms of the estimated PMF and the prior probability of a patient being a control or transfer patient, $p(y)$.

$$\widehat{y} = \underset{y}{\operatorname{argmax}} \frac{p(\mathbf{x}_0|y)p(y)}{\sum_{y'} p(\mathbf{x}_0|y')p(y')} \tag{4.2}$$

The denominator does not affect the argmax calculation as it is merely a normalization factor; it is only shown to illustrate the explicit rewriting of Equation 4.1 with Bayes's rule. If the prior probabilities are equal, the MAP rule is equivalent to maximizing the likelihood $p(\mathbf{x}_0|y)$ of the specific feature vector over the possible patient types.

While the MAP strategy holds for feature vectors of arbitrary dimension, where each element can take an arbitrary number of values, the state space from which $\mathbf{x}$ is drawn grows exponentially with the dimension. For example, assume classification is done on an N-dimensional feature vector $\mathbf{x}$, where each element can take K values, and the number of classes for categorization is M. Then the number of elements needed in each of the M PMFs is K^N. If we consider a feature vector that only uses the most recent CHEWS score, then $N = 1$; $K= 12$ because CHEWS ranges from 0 to 11. In estimating a histogram, a rule of thumb is to have 10 samples per bin, so already $12^1 \cdot 2 \cdot 10 = 240$ patients are needed to train this model.

MAP may not be a practical rule to actually implement. As with any decision rule, MAP may grossly overfit the data, which can lead to rules that do not generalize well. In other words, a low classification error can be obtained on the data set from which the model was trained. However, the model will have a high error when applied to new data. Overfitting can also lead to counter-intuitive decision rules, which will be discussed in Section 4.2.3.

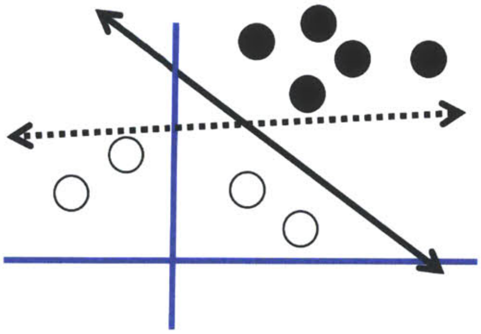


Figure 4.2: Two-dimensional feature space with two classes of feature vectors (black and white circles). The SVM boundary would be the solid black line. While the dotted line also is a linear separator that perfectly separates the classes, it is not the SVM boundary because another boundary exists that has a larger minimum distance to the feature vectors.

4.2.2 Support Vector Machines

Support vector machines (SVMs) construct a linear boundary in the feature space that maximizes the distance to the closest points in each of the two classes. An example of a notional SVM boundary in a two-dimensional feature space is shown in Figure 4.2. The boundary is specified by the implicit equation

$$\mathbf{w}^T\mathbf{x} + w_0 = 0. \tag{4.3}$$

where $\mathbf{w}$ and w_0 are weights that define the normal vector to the decision boundary.

The geometric margin for each point is the perpendicular distance of the point from the decision boundary. The goal is to maximize the minimum geometric margin among all the points subject to all the points being classified correctly. For now, we assume that the data can be perfectly separated. Then the optimization problem becomes [55]

$$\{\widehat{w_0}, \widehat{\mathbf{w}}\} = \underset{w_0,\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2 \tag{4.4}$$

subject to the constraints that for all points $\mathbf{x}_i$ in the training set

$$y\left(\mathbf{w}^T\mathbf{x}_i + w_0\right) \geq 1 \tag{4.5}$$

where y $\in \{-1, 1\}$. Equation 4.4 has a single local minimum that can be found by standard quadratic solvers.

Despite the straightforward computational problem, solvers can return weight vectors with small weights. To enforce reasonable sparsity in the solution, we can change the tolerance, *tol*, limit for our solver. Weights smaller in magnitude than *tol* are forced to zero. We will use a *tol* parameter in our SVMs.

Once the weights are found, an unknown point $\mathbf{x}_0$ is classified based on the sign of its margin:

$$y = \text{sign}(\widehat{\mathbf{w}}^T\mathbf{x}_0 + \widehat{w_0}) \tag{4.6}$$

In other words, a functional margin of zero acts as the decision boundary.

Unfortunately, the SVM problem as stated has no solution if no linear boundary can separate the classes [55]. To use an SVM with data that cannot be perfectly partitioned, the optimization problem is augmented to be

$$\{\widehat{w_0}, \widehat{\mathbf{w}}\} = \underset{w_0,\mathbf{w}}{\text{argmin}}\frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\zeta_n. \tag{4.7}$$

subject to the constraints

$$y_n\left(\mathbf{w}^T\mathbf{x}_n + w_0\right) \geq 1 - \zeta_n, \tag{4.8}$$

$$\zeta_n \geq 0. \tag{4.9}$$

The slack variables ζ_n reflect the distance of the misclassified training points from the boundary as well as the failure of some correctly classified points to not be sufficiently far away from the boundary. If $0 < \zeta_n < 1$, then $\mathbf{x}_n$ is correctly classified but incurs a penalty in the optimization, while if $\zeta_n \geq 1$, then $\mathbf{x}_n$ is misclassified and also incurs a penalty.

C is a tunable constant, chosen through cross validation, that is a regularization term. It controls the tradeoff between the margin associated with the decision boundary and the number of misclassified points. By letting C approach infinity, the hard margin classifier in Equation 4.4 is recovered. Larger values of C lead to better fits on training data but the resulting model may not generalize as well as a model with lower values of C. A small value of C allows some points to be misclassified in exchange for placing the decision boundary farther from the two main clusters of points from the different classes.

By reformulating the optimization problem into an alternate form called the "dual" form, SVMs permit an extremely flexible implicit feature vector expansion through a function called a kernel. A polynomial kernel that is quadratic can automatically expand a base set of features into all their squared and cross product terms. A linear SVM can then be fit in the expanded feature space, but when the linear SVM boundary is viewed in the original base feature space, it will appear as a non-linear decision boundary [55]. An example of classification with a quadratic kernel is shown in Figure 4.3.

Just as with MAP, the great potential flexibility afforded by SVMs can also lead to overfitting the data if the feature vector is expanded to high enough dimensions. Therefore, we will limit the order of our SVM kernels in our work.

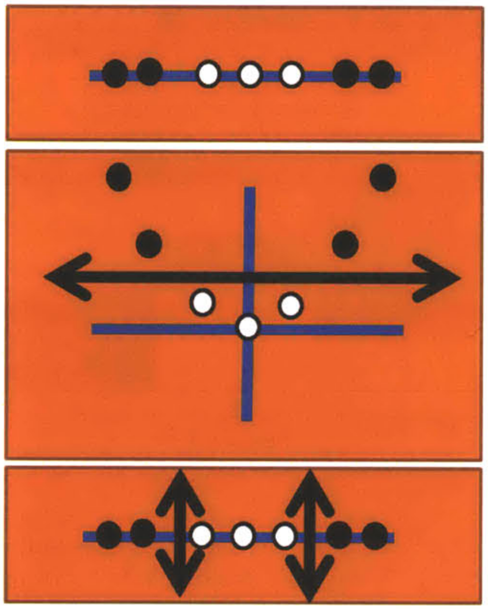


Figure 4.3: Top: One-dimensional feature space with two classes of feature vectors (black and white circles) each with value x_0. Middle: Two-dimensional feature space obtained by creating a new feature vector $[x_0, x_0^2]$. The SVM boundary would be the solid black line. Bottom: The linear SVM boundary from the two-dimensional space as it would appear in the original one-dimensional space.

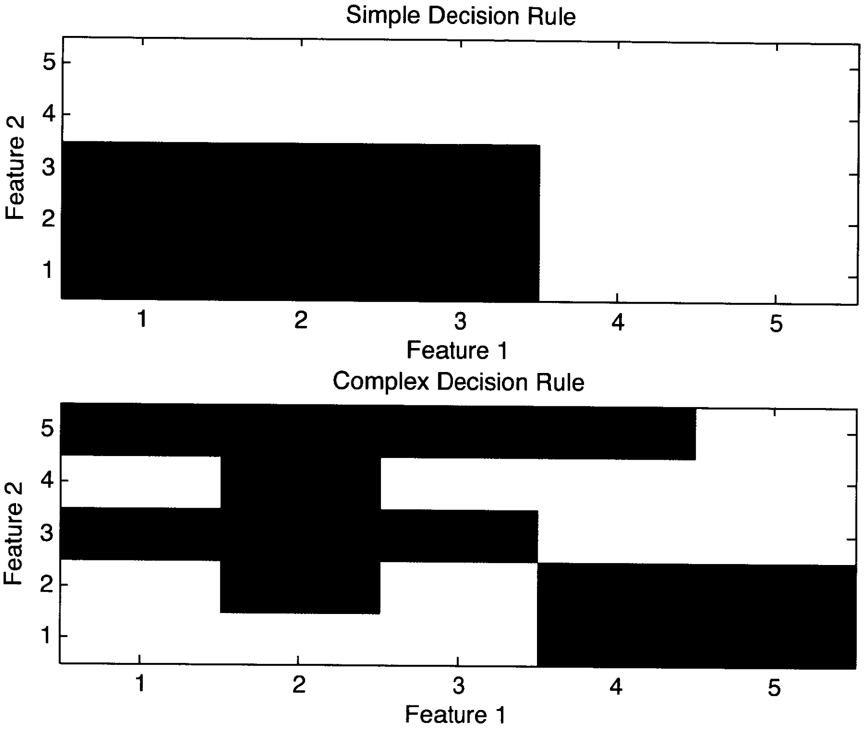


Figure 4.4: A simple decision rule (top) that may generalize well when deciding between two classes (black and white). A complex decision rule (bottom) that may overfit the data.

■ 4.2.3 Decision Rule Complexity

Figure 4.4 shows two notional decision rules based on two features in a feature vector. The possible values each feature element can take are along the axes, so a feature vector maps to a square in the grid. The feature vector belongs to the patient type determined by the square's color.

The MAP rule may give rise to a complex decision rule because it makes the optimal decision with the given feature vector, irrespective of the patient types that go with small perturbations of the feature vector. The decision can change several times instead of just once when reading along any row or column in a 2D decision graph. A decision rule with multiple decision changes might suggest that features were not optimally chosen. The chosen features might not be in accordance with physiology where we expect a continuum of responses. For example, if a patient is a transfer patient and has a score of 3, whether by CHEWS, rCHEWS, or some other method, then patients with higher scores than 3 should also be transfer patients.

Ideally, a decision rule should separate the patient types into two distinct but internally connected clusters. Heuristics that violate the MAP rule can enforce a two-cluster requirement. As this thesis progresses into using more than two dimensions in the feature vector, it will no longer be possible to visualize a natural boundary between two clusters.

■ 4.3 CHEWS and rCHEWS Classification

We begin by examining classification of control and transfer patients using only the CHEWS score, and then turn to the use of the rCHEWS score. For transfer patients, the window considered (observation window) ends at the call time and extends backwards up to 48 hours, provided data is present. For control patients, the observation window could start at an arbitrary location and extend back up to 48 hours, provided data is available. Typical control patient time points could be the last CHEWS score or the end of the record. Here, the time of the last CHEWS score will be used as the analog of call time for control patients.

The current CHEWS system is a memoryless system. Previous scores are not considered explicitly when deciding if a patient should be transferred. As pointed out, patient history is taken into account by the clinical staff but that process is not captured in the CHEWS algorithm.

Our problem then is to classify patients as either control or transfer patients, using only the patient's history of CHEWS in the last 48 hours. Our objective is a probability of correct classification (Pcc) of 80 percent, which would be on par with the human classification performance in Chapter 3. We also set ourselves the additional challenge of a Pcc of 80 percent using 48 hours of data that ends not at the call time but six hours prior to the call time. The six hour buffer is called the warning time (WT). The warning time is the minimum interval caregivers have, if a transfer event is predicted, in which to take preventative action. Six hours is chosen for its medical significance based on conversations with BCH clinical staff.

Only patients for whom good trend data is present and who have CHEWS scores recorded are evaluated. While the trend data plays no role in this section for classification, by only using patients for whom trend data is present, the data set can be reused for the next section, where rCHEWS is computed on the basis of the trend data and used for classification. Our set of 50 control and 50 transfer patients for whom good trend data is present is reduced to 40 control patients and 34 transfer patients.

The first subsection reviews performance using the MAP rule with several heuristically motivated feature vectors while the second reviews performance using SVMs of various complexities. For ease of comparison, we have placed the results from both subsections together in Tables 4.1 and 4.2.

■ 4.3.1 MAP Classification

We will consider all the patients in our dataset and retrospectively choose the decision rule for each one of the possible feature vectors to minimize the probability of error. In this sense, we will present results using MAP classification as an upper bound on performance. When applied this way, the MAP rule loses connection to the probabilistic framework from which it comes, so "MAP" as used in this thesis is a misnomer. A better name might be the "Omniscient Rule."

We consider three feature vectors. The first feature vector uses only the most recent CHEWS score; it best emulates current BCH practice. The second feature vector uses the two most recent CHEWS scores, and the third feature vector uses the three most recent CHEWS scores. Because CHEWS are generally collected at four-hour intervals, a history of three CHEWS can allow trend identification over approximately one nursing shift.

■ 4.3.2 SVM Classification

We used support vector machines that are trained and tested using K-fold cross validation. The resulting Pcc of each SVM's performance is averaged across the K trials. K$-$1 partitions of the data are used for training and the Kth fold is used for testing. The training and testing is done K times in order to test on each of the partitions. We chose K=3 to balance rounds of testing with a sufficient number of patients during each round.

Four different SVMs were considered. The first two use the two most recent CHEWS scores; one uses a linear separator while the other uses a quadratic separator. They are called "SVM: Lin. 2 Latest" and "SVM: Quad. 2 Latest", respectively. The other two SVMs use the three most recent CHEWS scores and the maximum score among the three latest scores, respectively "SVM: Lin. Max & 3 Latest" uses a linear separator and "SVM: Quad. Max & 3 Latest" uses a quadratic separator. The maximum score over a window was motivated by the hypothesis that if at some point a patient crosses some severity of sickness, even if they appear to get better, ultimately they might end up in the ICU.

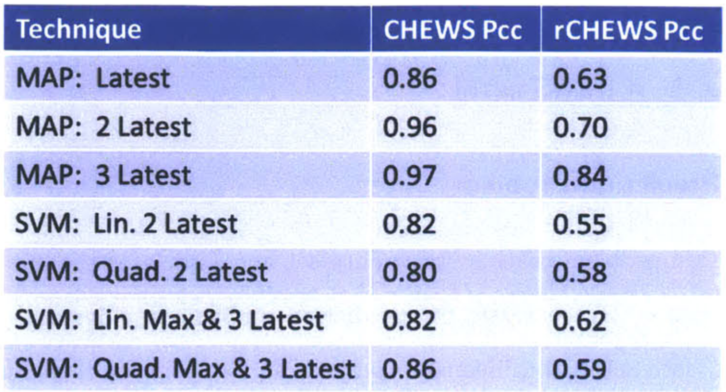
■ 4.3.3 rCHEWS Classification Method

The rCHEWS procedure matches that of the CHEWS process, except for one difference related to the sampling times. Trend data are available at the minute level, so an instantaneous rCHEW can be computed every minute. Because we desire initially to emulate a CHEWS application, we downsample the trend data to four-hour intervals. The most recent score is taken to be one hour before call time for transfer patients and one hour from the record end for control patients. Unlike CHEWS, where the most recent point may or may not be close to the call time and the spacing between points may be uneven, the rCHEWS is always computed at the specified time points. Because the trend data and consequently the computed rCHEWS is noisy, we first filtered the trend data with a ten-minute median filter. Additionally, once the final rCHEWS was computed, we filtered the rCHEWS sequence with a four-hour averaging filter. This filter performs an integration of the signal, so after downsampling, information about rCHEWS behavior between four-hour samples is encapsulated in the four-hour samples.

■ 4.3.4 Classification Results

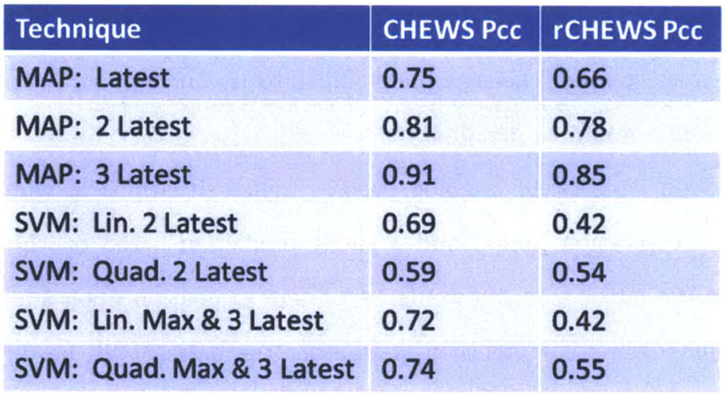
Tables 4.1 and 4.2 present the CHEWS and rCHEWS classification results for a WT of 0 and 6 hours. As expected, overall performance as judged by probability of correct classification is better at WT=0 hours than WT=6 hours, which suggests that discriminating information is most obvious the closer transfer patients are to call time. Classification performance using the MAP rule steadily improves as more features are added, but the performance is largely determined by only a single score with additional scores improving performance but significantly adding to complexity. The classification performance with CHEWS using WT=0 compares favorably to the performance of the human experts, as Pcc > 0.80.

In summary, a simple linear SVM that uses the two most recent CHEWS can perform nearly as well as the upper bound given by MAP classification at a WT of 6 hours. Therefore, using complex decision rules that attempt to exploit trends in the CHEWS history does not appear beneficial. Unfortunately, Pcc using the rCHEWS is about



Technique	CHEWS Pcc	rCHEWS Pcc
MAP: Latest	0.86	0.63
MAP: 2 Latest	0.96	0.70
MAP: 3 Latest	0.97	0.84
SVM: Lin. 2 Latest	0.82	0.55
SVM: Quad. 2 Latest	0.80	0.58
SVM: Lin. Max & 3 Latest	0.82	0.62
SVM: Quad. Max & 3 Latest	0.86	0.59

Table 4.1: MAP and SVM classification with WT=0 hours on 40 control and 34 transfer patients. The same patients were used for both CHEWS and rCHEWS classification. Performance evaluation for the MAP rule was done without cross-validation, and the decision rule for each feature vector was chosen retrospectively to provide the highest possible Pcc. Different linear and quadratic SVMs were used and evaluated with 3-fold cross-validation.



Technique	CHEWS Pcc	rCHEWS Pcc
MAP: Latest	0.75	0.66
MAP: 2 Latest	0.81	0.78
MAP: 3 Latest	0.91	0.85
SVM: Lin. 2 Latest	0.69	0.42
SVM: Quad. 2 Latest	0.59	0.54
SVM: Lin. Max & 3 Latest	0.72	0.42
SVM: Quad. Max & 3 Latest	0.74	0.55

Table 4.2: MAP and SVM classification with WT=6 hours on 40 control and 34 transfer patients. The same patients were used for both CHEWS and rCHEWS classification. Performance evaluation for the MAP rule was done without cross-validation, and the decision rule for each feature vector was chosen retrospectively to provide the highest possible Pcc. Different linear and quadratic SVMs were used and evaluated with 3-fold cross-validation.

equal to chance. Based on these results, it appears that the CHEWS captures significant information that is not present in the automatically collected data when the data is analyzed using the rCHEWS metric.

■ 4.4 The Prediction Problem

The classification problem takes a static data set and associates the data with one of two patient groups. By contrast, the prediction problem is a dynamic classification problem. The data set is continuously updated, and at periodic intervals the patient must be declared to belong to the transfer or the control group.

A performance metric for a prediction algorithm is a necessary part of the research process. However, many early warning systems neglect that component, which undermines their purported benefits [17]. Perhaps part of the reason a metric is not reported is that the prediction is difficult to quantify. Ideally, an early warning score should predict with reasonable lead time and high sensitivity and specificity whether or not a patient will need to be transferred from the floor to the ICU. Quantitatively this is ambiguous because "predict" needs to be qualified by how many hours ahead of transfer the warning should or can be given.

The prediction itself is an ongoing process unlike situations in which there is a once-and-done observation period and a single prediction based on the observation. An example of the latter scenario would be patient evaluation upon admission to the emergency department. The decision is whether the patient will be directed to the general floor or ICU based on a collection of data acquired upon admission [52]. Another example is the prediction of mortality after a 12- or 24-hour observation window that began as soon as the patient was admitted to the ICU [53, 56]. Instead, an early warning algorithm must constantly make decisions as time advances and new data becomes available. Consequently, a performance metric must account for the sliding window aspect of the prediction problem.

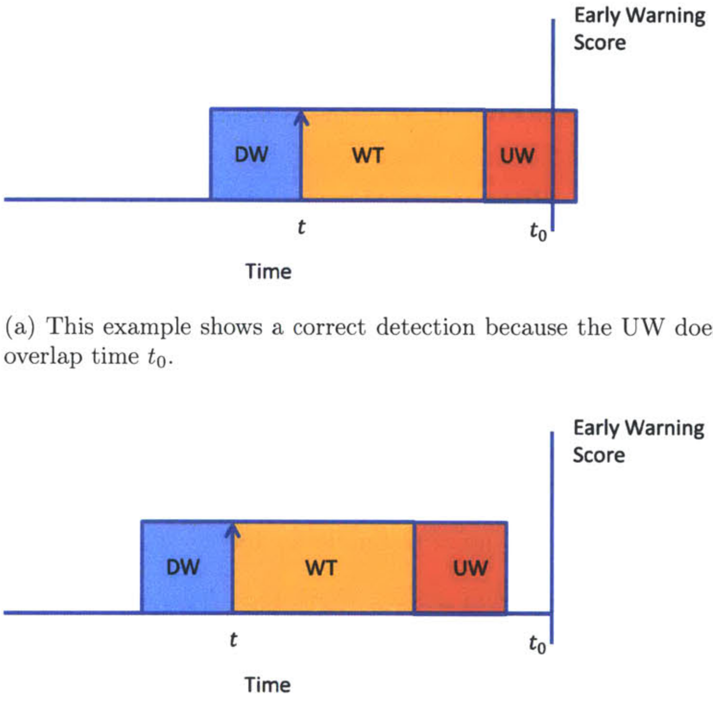
Section 4.4.1 begins by defining important segments of time in a data record. We will refer to these definitions during a literature survey of how other authors have approached the prediction problem in Section 4.4.2 and throughout the thesis. Sections 4.4.3, 4.4.4, and 4.4.5 will use the definitions from 4.4.1 to explain a prediction framework that concretely measures prediction performance. Section 4.4.6 uses this framework to understand some of the results presented in the literature survey.

■ 4.4.1 Data Window, Uncertainty Window, and Observation Window

Figure 4.5 illustrates the data window, uncertainty window, and warning time. Recall that WT is the minimum amount of notice caregivers have to take action if a transfer event is predicted. The data window (DW) is the length of data prior to time t that is used to make a prediction at time t. The maximum size of the data window is bounded by the length of time between the current moment, t, and the start time of the data, t_s. The time t is the right most point of the data window. The uncertainty (or event) window (UW) is the length of time during which the predicted event may occur after the warning time has expired.

Figures 4.5a and 4.5b schematically outline the prediction problem and how it relates to DW, WT, and UW. The vertical arrow at time t represents a threshold crossing, or the moment when a future transfer event is predicted. The fiducial time t_0 represents the time a call for transfer was made. The actual early warning score values are not shown. The WT and UW regions are schematically represented by colored boxes. Because the UW in Figure 4.5a includes t_0, the prediction situation is a correct prediction. In Figure 4.5b, the threshold crossing occurs prematurely, so UW does not overlap with t_0. Consequently, the prediction in Figure 4.5b is a false positive.

Figure 4.6 illustrates the notations of an observation window (OW) and a warning



(a) This example shows a correct detection because the UW does overlap time t_0.

(b) This example shows a false alarm because the UW does not overlap time t_0.

Figure 4.5: Schematic of warning time (WT) and uncertainty window (UW) regions relative to the threshold crossing (arrow) and call time for transfer to the ICU (time t_0).

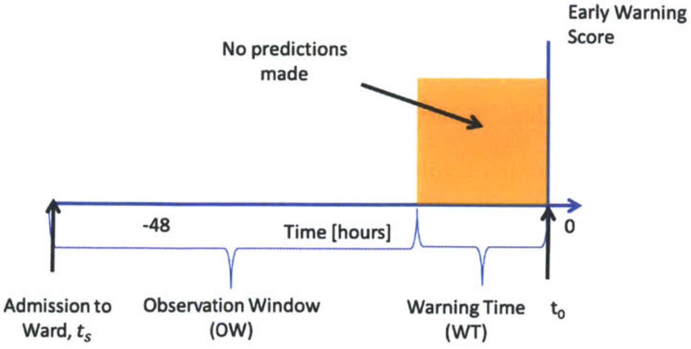


Figure 4.6: An observation window, and a buffer region (warning time) before the call time at t_0. During the warning time, predictions do not take place, and the goodness of the early warning score is not evaluated.

time (WT). The observation window (OW) is the data interval during which we are actively making transfer predictions. The interval begins as soon as a patient is admitted at time t_s, and continues up until no more predictions are to be made. Figure 4.6 shows how the OW is shifted back relative to t_0 because WT is greater than zero. No data is used from the WT, and no predictions are made. In order to have some measure of consistency among patient records of various lengths, we will truncate records that contain more than OW+WT hours of data to lengths of OW+WT. Then t_s becomes the start of the truncated record, and t_0 remains the right most point of the record.

The ideas of UW and WT actually appear in the seizure prediction literature. Winterhalder calls uncertainty window the seizure occurrence period, and warning time the seizure prediction horizon [57]. We will analyze his definitions of UW, WT, sensitivity, and specificity to evaluate performance from the receiver operating characteristic (ROC) point of view in Section 4.4.3

■ 4.4.2 Literature Survey

The early warning score that we have in mind is different than other severity of disease metrics such as the Simplified Acute Physiology Score (SAPS II) [58] and the Acute Physiology and Chronic Health Evaluation II (APACHE II) score [59]. The latter two compute a single score once after admission to the ICU. These single scores may be used to compare the morbidity of the patient to other patients, determine if certain procedures are warranted, or predict the patient's mortality. By contrast, an early warning score is a continuously evolving metric where not only the final outcome of the patient must be accounted for in the score (transfer or not transferred) but also the value of the score through time towards the end point.

Three early warning score studies involving pediatric populations show two approaches to the problem of quantifying prediction performance. The first by Akre *et al.* [18] considered early warning scores for pediatric patients before a rapid response team (RRT) or code blue (respiratory arrest) call was made. In this study, there were no control patients. Akre looked only at the 24 hours preceding the event right up to the event itself. Akre's OW was 24 hours and WT was 0 hours. Because no controls were present, no attempt was made to define specificity. For sensitivity, the maximum score during the previous 24 hours was compared to a threshold of four. If the maximum score was greater than or equal to the threshold, a "transfer" was declared. A transfer was also declared if a single domain score in the rubric was equal to three (the maximum domain score). Sensitivity was then defined as the ratio of the number of patients declared as transfer to the total number of patients. Akre reported a sensitivity of 85.5% [18].

There are several concerns with Akre's approach. The most obvious is the lack of a control group and the consequent inability to quantify the specificity of the approach. One suggestion for overcoming the lack of a control group is to use data from the

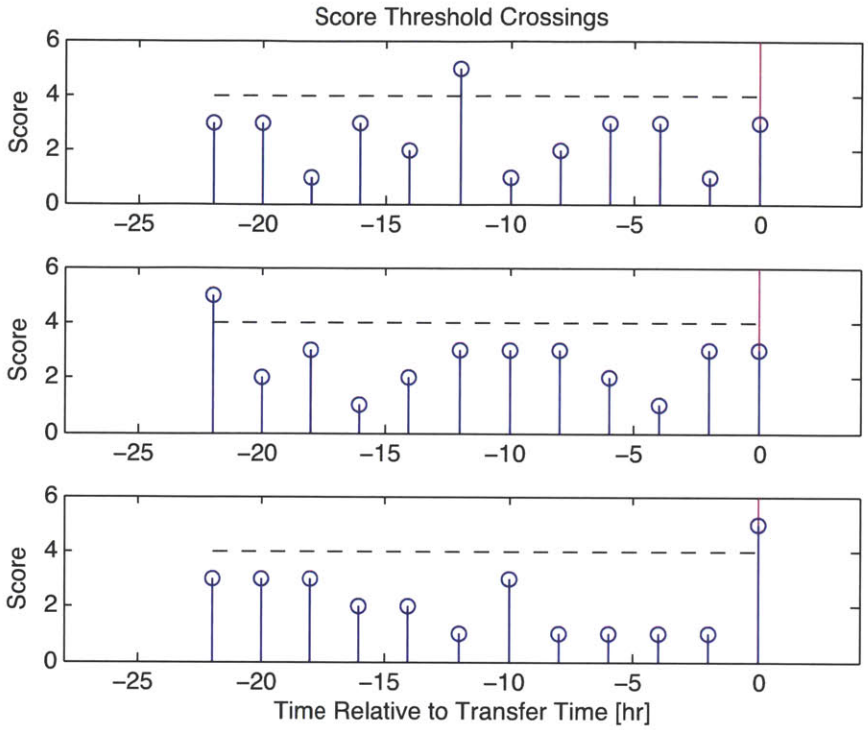


Figure 4.7: Three different threshold crossing locations all of which yield a sensitivity of 1.0 under Akre's definition [18].

transfer patients prior to the 24 hours before each is transferred. The transfer patients could act as self controls because any alarms earlier than 24 hours from the transfer time could be considered false alarms. Second, the clinical utility of this approach is questionable. Figure 4.7 shows three different scenarios all of which would yield a sensitivity of 1.0 but with very different intervals between the threshold crossing and the actual transfer time. When the latest score crosses the threshold, Akre's definition of sensitivity predicts a transfer anytime between the threshold crossing and 24 hours into the future.

Akre attempted to address the 24 hours of uncertainty surrounding a threshold crossing by calculating the median time of the earliest and of the latest threshold crossings. The median earliest instance was 696 minutes and the median latest was 30 minutes. The respective ranges were 5 to 1439 minutes and 1 to 1438 minutes [18]. However, "earliest" and "latest" determinations do not appear to map to a clinically actionable event. For example, if a patient checks into the floor and one hour later has a threshold crossing, does that mean they will have an event twenty three hours from that point or 1 minute from that point? Clearly a more precise prediction is necessary.

Two other studies, Duncan in 2006 [16] and Parshuram in 2011 [51], addressed some of the issues raised by Akre's approach. As in Akre, both of these studies included a transfer group, where a transfer was an unplanned transfer to the ICU or a code call. They also included a control group. Control patients did not experience a transfer or code for at least 48 hours following the end of the data segment used for the control group. For transfer patients, the OW included 12 hours (Parshuram) or 24 hours (Duncan) of data up to one hour preceding the transfer. The warning time for both was therefore 1 hour. For control patients a stretch of data equal in duration to the transfer patient observation window was selected. The one-hour buffer is shown for three transfer patient scenarios in Figure 4.8.

As in Akre, the maximum score during the observation window was compared to a threshold. An "event" was predicted if the maximum score was above the threshold. However, this was done for both transfer and control groups, so specificity and sensitivity could be calculated. Furthermore, the one-hour gap adds a buffer before the predicted event during which clinicians have a chance to take action. For example, the patient who has a threshold crossing during the last 12 or 24 hours might have an event exactly one hour from the end of the observation period. Unfortunately, this still leaves uncertainty in the form of the observation window's length. Using the same example described above for Akre, a patient arrives on the floor and one hour later has

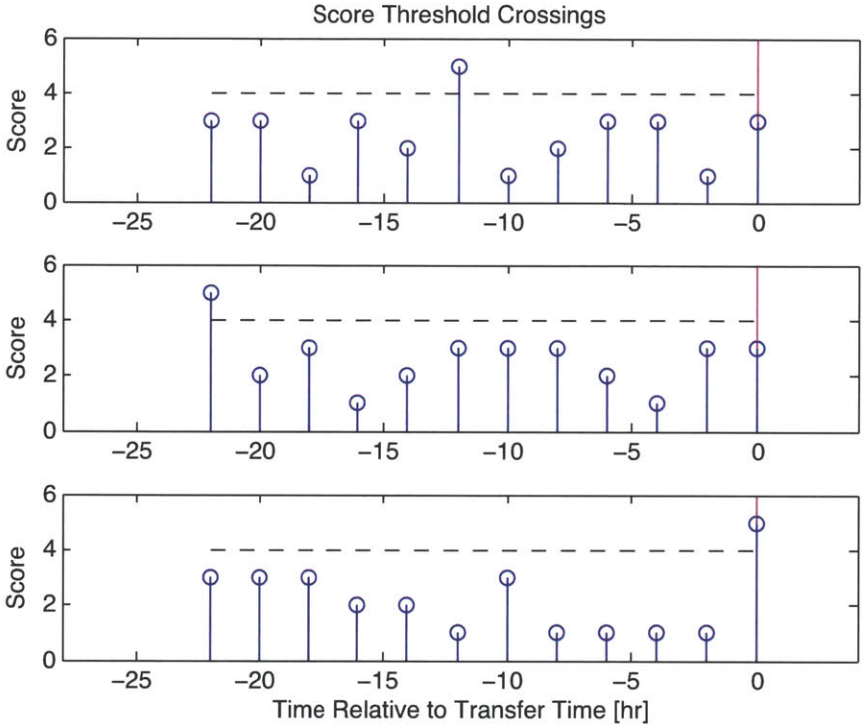


Figure 4.8: Three different threshold crossing locations all of which yield a sensitivity of 1.0 under Duncan's definition [16]. Time zero is the transfer time. Unlike in Figure 4.7, there is a warning interval of one hour between the red and magenta lines. The score threshold is set at 4.0. The observation window extends back in time from the red line.

a threshold crossing. Then the patient may have an event one hour from that point. Caregivers have a minimum one-hour warning time unlike in Akre where there is a zero-hour minimum. However, the event might not happen for as many as 13 hours (Parshuram) or 25 hours (Duncan) from that point because the observation window essentially becomes the uncertainty window. Ideally, one would like to be more precise in when the event will actually occur.

■ 4.4.3 The ROC with UW and WT

The ROC is a plot of the sensitivities and specificities for a variety of thresholds, where varying the threshold trades increased sensitivity for decreased specificity. The sensitivity is on the ordinate, and 1−specificity is on the abscissa. The area under the curve (AUC) traced by varying the thresholds is a one-number summary of the algorithm's performance. A ROC area close to 1.0 corresponds to an algorithm with good sensitivity and specificity for a range of thresholds, while an ROC area of 0.5 represents performance on par with random guessing. In particular, the area enclosed by a box whose upper left hand corner is a point with a given sensitivity and specificity is equal to the point's sensitivity multiplied by its specificity. Therefore, a point whose corresponding area is 1.0 translates to perfect sensitivity and specificity.

Just as a threshold is varied to generate a ROC, the warning time and uncertainty window should also be swept independently to generate a family of ROCs. Sweeping the WT and UW explores the full range of ROC areas as a function of WT and UW. If the area under each ROC for a given UW and WT represents that ROC, then the family of ROCs can be visualized as a two dimensional heat map with UW on the ordinate, WT on the abscissa, and pixel color proportional to the underlying ROC's area. Figure 4.9 shows a notional heat map. In this fictitious example, the performance drops as the WT is increased and UW is decreased because both changes correspond to increasing

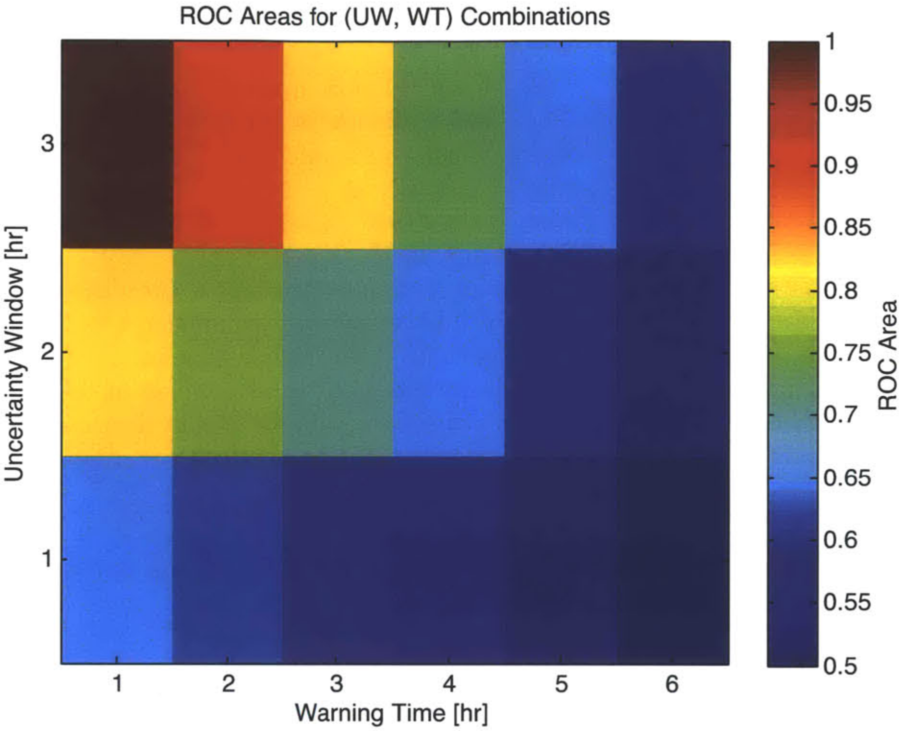


Figure 4.9: Generic heat map where each cell's color corresponds to the ROC's area under the curve for the given UW and WT.

the difficulty of the prediction problem.

By choosing a UW and WT point on the heat map to select a particular ROC and by choosing a threshold on that ROC, a precise statement can be made: WT hours from the prediction of a transfer, there begins a period of UW hours during which the patient will be transferred with sensitivity P_d, and specificity $1-P_{fa}$, where P_d is the probability of detection and P_{fa} is the probability of a false alarm. Thus, if a transfer event is predicted at time t, P_d is the probability that a transfer patient will actually be transferred during [t+WT, t+WT+UW]. P_{fa} is the probability that the algorithm declares a transfer to occur during [t+WT, t+WT+UW] and it does not. False alarms

Table 4.3: Selected ROC Notation

Term	Definition
OW	Observation Window in hours, > 0
UW	Uncertainty Window in hours, ≥ 0
WT	Warning Time in hours, ≥ 0
DW	Data Window, ≥ 0
t_s	Record start time, $\geq -\text{OW} - \text{WT}$
P_d	Sensitivity, [0, 1]
P_{fa}	Probability of a false alarm (1 $-$ specificity), [0, 1]
t_0	$t = 0$, call time for transfer patients
$S(t)$	Early warning score value at time t
$\lvert S(t)\rvert : t \in$ time interval	Cardinality (size) of the set of scores in the time interval
$\lvert S(t) \geq \eta\rvert : t \in$ time interval	Number of scores $\geq$ threshold η in the time interval

may occur for control patients because they are never transferred, but they can also occur for transfer patients if the transfer does not occur during the specified interval. In addition to a high sensitivity and specificity, a good algorithm will have a large WT and a small UW. A large WT will give clinicians sufficient lead time to intervene; a small UW precisely localizes in time the need to transfer.

4.4.4 ROC Calculation

We will present our approach to the ROC calculations in a step by step manner. Then we will follow this section by applying the definitions to our data. Table 4.3 summarizes some of the principal terms that have been and will be encountered.

The ROC calculation as applied to a single transfer patient is the following: given a WT, UW, and threshold, η, at each time, t, compare the early warning score, $S(t)$, to the threshold, η. If the score is greater than or equal to the threshold, check if the patient's call time for transfer occurred during $[t + \text{WT}, t + \text{WT} + \text{UW}]$. If the call time occurs in that window, the threshold crossing was a correct detection and is counted as a true positive. If not, the threshold crossing was a false positive. We denote the number of times a threshold crossing occurs during an interval as $|S(t) \geq \eta| : t \in$ time interval.

Following Akre [18], we define an indicator or binary sensitivity P_d, which we have explicitly written in Equation 4.10 using the UW, WT framework. For the i^{th} transfer patient, if there is *at least one* threshold crossing in [−WT−UW, −WT] hours, then $P_d(i) = 1$, and if there are no threshold crossings then $P_d(i) = 0$:

$$P_{d,transfer}(i) = \begin{cases} 1, & \text{if } |S(t) \geq \eta| \geq 1 : -\text{WT} - \text{UW} \leq t \leq -\text{WT} \\ 0, & \text{otherwise} \end{cases} \tag{4.10}$$

For transfer patients, opportunities for correct detection occur only during the window [−WT−UW, −WT]. For any time $t <$ (−WT−UW) there is no call time during the interval [t+WT, t+WT+UW]. Therefore any threshold crossings for $t <$ (−WT−UW) will be declared false alarms. P_{fa} for a single transfer patient can then be calculated as

$$P_{fa,transfer} = \frac{|S(t) \geq \eta|}{|S(t)|} : t_s \leq t \leq -\text{WT} - \text{UW}. \tag{4.11}$$

With the definition of sensitivity in Equation 4.10, it is possible to determine retrospectively where threshold crossings should occur and where they should not occur (Figure 4.10a). As mentioned above, threshold crossings should occur only during the UW region [−WT−UW, −WT] in Figure 4.10a because these crossings result in true positives. UW and WT are flipped in Figure 4.10a to graphically show where threshold crossings result in true positives. Sample scenarios corresponding to the threshold crossings are illustrated in Figures 4.10b, 4.10c, and 4.10d. Threshold crossings should not occur earlier than −(UW+WT), else they are false alarms. Threshold crossings cannot occur at −WT< $t < t_0$ because prediction with a buffer of WT requires truncating WT hours from the data prior to call time. Under a hypothesis of WT hours, data from [−WT, t_0] would not be collected, so the algorithm will not be tested on it.

For control patients, every threshold crossing is a false alarm because the patient is never transferred from the floor. There are never correct detections of an impending call time because call times do not occur by definition of being a control patient. Therefore, for a single control patient, only P_{fa} has meaning, and it is computed as

$$P_{fa,control} = \frac{|S(t) \geq \eta|}{|S(t)|} : t_s \leq t \leq t_0 \tag{4.12}$$

To compute the algorithm's overall performance, we give the following definitions:

$$M_{transfer} = Number\ of\ transfer\ patients \tag{4.13}$$

$$P_d = \frac{\sum_{i=1}^{M_{transfer}} P_{d,transfer}(i)}{M_{transfer}} \tag{4.14}$$

$$N'_{transfer} = |S(t) \geq \eta| : t_s \leq t \leq -\mathrm{WT} - \mathrm{UW} \tag{4.15}$$

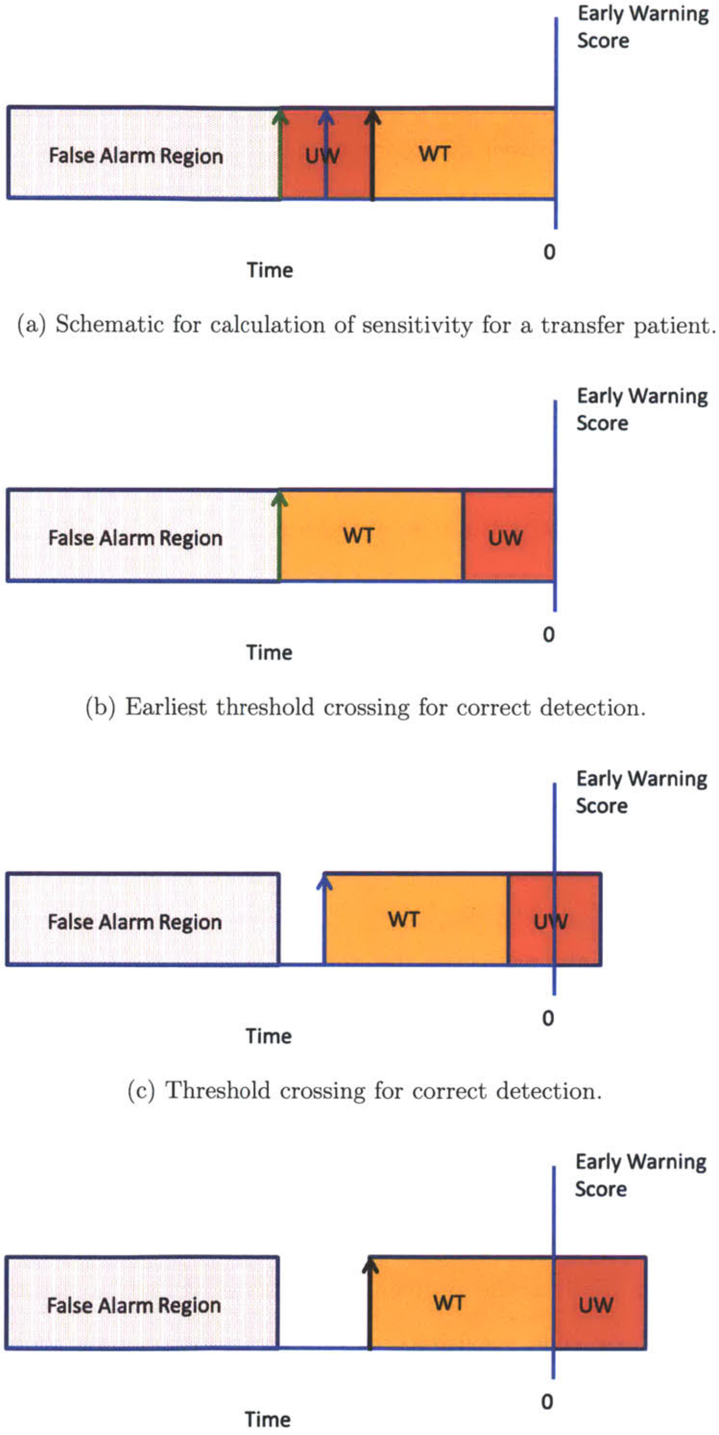
$$N_{transfer} = |S(t)| : t_s \leq t \leq -\mathrm{WT} - \mathrm{UW} \tag{4.16}$$

$$N'_{control} = |S(t) \geq \eta| : t_s \leq t \leq t_0 \tag{4.17}$$

$$N_{control} = |S(t)| : t_s \leq t \leq t_0 \tag{4.18}$$

$$P_{fa} = \frac{\sum_{All\ transfer} N'_{transfer} + \sum_{All\ control} N'_{control}}{\sum_{All\ transfer} N_{transfer} + \sum_{All\ control} N_{control}} \tag{4.19}$$

P_d represents the proportion of transfer patients that are correctly detected as transfer patients. $N'_{transfer}$ is the number of times that the score matched or exceeded the threshold too early for the given UW, WT pair in the transfer population. $N_{transfer}$ is the total number of scores from transfer patients that could have resulted in a false alarm if all the scores had matched or exceeded η. $N'_{control}$ and $N_{control}$ are defined similarly. P_{fa} represents the proportion of scores that yielded false alarms relative to the total number of scores among all transfers and controls that could have yielded false alarms if they had been above threshold. In subsequent sections and plots, we will refer



(a) Schematic for calculation of sensitivity for a transfer patient.

(b) Earliest threshold crossing for correct detection.

(c) Threshold crossing for correct detection.

(d) Latest threshold crossing for correct detection.

Figure 4.10: Examples of correct detection of need to transfer.

to this method of computing the ROC as ROCv2.

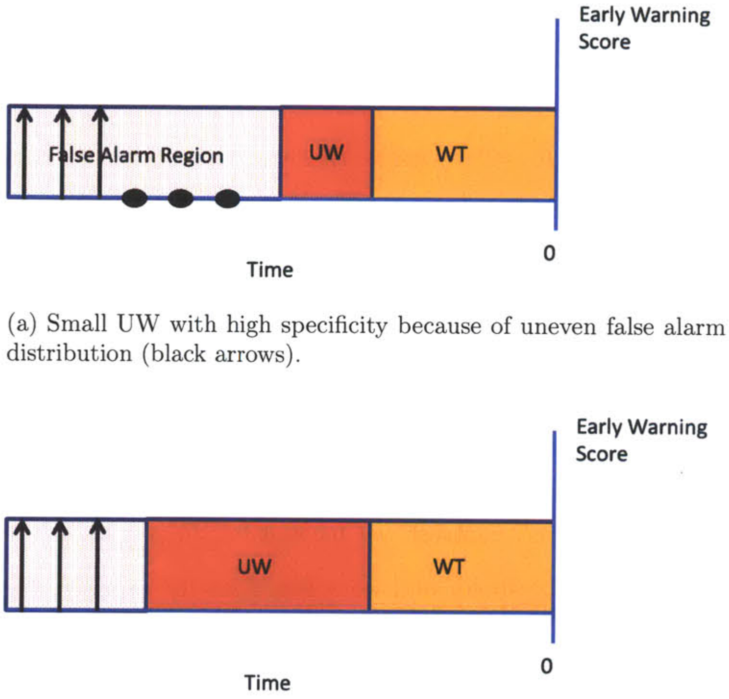
4.4.5 ROC Discussion

ROCv2 requires a UW, WT sweep to observe performance trade-offs. By using ROCv2, an algorithm's performance should trend in an intuitive manner if false alarms are uniformly distributed through the false alarm region. Defining sensitivity as perfect given at least one threshold crossing means that increasing the UW will at least not decrease, and probably will increase, sensitivity. There are more opportunities to catch a single threshold crossing during [−WT−UW, −WT].

We also expect that early warning scores should increase towards t_0. Consequently, a decrease in WT, which shifts the true positive region [−WT−UW, −WT] closer to t_0, is likely to improve the probability that the true positive region includes a threshold crossing score, and therefore improve the sensitivity of the algorithm.

However, as UW is increased the ROC area under the curve (AUC) may or may not increase because the specificity for transfer patients could increase or decrease. If false alarms are independent and uniformly distributed over the false alarm region [−OW−WT, −WT−UW], then specificity should stay the same on average. However, if false alarms are clustered towards t_s, the beginning of the false alarm region, then increasing the UW actually increases the density of false alarms, so specificity will decrease. For example, Figure 4.11a has higher estimated specificity than Figure 4.11b because there are three false alarms out of six evaluation instances as opposed to three false alarms out of three evaluation instances. Therefore, while there is no guarantee that increasing the UW and decreasing the WT will increase performance using ROCv2, a reasonable expectation is that the ROC area would increase.

A heat map also resolves the problem that a binary $P_d(i)$ prohibits comparing two algorithms based on their sensitivity. As shown in Figure 4.12, algorithm A may



(a) Small UW with high specificity because of uneven false alarm distribution (black arrows).

(b) Large UW with small specificity because of uneven false alarm distribution (black arrows).

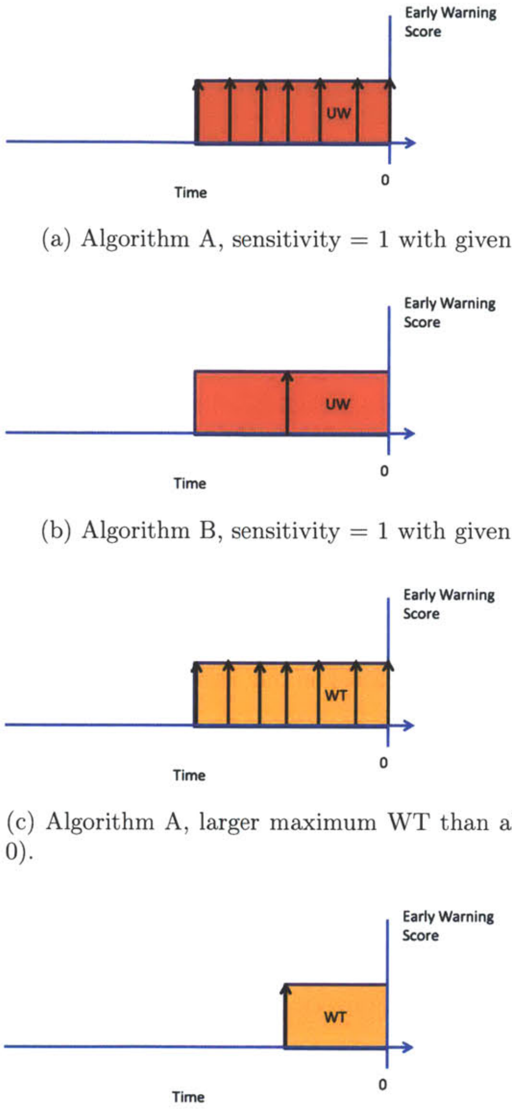
Figure 4.11: Relationship of small and large UW with specificity for ROCv2.

detect a threshold crossing ("trigger") and stay triggered while algorithm B may only trigger once during [−WT−UW, −WT]. However, both algorithms will produce perfect sensitivity because at least one threshold crossing occurred. On the other hand, a sweep through UW and WT will reveal that algorithm A is superior because it has a sensitivity of 1.0 with twice the WT and the same UW (UW = 0) as algorithm B.

It is tempting to increase sensitivity by increasing the frequency of evaluation $1/\Delta T$. More evaluation instances during the true positive region should increase the chance that at least one evaluation instance will be above threshold. Furthermore, increasing the evaluation frequency may not increase P_{fa} because the absolute number of false alarms would increase but the number of opportunities for false alarms would increase. The ratio may be unchanged. However, the cost of increasing the evaluation frequency is in the false alarm rate, $\dot{P}_{fa} = \frac{P_{fa}}{\Delta T}$, which is the number of false alarms per unit time.

While our definition of sensitivity follows that of Akre, Duncan, and Parshuram, our definition of specificity differs slightly. They compute specificity as the ratio of the number of control patients that did not have any scores above threshold to the total number of control patients. By contrast, we account for the number of times a control patient's score was above threshold, and we account for the number of a times when a transfer patient's score was above threshold that would have resulted in false alarm for the given UW, WT. We believe our approach is similar to how an EWS system would actually be implemented. If a control patient always is above threshold, that EWS rubric should have worse specificity than an EWS rubric that occasionally rates a control patient above threshold.

The framework we developed here may not really fit the probabilistic framework of an honest-to-goodness sensitivity/specificity calculation. By choosing our approach to specificity, our computations involve multiple contributions from each patient and unequal contributions from all patients because patients have different record sizes.



(a) Algorithm A, sensitivity = 1 with given UW (WT = 0).

(b) Algorithm B, sensitivity = 1 with given UW (WT = 0).

(c) Algorithm A, larger maximum WT than algorithm B (UW = 0).

(d) Algorithm B, smaller maximum WT than algorithm A (UW = 0).

Figure 4.12: Algorithm sensitivity comparison under ROCv2.

Furthermore, when the data window grows beyond just using the latest score and uses multiple past scores in determining if a transfer event will occur, the moment to moment predictions will not be independent predictions. Each prediction will depend on a segment of data within the data window that also is used in an adjacent prediction instance.

4.4.6 Understanding Published Results with UW and WT

With the ROCv2 developed in the preceding sections and the ideas of UW and WT, we can interpret the published values of Akre [18], Duncan [16], and Parshuram [51], and investigate how their results might have changed with a different OW. We did not try to re-score patients according to the Akre, Duncan, and Parshuram rubrics. Instead, the reproduction is done using the BCH CHEWS score, which is very similar to the criteria used by Akre and Duncan. Akre's score and CHEWS range from 0 to 11, and Duncan's score ranges from 0 to 9. The Parshuram score goes from 0 to 26. We used CHEWS as our surrogate score and evaluated performance on our data set using the OW, UW, and WT implied by the published papers. We then lengthened the OW and re-evaluated performance. We labeled the reproduced results using CHEWS as "mimic" in Table 4.4 and provided the original published performance values for reference. To reiterate, the only aspect of the authors' studies that we are explicitly mimicking is the respective UW and WT values.

Our data set for this experiment used between 207-211 control patients and 162-192 transfer patients. Transfer patients were from October 2010 through March 2012 and controls were from a similar time frame. The range in numbers exist because some patients do not have CHEWS scores more than six hours before the call time or the record end. This is a large data set because only the meta data, not the trend data was needed. To be consistent with published results, no cross-validation was performed nor was a MAP rule or SVM used as the decision rule. Instead, each ROC whose AUC is in Figures 4.13 and 4.14 was computed by sweeping the CHEWS threshold from 0 to 11 in integer steps, and P_d and P_{fa} were computed via ROCv2. The values in Table 4.4 are taken from the heat maps. When comparing specificities between published and mimic results, recall that the ROCv2 version of specificity is slightly different than the Duncan and Parshuram version.

As published, the observation window is equal to the uncertainty window for Akre, Duncan, and Parshuram. Equality of OW and UW allows excellent results as reproduced below in the second row for each of the respective authors' sections. However, when the OW is extended to 48 hours and the UW stays the same, there is a noticeable performance decline.

When OW is increased for controls, there is a larger interval during which false alarms may be recorded, but there is no reason to believe that the density of false alarms would change depending on the duration of the control record. Therefore, using only controls, the specificity should be unchanged (recall Equation 4.12).

For transfer patients, the time from $[t_s, -\mathrm{UW}{-}\mathrm{WT}]$ is a "control" period because any predictions during that period about future events are a false alarm (recall Equations 4.11 and 4.12). Chapter 3 showed that transfer CHEWS scores are obviously different than control scores even as much as 48 hours prior to call time. When UW and OW were equal, an algorithm only had to distinguish between transfer and control scores which was comparatively easy as suggested in Chapter 3. However, when OW was made larger than UW, an algorithm additionally must distinguish between transfer scores that are close to call time (within UW + WT hours) and transfer scores that are far from call time. The challenge associated with this additional requirement caused the performance decline in terms of specificity. Sensitivity should remain unchanged because the true positive region is unchanged. The decline in performance is exactly

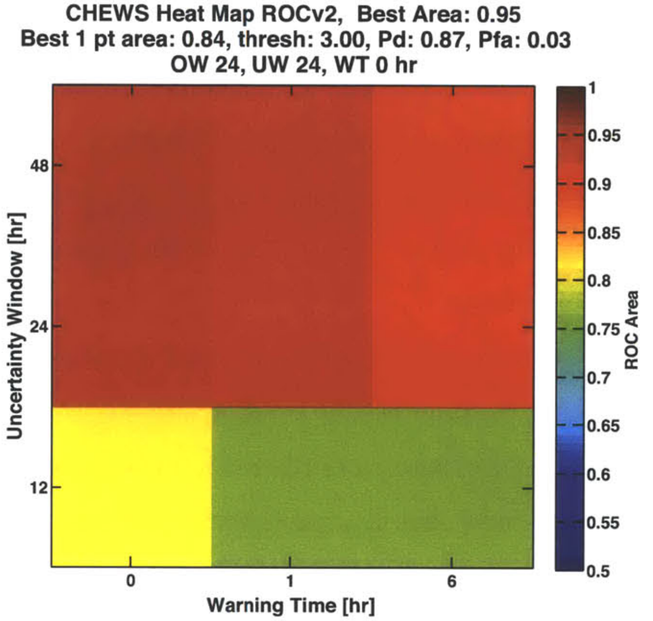


Figure 4.13: ROCv2 heat map for OW=24 hours. Approximately 200 transfer and 200 control patients' CHEWS scores were used to generate the ROC associated with each cell by varying the CHEWS threshold from 0 to 11. The top two rows are identical because a UW of 24 hours already encompasses the whole data record which is at most OW=24 hours long. Therefore, a UW of 48 hours does not bring any new chances for true positives or false positives.

the expected behavior when using ROCv2.

In conclusion, our analysis has led us to a prediction metric described by ROCv2, which uses the intuitive notions of warning time and uncertainty window along with the probabilistic metrics of sensitivity and specificity. We have taken the time to define the ROCv2 metric in order to provide a clinically meaningful understanding of results that have been published and to understand results for the rest of this thesis. The results presented in this subsection constitute a retrospective evaluation of CHEWS. In the following section, we will evaluate the performance of decision rules based on CHEWS and rCHEWS.

Table 4.4: ROC performance from literature as well as ROCv2 results ("mimic") using BCH CHEWS scores.

Source	OW	UW	WT	ROC Area	Threshold	Sens. at Threshold	Specf. at Threshold
	[hour]	[hour]	[hour]	[0, 1.0]		[0, 1.0]	[0, 1.0]
Akre 2010	24	24	0	Undefined	4*	0.86	Undefined
Akre 2010 ROC Mimic	24	24	0	0.95	3	0.87	0.97
Akre 2010 ROC Mimic	48	24	0	0.94	3	0.86	0.94
Duncan 2006	24	24	1	0.90	5	0.78	0.95
Duncan 2006 ROC Mimic	24	24	1	0.93	3	0.78	0.97
Duncan 2006 ROC Mimic	48	24	1	0.89	3	0.78	0.88
Parshuram 2011	12	12	1	0.87	7	0.64	0.91
Parshuram 2011 ROC Mimic	12	12	1	0.91	2	0.86	0.88
Parshuram 2011 ROC Mimic	12	12	1	0.91	3	0.75	0.96
Parshuram 2011 ROC Mimic	48	12	1	0.86	3	0.75	0.84

* Or single domain score ≥ 3

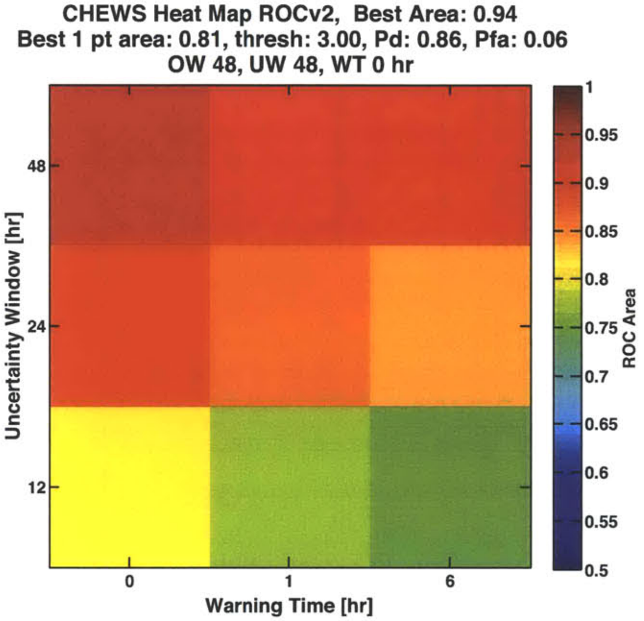


Figure 4.14: ROCv2 heat map for OW=48 hours. Approximately 200 transfer and 200 control patients' CHEWS scores were used to generate the ROC associated with each cell by varying the CHEWS threshold from 0 to 11.

4.5 CHEWS and rCHEWS Prediction

With the prediction metric of ROCv2 presented in Section 4.4, we can turn to the prediction problem in which we can now evaluate algorithms. In the prediction problem, we use a SVM detailed in Section 4.2.2 to decide if a patient is going to be transferred during a period of UW hours starting WT hours after a threshold crossing.

4.5.1 SVM Training for Prediction

SVM training is done with the ROCv2 metric in mind, which means that each patient actually contributes multiple training vectors unlike in classification where each patient contributes one training vector. The feature vector's time span may be less than the amount of data for a patient. For example, a patient may have 48 hours of scores but the feature vector only considers a sequence of N_{tap} equal to three scores which span 8 hours. N_{tap} is the number of scores in time needed to make a decision. The name is by analogy with the number of filter taps or coefficients in a finite impulse response filter. For a patient of either type with N_{score} values, the total number of contributed feature vectors, N_{total}, is $N_{score} - N_{tap} + 1$. N_{total} comes from sliding a window of length N_{tap} across the N_{score} length sequence and only counting points where the window completely overlaps the sequence.

For control patients, each of the points where the window fully overlaps the sequence is a control feature vector. If an algorithm is presented with that feature vector, it should declare a control patient.

For transfer patients, feature vectors whose most recent time point is less than −WT−UW hours are control features while feature vectors whose most recent time point is greater than −WT−UW are transfer feature vectors. The distinction comes from how decisions for transfer made outside of −WT−UW are actually false alarms because the transfer does not take place within the ensuing WT+UW hours. Only decisions for transfer made during the interval [−WT−UW, 0] are true positives.

The SVM parameters of C and tol were 10^6 and 10^{-5} respectively. The SVM was re-trained and re-sampled. These terms will be discussed in Chapter 5.

■ 4.5.2 Evaluation Method

The process of dividing a dataset into non-overlapping segments, training on some of these and testing on others, and then changing which segments are used for training and testing, is called cross-validation. K-fold cross-validation divides the data into K segments or 'folds', and runs K different trials, where K-1 folds are used as training and the last fold is used for testing. Our experiments used 3-fold cross validation.

In evaluating prediction, we will use a single value called AUC_pt_zero. AUC_pt_zero is the area under the ROC curve when an SVM threshold of zero is used, which comes from the SVM decision rule (Equation 4.6). The area associated with a point on the ROC is a lower bound on both the sensitivity and specificity for that point. Because the area is the product of sensitivity and specificity, and sensitivity and specificity are upper bounded by one, the area cannot be lower than sensitivity or specificity. The area can only equal sensitivity if specificity is one or vice versa.

We use up to the three most recent CHEWS and the maximum CHEWS of those three as inputs to the SVM. The chance predictor is provided as a reference. Rather than using the output from an SVM as the decision rule for a feature vector, we used a Bernoulli random variable with $p = 0.5$. Otherwise, P_d and P_{fa} were still computed using ROCv2.

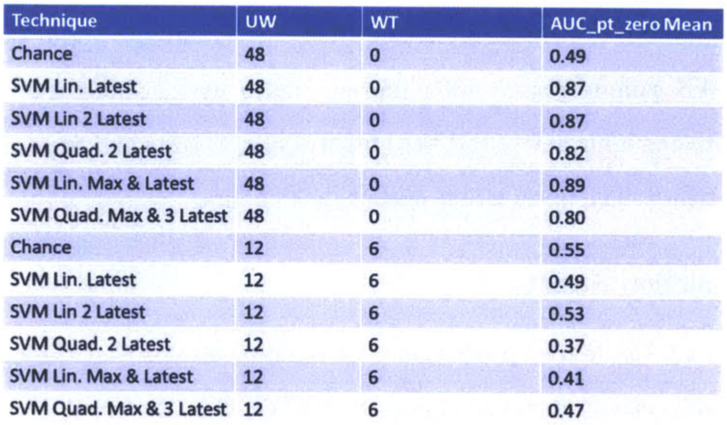
Because not all patients that have good trend data also have CHEWS scores, the patients used for the CHEWS SVM prediction performance are a subset of the 50 control and 50 transfer patients used for the rCHEWS performance. rCHEWS values are computed using the BCH rubric, which accounts for the heart rate, respiration rate, and SpO_2 values. Before a rCHEWS value is computed, the trend data is filtered with a ten-point (ten minute) median filter, and then passed through a four-hour averaging filter. rCHEWS sample points are uniformly taken at four-hour intervals stepping backwards from one hour before call time for transfer patients and from the record end for control patients. The observation window is 48 hours.

4.5.3 Prediction Results

Tables 4.5 and 4.6 summarize prediction performance. We see that the CHEWS based SVM does significantly better than chance with UW=48 hours and WT=0 hours, while rCHEWS does not. Additionally, we see that CHEWS performance is essentially the same regardless of the complexity of the SVM. Unfortunately, CHEWS and rCHEWS performance both decline to essentially chance when the UW is decreased and the WT is increased. CHEWS performance declines because for some patients, CHEWS scores are not elevated 6 hours prior to transfer while for other patients CHEWS becomes elevated more than 18 hours prior to transfer. Therefore, a CHEWS based SVM is not able to localize the transfer decision in time.

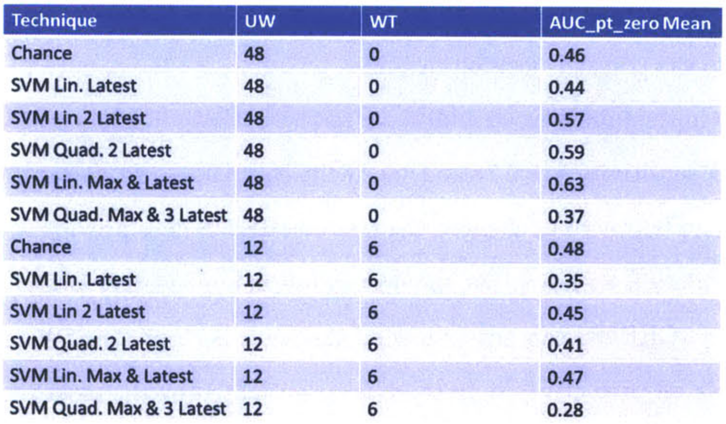
Not shown in the table was an additional test with UW=48 hours and WT=6 hours. With CHEWS, sensitivity was 0.78 and specificity was 0.90, so AUC was 0.70. However, rCHEWS was no better than chance. Therefore, CHEWS does provide a warning time but the uncertainty is so large that the overall use as applied here is limited.

In summary, CHEWS can act as a good distinguisher but the UW must be large. Also, essentially all predictive information is encapsulated by the most recent CHEWS for both a WT of 0 and 6 hours. rCHEWS fails as a predictor even if warning time is 0 hours.



Technique	UW	WT	AUC_pt_zero Mean
Chance	48	0	0.49
SVM Lin. Latest	48	0	0.87
SVM Lin 2 Latest	48	0	0.87
SVM Quad. 2 Latest	48	0	0.82
SVM Lin. Max & Latest	48	0	0.89
SVM Quad. Max & 3 Latest	48	0	0.80
Chance	12	6	0.55
SVM Lin. Latest	12	6	0.49
SVM Lin 2 Latest	12	6	0.53
SVM Quad. 2 Latest	12	6	0.37
SVM Lin. Max & Latest	12	6	0.41
SVM Quad. Max & 3 Latest	12	6	0.47

Table 4.5: Different linear and quadratic SVMs were used with CHEWS scores and evaluated with 3-fold cross-validation. The mean AUC associated with an SVM threshold of 0 is reported. Chance performance is provided as a reference.



Technique	UW	WT	AUC_pt_zero Mean
Chance	48	0	0.46
SVM Lin. Latest	48	0	0.44
SVM Lin 2 Latest	48	0	0.57
SVM Quad. 2 Latest	48	0	0.59
SVM Lin. Max & Latest	48	0	0.63
SVM Quad. Max & 3 Latest	48	0	0.37
Chance	12	6	0.48
SVM Lin. Latest	12	6	0.35
SVM Lin 2 Latest	12	6	0.45
SVM Quad. 2 Latest	12	6	0.41
SVM Lin. Max & Latest	12	6	0.47
SVM Quad. Max & 3 Latest	12	6	0.28

Table 4.6: Different linear and quadratic SVMs were used with rCHEWS scores and evaluated with 3-fold cross-validation. The mean AUC associated with an SVM threshold of 0 is reported. Chance performance is provided as a reference.

■ 4.6 Summary

This chapter has shown that classification using the CHEWS results in very high performance. If only the most recent CHEWS is used, classification is on par with or slightly better than the human classification performance in Chapter 3. For the prediction problem, we have introduced a clinically meaningful prediction metric, ROCv2, and applied it to several different prediction algorithms. If WT is 0, a simple threshold on the CHEWS yields performance on par with the more complex SVM algorithms. Even if WT is 6 hours, a one feature algorithm still appears best. The rCHEWS does not perform well as a predictor, so Chapter 5 will consider finer grained features that exploit the trend data directly.

Chapter 5

Modified rCHEWS

The previous chapter focused on methods to improve classification and prediction performance by a considering a history of the CHEWS score and also to test performance against an automated reduced CHEWS score (rCHEWS). This chapter focuses on changing the reduced CHEWS score at the level of the trend data to improve performance.

Various forms of temporal integration at the rCHEWS level only marginally increased performance despite significant increases in complexity. Therefore, it is necessary to move from the temporal axis of the feature space to the physiological feature axis. We need features other than rCHEWS that can be derived from the raw trend data if we are to improve classification and prediction performance.

In Sections 5.1 and 5.2 we will consider features at the level of the age-normalized vital signs. In Sections 5.3 and 5.4 we will consider incorporating meta-data directly as a feature. Finally in Section 5.5 we will return to the classification problem using a variety of features derived from the trend data.

■ 5.1 BCH Age-Normalized Trend Data

The conventional BCH rubric takes the unnormalized trend data and applies an age appropriate normalization in order to calculate each vital sign's percent deviation from normal. The percent deviation is then converted to a CHEWS subscore using coarse quantization boundaries. A deviation greater than 10% above normal is needed for a score of 1, a deviation greater than 25% above normal is needed for a score of 2, and a deviation greater than 50% above normal is needed for a score of 3. The CHEWS rubric does not allow distinctions finer than the 10, 25, and 50% levels. The CHEWS also enforces a strong asymetry because any deviation below normal, no matter how small or large, is immediately scored as 3; there is no gradual transition from 0 to 3.

In this section, we propose using the percentage deviation itself as the feature for the heart rate, respiration rate, and SpO_2 categories. First we will consider each vital sign on its own, and then we will include all three vital signs in the feature vector. As in Chapter 4, the vital signs will be preprocessed with a 10-point median filter, followed by a four-hour averaging filter. Samples will be taken at four-hour intervals, working backwards from one hour before call time for transfer patients (if WT=0 hours) and one hour before the record end for control patients. We used the SVM decision approach introduced in Section 4.5, with a warning time (WT) of 0 hours and an uncertainty window (UW) of 48 hours, because we wanted to test performance under the most favorable circumstances.

Figures 5.1 and 5.2 present prediction performance using only heart rate and respiration rate trend data. Chance is provided as reference for the SVM performance. The titles of these and similar figures in this chapter include the phrase "c:t". The 'c' refers to the number of control feature vectors, which can come from both control and transfer patients. The 't' refers to the number of transfer patients. These counts are the denominators of P_{fa} and P_d from the ROCv2 definitions in Chapter 4.

Chance performance was computed by randomly classifying each feature vector as either control or transfer with probability 0.5 and then evaluating prediction performance using ROCv2. Because control features are labeled with probability of 0.5 as control, we expect specificity to be 0.5. Transfer vectors are labeled as transfer with

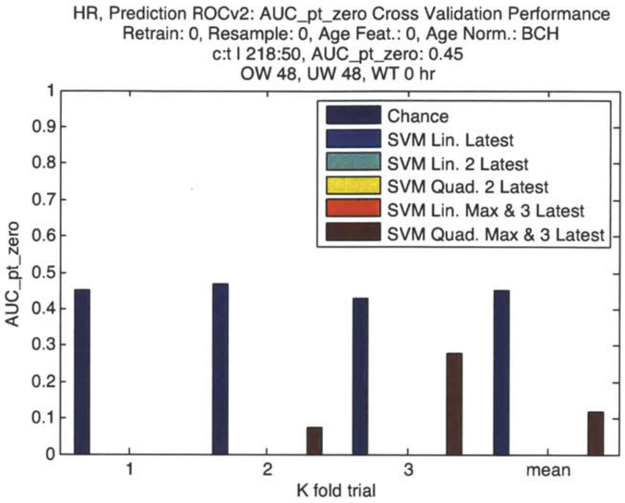


Figure 5.1: Area under the curve (AUC) for prediction problem using only heart rate and a number of SVM feature choices. SVM parameters: $C = 10^6$ and $tol = 10^{-5}$.

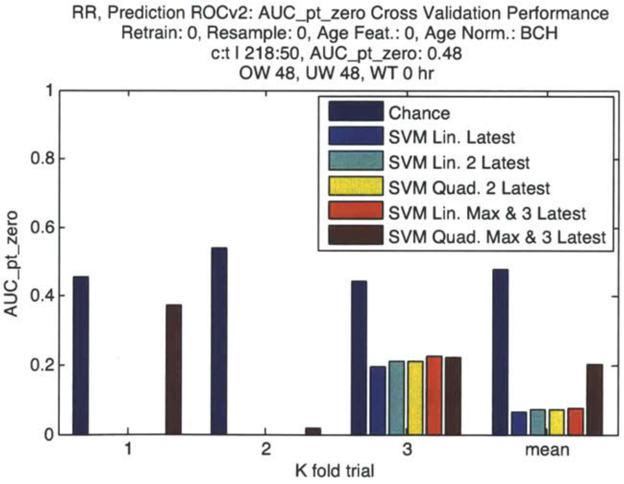


Figure 5.2: Area under the curve (AUC) for prediction problem using only respiration rate and a number of SVM feature choices. SVM parameters: $C = 10^6$ and $tol = 10^{-5}$.

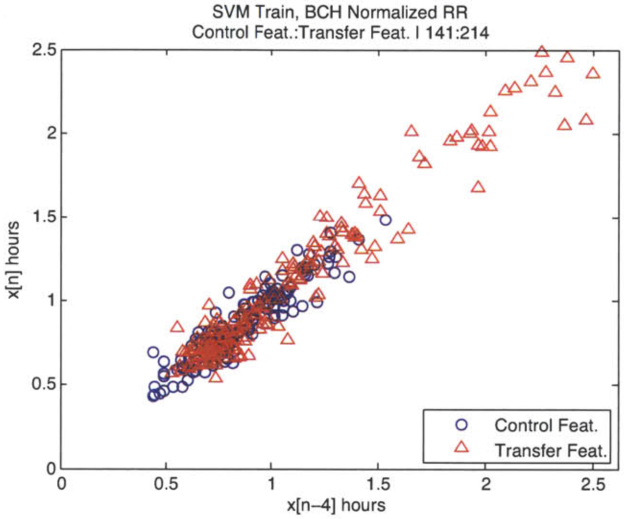


Figure 5.3: BCH age-normalized respiration rate for transfer (triangle, +1) and control (circle, −1). The plot of most recent value (x[n]) vs. the value four hours earlier (x[n-4]) shows extreme overlap. Because the data lies along a forty-five degree line, the values at x[n] are correlated with the values at x[n-4]. The red crosses represent the SVM classification decision for each feature vector. A red cross means the surrounding red triangle or blue circle was declared a transfer vector. Only red crosses are present because the SVM labeled all vectors as transfer vectors, so the sensitivity is 1 and the specificity is 0.

probability of 0.5, and ROCv2 only needs one transfer vector to be labeled as transfer in order for the sensitivity for that patient to be 1.0. Therefore, as shown in Chapter 4, the overall sensitivity of chance prediction for all the patients can approach 1.0. The chance area under the curve (AUC) is the product of the sensitivity and specificity, so chance performance should be approximately 0.5.

Under the SVM scenarios, feature vectors were labeled according to the sign of the computed margin, as discussed in Chapter 4.

Performance is terrible in both cases. None of the SVMs can do better than chance, regardless of whether just the most recent vital sign datum ("Latest") or several vital sign points are used ("2 Latest", "Max & 3 Latest"). When considering two dimensions only, Figure 5.3 shows that respiration rate at the most recent time point (ordinate) and four hours prior (abscissa) are almost completely intertwined, so every classifier is going to struggle with separating the points into two classes. Consequently, we observe that there is almost no predictive information encoded by the respiratory rate.

It is possible for two features together to be better predictors than either alone. For example, there might be equal numbers of control and transfer patients with symptoms A and equal numbers with symptoms B, but there might only be transfers with both symptoms A and B. Therefore, we moved to joint classification using both heart rate and respiration rate (Figure 5.4). Unfortunately, we still see very poor performance. The feature vector "Latest" refers to using the most recent heart rate and respiration rate jointly at each evaluation. "Max & 3 Latest" means we used the three most recent heart rate and respiration rate age-normalized values in addition to the maximum of each of these vital signs at each evaluation. The maximum was taken over the three evaluation instances of the vital sign during the previous 12 hours. While there is no apparent benefit to jointly using features, we did notice that the area associated with a point (AUC point) was exactly zero for many trials. We sought to explain this

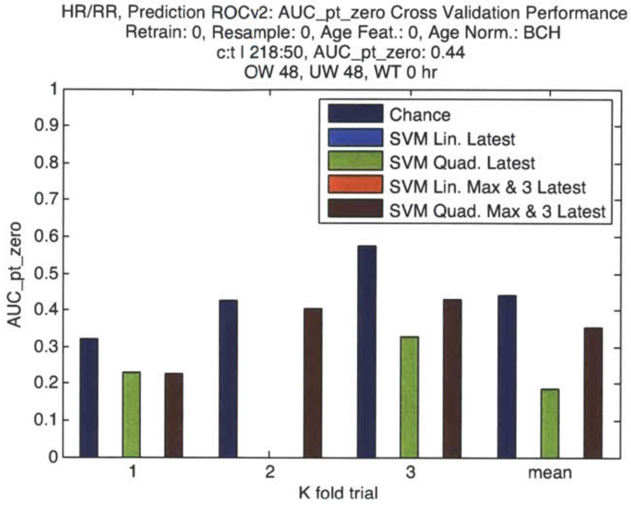


Figure 5.4: BCH age-normalized heart rate and respiration rate used jointly for prediction.

phenomenon in order to gain insight into the poor overall performance.

■ 5.2 Resampling and SVM Retraining

During training it is possible by chance that there are many more training vectors for one patient class than another because representative patients are chosen at random during each trial and some patients have more data than others. A class imbalance combined with poor separability between classes can force the SVM's optimal solution to simply declare all vectors as belonging to only one class. Therefore, sensitivity may be one, but specificity will be zero, so the area associated with the ROC point will be zero. This phenomenon explains why so many trials resulted in AUC values of zero in the preceding figures.

We mitigate the issue of unequal class representation with a modified sample with replacement scheme. For the predominant class, we take all the feature vectors as they

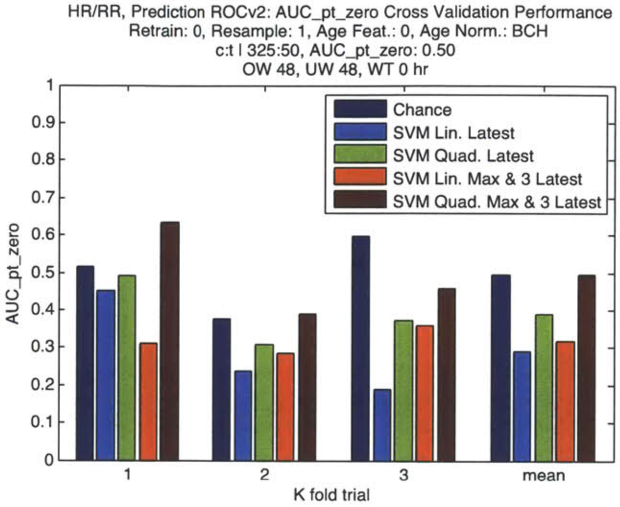


Figure 5.5: BCH age-normalized heart rate and respiration rate used jointly for classification. Resampling for class equalization was used, and overall performance is better than in Figure 5.4.

are. For the underrepresented class, we take all of the feature vectors available, and then we sample the underrepresented class with replacement until the number of feature vectors in both classes is equal. We can then proceed to cross-validation training and testing as usual. Figure 5.5 shows the result of sampling with replacement. We can see that overall AUC point area is increased for all techniques (except for chance) because the SVM is less likely to declare all patients as the predominant class.

In an additional attempt to improve performance, we have tried training the SVM twice. Training the SVM, and then modifying the SVM constraints before retraining, allows us to account for the fact that ROCv2 does not penalize a decision rule if the decision rule misses all but one opportunities to declare a transfer. In the first round of SVM training, the SVM tries to separate all transfer vectors from all control vectors. However, this task may be unnecessarily difficult because the SVM really only has to separate one transfer vector for each patient from all the control vectors. In trying to separate all the transfer vectors, the SVM's specificity may suffer.

Our solution is to train the SVM once using all the transfer and control data and then compute the margin for all the transfer vectors (recall from Section 4.2.2 that the margin is a distance of the point from the decision boundary). For each patient, we keep only the largest margin vector and then retrain using all control vectors but only the kept transfer vectors. This process should make the SVM classification problem easier by accentuating the extreme feature-vectors for the transfer patients. Because the retraining has many more control vectors than transfer vectors, we avoid the problem of the SVM declaring all patients as control by changing the SVM parameter C. Recall that a larger C more heavily penalizes misclassified points than a smaller C. We actually make C a vector, where C for the kept transfer vectors is multiplied by the total number of transfer vectors that were removed divided by the number that were kept. Consequently, the penalty for misclassifying the kept transfer vectors is substantially increased. While training twice works on simulated data, its impact on real data is both variable and limited.

Upon review of the different SVM classifiers' performances, the quadratic classifier that uses all the time points may at first appear to have potential because it never declares all features as belonging to one class or another. We considered whether or not we should focus our efforts on improving the quadratic SVM. Unfortunately, a simple explanation is available upon visual inspection of two-dimensional plots which likely extends to the higher dimensions used by the "Max & 3 Latest" version. The boundaries the classifier creates in higher dimensions are so complicated that they mimic a chance partitioning of the test data into two categories. Therefore performance is on par with chance, but the classifier actually is not partitioning the data in a meaningful way.

In summary, we have attempted SVM prediction using the vital-sign percentage deviations from normal for both heart rate and respiration rate and met with disappointing performance due to the large feature overlap. We have accounted for one weakness of the SVM by resampling data to ensure equal class representations to good effect, and we have discussed a method of training the SVM that is appropriate for the prediction metric ROCv2. Nonetheless, weak performance across patients motivates a closer look at the patients who are misclassified.

■ 5.3 The Misclassified Misfits

The rCHEWS has shown little overall discriminating ability, especially at the sub-CHEWS level. This could be because it is inherently not a good discriminator, but this could also be because it only discriminates certain types of patients well. Patient meta-data includes the age, gender, transfer reason, type of call, and hospital admission reason.

The underlying admission reason could potentially be a useful discriminator since it is unlikely that certain admission reasons will lead to life threatening deterioration and transfer. However, admission reasons are diverse and entered in clinical shorthand notation. Due to the difficulty of working with them, they will not be considered further. The type of call and transfer reason could be useful, except that Chapter 3 has shown that nearly all transfer patients are in respiratory distress. Gender is an interesting possibility but there are currently no grounds for believing that the rCHEWS would be favoring one gender over another.

Finally, there is age. We hypothesize that age is the most significant of the meta-data available; it can also be readily incorporated into testing. While the rCHEWS does adapt threshold levels to age specific norms, it is possible that the rCHEWS norms are only working well for particular age groups. For example, the physiology of newborns is not as developed as older patients, and the rCHEWS overall performance could be suffering by including children less than four years old.

■ 5.4 Using Age

Up to this point, the BCH quantization, which is based on BCH scores and age-dependent normal values, has been used. However, while the BCH quantization map has the advantage of being accepted clinical practice, it may be possible to use age as a feature explicitly in conjunction with the unnormalized vital-sign values. To use the unnormalized vital-sign values, we actually do have to normalize them because heart rate and respiration rate span widely different values and an SVM has heuristically been shown to perform better when the input feature vectors are all of the same comparable magnitude. However, the applied normalization will not be age dependent, it will simply be an age independent rescaling for each vital sign.

A common method of rescaling is to treat the data as if it were Gaussian, and then to convert it to standard normal by subtracting the mean and dividing by the standard deviation. For each vital sign, we estimated the mean and standard deviation for a single Gaussian by creating an empirical probability mass function using the median filtered trend data from all training patients (both control and transfer). We also estimated the mean and standard deviation of the age for all training patients.

Using the normalized feature vector of vital signs and age, we ran the SVM as usual and obtained the results in Figure 5.6. Unfortunately, even explicitly including age is not sufficient to reliably achieve performance on par with chance.

■ 5.5 Custom Features

Given the lackluster predictive performance of the raw bedside data, we take a step back and return to the classification problem. Given all the data over the past 48 hours, can we improve upon the rCHEWS performance in Chapter 4? If we can, perhaps those

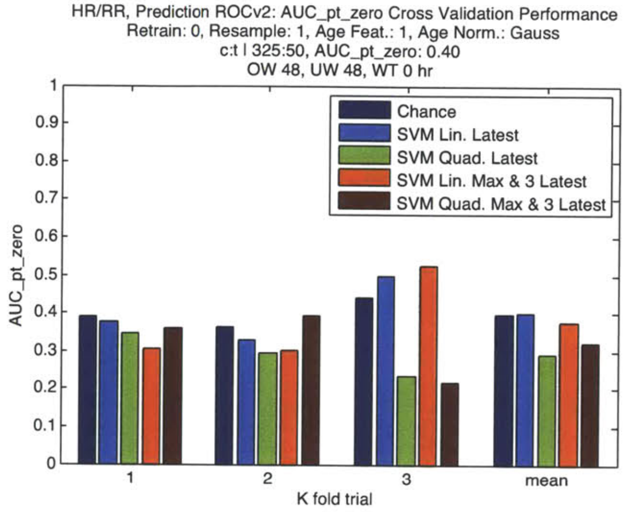


Figure 5.6: Heart rate and respiration as vital signs, Gaussian normalized values, age is used as a feature.

features may apply to the prediction problem. We computed a series of possible features for each vital sign and then derived several more features. Using age plus one feature, we tested every age-feature pair using a linear SVM. We did not try to jointly fit all features in order to keep the model simple and interpretable. We wished to identify the classification ability of each feature by itself, but we did include age as a second feature in each SVM because we did not age-normalize the raw values by the BCH scale factors. Instead we used the approximate Gaussian normalization.

We used the bedside data of heart rate, respiration rate, SpO_2, systolic pressure, diastolic pressure, and mean blood pressure to compute the average value, interquartile range (IQR), length transform, and least-squares line slope of the vital sign over the past 12 hours before transfer. We also computed the pulse pressure and values proportional to cardiac output and total peripheral resistance. The derived features such as pulse pressure were included because they encode dependencies that might not be obvious if vital signs were taken at face value by themselves.

We considered mean values and simple measures of variability such as the IQR as natural starting points because of their ease of interpretation. We used the slope of a trend line in order to characterize trends if they were present.

A more advanced measure of variability than IQR is the length transform, $L(y)$, in Equation 5.1 [60]. The length transform measures the length of the path as one moves along the path:

$$L(y) = \frac{1}{N} \sum \sqrt{1 + \left(\frac{y[n] - y[n-1]}{t[n] - t[n-1]} \right)^2} \tag{5.1}$$

N is the number of points in the record, y is the signal value and t is the time. We normalized the length transform of the signal by the number of points used in the computation to account for records that did not have 12 hours of data. Unlike IQR, the length transform captures information about how the signal evolves with time. Reordering points will not change the IQR but will change the length transform. The length transform may be especially suited for detecting transient events among an otherwise steady signal such as desaturations in the SpO_2 signal.

We trained and tested on the same data because we have insufficient patients per age group to effectively use age as a feature unless we use all available patients. This method will therefore overestimate performance. We used a linear SVM with parameters C=1e6 and tol=0.01. The data was not median filtered in order to accentuate variability, especially with respect to the SpO_2 signal. We also limited ourselves to patients greater than or equal to four years of age, and normalized the age as detailed in Section 5.4. Of the original 50 transfer and 50 control patients, the age limitation reduced the set to 21 transfers and 23 controls. The decision to restrict to older patients was motivated by signs that classification would be easier for older patients and because we wished to see performance under "good circumstances." If performance is bad even under the

Table 5.1: Custom Feature Classification Performance via SVM

Pcc SVM	Feature
1) 0.75	Average Respiration Rate
2) 0.73	Average Mean BP
3) 0.73	SpO_2 Length Transform
4) 0.68	Average Systolic BP
21) 0.59	Average Heart Rate
Pcc: probability of correct classification	

most favorable conditions, we would not expect good performance among the younger age group.

Table 5.1 shows the performance of the top four features and the performance of the average heart rate for reference. Two important observations can be made. First, the mean respiration rate is actually a much better classification feature than the mean heart rate. Despite the fact that clinicians note high heart rates, the respiration rate actually contains more discriminating information. This is reasonable, given that most patients are transferred for respiratory distress. Second we see that the length transform of the SpO_2 signal yields more information than its mean. This is again reasonable because the length transform is a measure of the variability of the signal. A mean over twelve hours will easily smooth out intermittent desaturations, but the length transform will register larger values because of the increased signal variability.

We also attempted to combine features using a voting scheme to capitalize on the individual feature performances. Each feature cast one vote for each patient and the majority of votes determined the declared patient type. Voting was not able to significantly change overall performance when compared to the individual performance of just the mean of respiration rate. As more features were incorporated, performance actually declined. The low classification abilities of most of the features resulted in essentially adding noise to the moderate classification abilities of the top features.

To summarize, we have investigated various methods of improving prediction using the raw bedside monitoring data. We have addressed potential problems associated with the SVM training but still see poor predictive performance. We returned to the classification problem and discovered that the overlooked respiration rate signal does contain classification value greater than that of heart rate. Unfortunately, many other features have no substantial classification value, so there is not an expectation that prediction performance can be improved with them.

Overall, the results associated with the prediction and the classification problems do not suggest that clinically acceptable performance can be achieved just from the readily available vital sign data above. Note that many results in this chapter are best-case results, and performance is likely to be even worse in real-life applications.

Chapter 6

Conclusion and Future Work

In approaching the problem of distinguishing which patients might be transferred from the general ward to the ICU, this thesis had two aims. First, the thesis set out to evaluate the capability of the BCH CHEWS score in identifying patients at risk of decompensation and for prediction of transfer. Second, it aimed to explore the utility of routine, automatically collected patient bedside data for the same tasks. Bedside data included intermittently sampled blood pressure determined by the oscillometric method, and heart rate, respiration rate, and SpO_2 levels, sampled at one-minute intervals for pediatric patients, 0-18 years of age and with diverse pathologies. Prediction is desirable in order to prevent transfer, through appropriate treatment, or to preemptively transfer in order for an impending decompensatory event to take place in a properly equipped environment.

In order to evaluate the predictive abilities of CHEWS and the bedside data, we first needed to introduce a clinically meaningful, probabilistically based prediction metric that encapsulated sensitivity, specificity and a measure of the prediction's temporal localization. We believe the UW-WT framework discussed in Chapter 4 is an essential element for early warning score evaluation that is currently lacking in the early warning literature. Without it, published results can claim success without rigorously showing if an EWS rubric is providing both sufficient warning time to take action and sufficient localization in time to be meaningful.

Our advice on improving the use of CHEWS is limited, given its good classification performance. However, our analysis suggests that a CHEWS greater than or equal to 3 is indicative of a patient being a transfer patient, rather than the currently used threshold of 4 or 5. Moving the threshold for transfer to 3 could advance the transfer decision by 4-6 hours. Furthermore, we see that CHEWS is more specific than sensitive. There are many more controls than transfers, so a high specificity will give the appearance of good performance because probability of correct classification will be high. However, the low sensitivity shows that the performance comes at the expense of missing transfer patients. We note that CHEWS underscoring occurs primarily in the respiration rate category, so one avenue for improving sensitivity would be to pay particular attention to respiratory health when scoring that section of the CHEWS rubric. While the EWS literature describes some indicators of impending transfer, a thorough BCH-specific, retrospective evaluation of each transfer when it occurs might reveal patterns of care that are more sensitive indicators of a future transfer than the currently used vital signs. These other indicators might include admission reasons, medications, and response to medications.

The utility of trend data from bedside monitoring for early warning is more questionable than the utility of CHEWS. We see that classification with rCHEWS is slightly better than chance with a WT of 0 hours (Pcc = 0.69), but it is no better than chance with a WT of 6 hours (Pcc = 0.47). However, there exists a clear positive correlation of rCHEWS performance with age. The oldest age bracket of 12-18 years has a Pcc greater than 0.80 in both cases. Currently, we cannot say that rCHEWS (or bedside trend data in general) are not beneficial. We see more promise in older patients who have more variation in vital signs than younger patients. Increasing the numbers of patients in each age group could provide evidence as to whether the rCHEWS has clinical utility for older patients or only has potential for older patients. Therefore, data availability rather than fundamental limitations of rCHEWS currently limits conclusions in this regard.

This thesis reveals the challenges in using monitoring data for predicting the need to transfer patients. In younger patients, vital sign trend values show nearly perfect overlap between control and transfer distributions. Completely overlapped distributions are a fundamental limitation that more data of the same kind cannot help overcome. For older patients, vital-sign segregation between the two groups is more pronounced, yet we see that the vital signs (heart rate and respiration rate) generally stay elevated for tens of hours before transfer occurs, if it occurs. While monitoring provides a graphic record of this phenomenon, charted four-hourly vital signs would provide a nearly identical picture. A monitor's strength may be in emphasizing to a clinician just how long a vital sign has been elevated rather than searching for a quick decompensation or an indicative trend.

Among the vital signs to which we had access, the data suggests that respiration rate, systolic blood pressure, and mean SpO_2 values have more utility than heart rate. Heart rate has high positive predictive value for transfer patients older than 4 years, but it is not specific. On the other hand, high respiration rates, high systolic blood pressure, and low SpO_2 values are more sensitive and specific than heart rate, but none of them are decisive. Therefore, clinicians may want to pay particular attention to these values during rounds, especially given that CHEWS underscoring occurs primarily in the respiratory rate category and most patients are transferred for respiratory distress.

The classification and prediction problems approached in this thesis are challenging for several additional reasons. First, the majority of transfers occur for respiratory distress, which is poorly defined symptomatically. The underlying deterioration could be from a multitude of pathologies. Second, there is no objectively defined gold standard for control and transfer patients. Classification is difficult because we do not know if some patients who were transferred by one medical team would not have been transferred by another. Prediction is difficult because we also do not know if the actual time of transfer would have varied if different staff had been on call. There may be context-dependent decisions such as inability to transfer because the ICU is at capacity, or decision to delay transfer because the patient is in a higher care, but still non-ICU bed. These limitations on algorithm evaluation will exist because of the subjective component of evaluation. If some patients were associated with a certain clinical team, some insight might be gained into whether particular clinicians had above average sensitivity and specificity. Unfortunately, simply collecting more data will not solve the subjectivity associated with assessing the severity of symptoms such as "work of breathing."

Additional cases could help with the problem that vital-sign values (or features extracted from them) are age dependent. Currently, we have insufficient data to properly characterize any age bracket, except for the patients under four years old. We do not even know for very young patients (less than one year) if these are term infants. We suggest both increasing the number of patient records for all age groups and performing our analysis for each age bracket separately.

We have mentioned some deficits with the data set. Important vital signs are not continuously monitored (blood pressure) or are missing entirely (temperature). We cannot observe a patient's physical or mental state. Additionally, we do not know what interventions have taken place. Consequently, SpO_2 values or changes in heart rate or breathing rate can be misleading. We would expect temperature to be a more useful indicator than blood pressure because blood pressure is known to be a late indicator of decompensation whereas temperature may herald a low-level infection before it becomes a dangerous and potentially life-threatening case of sepsis.

We also suggest focusing on a specific, well defined pathology with an unambiguous diagnosis, and preferably an unambiguous time of the event. For example, respiratory arrest is specific, and its time of onset, unlike the transfer time, is not debatable. However, doing so avoids the larger question of predicting transfer in the general pediatric inpatient population. On the other hand, if prediction could be done for this small subset, that would already be beneficial because respiratory and cardiac arrests are life threatening.

The ECG, respiratory waveform, and plethysmogram carry their own challenges but also the potential of very fine grained temporal features. We have done preliminary investigations into heart rate variability, but essentially all the waveform data have yet to be exploited. On the other hand, we see significant noise for stretches of time in these waveforms, and the sampling rate for the ECG is only sufficient to pick up large ECG features because a child's heart rate can be twice as fast as an adult's heart rate.

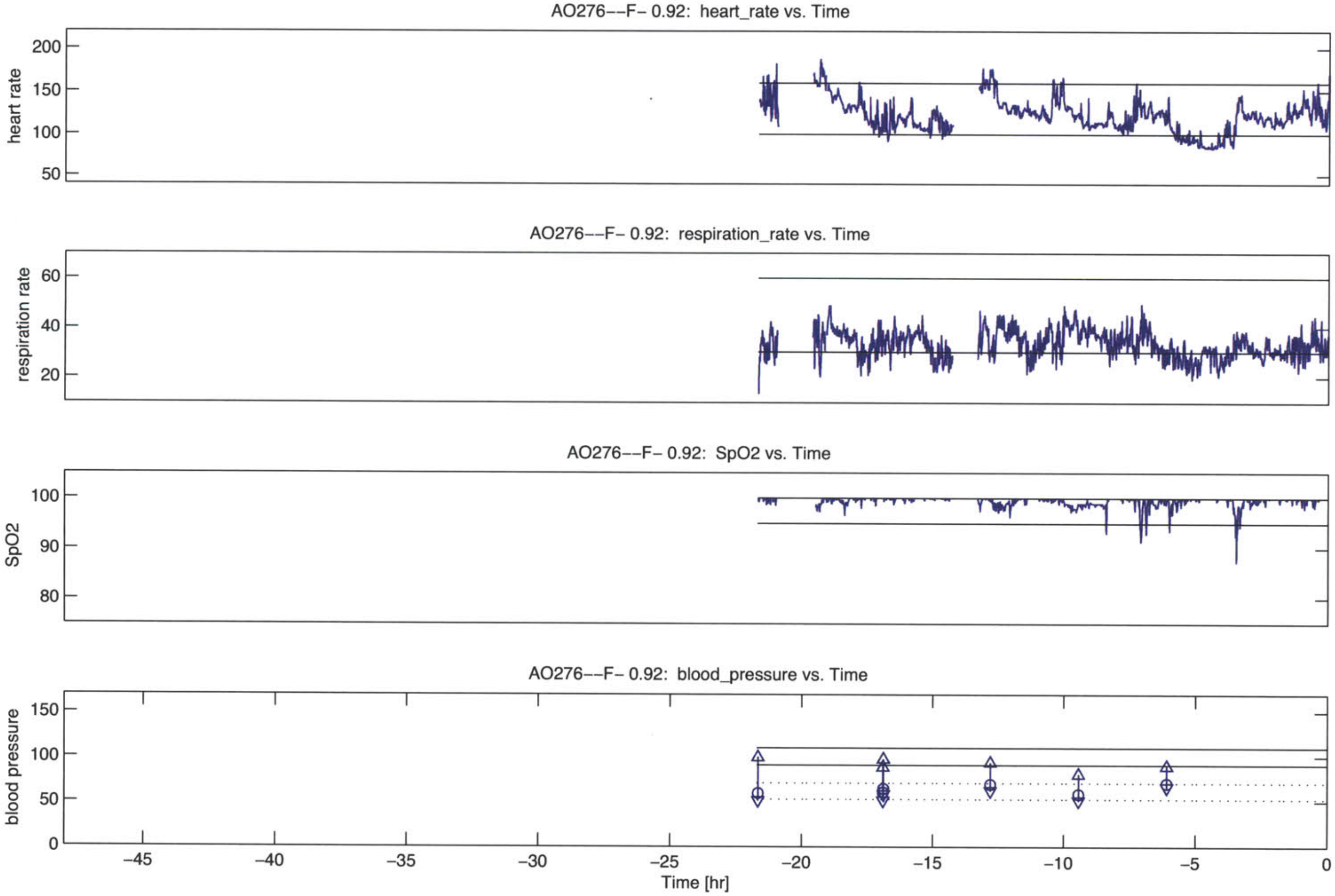
The utility of prediction is rooted strongly in what interventions, if any, can be taken to alter outcomes. It could be insightful to consider the interventions that take place immediately upon transfer to the ICU to determine the most common interventions and their outcomes. It would also be beneficial to retrospectively determine for transfer patients what interventions applied at what time could have prevented transfer. For example, if simply giving supplemental oxygen or another low-cost and low-risk treatment can prevent transfer, then even a non-specific predictor may still be beneficial.

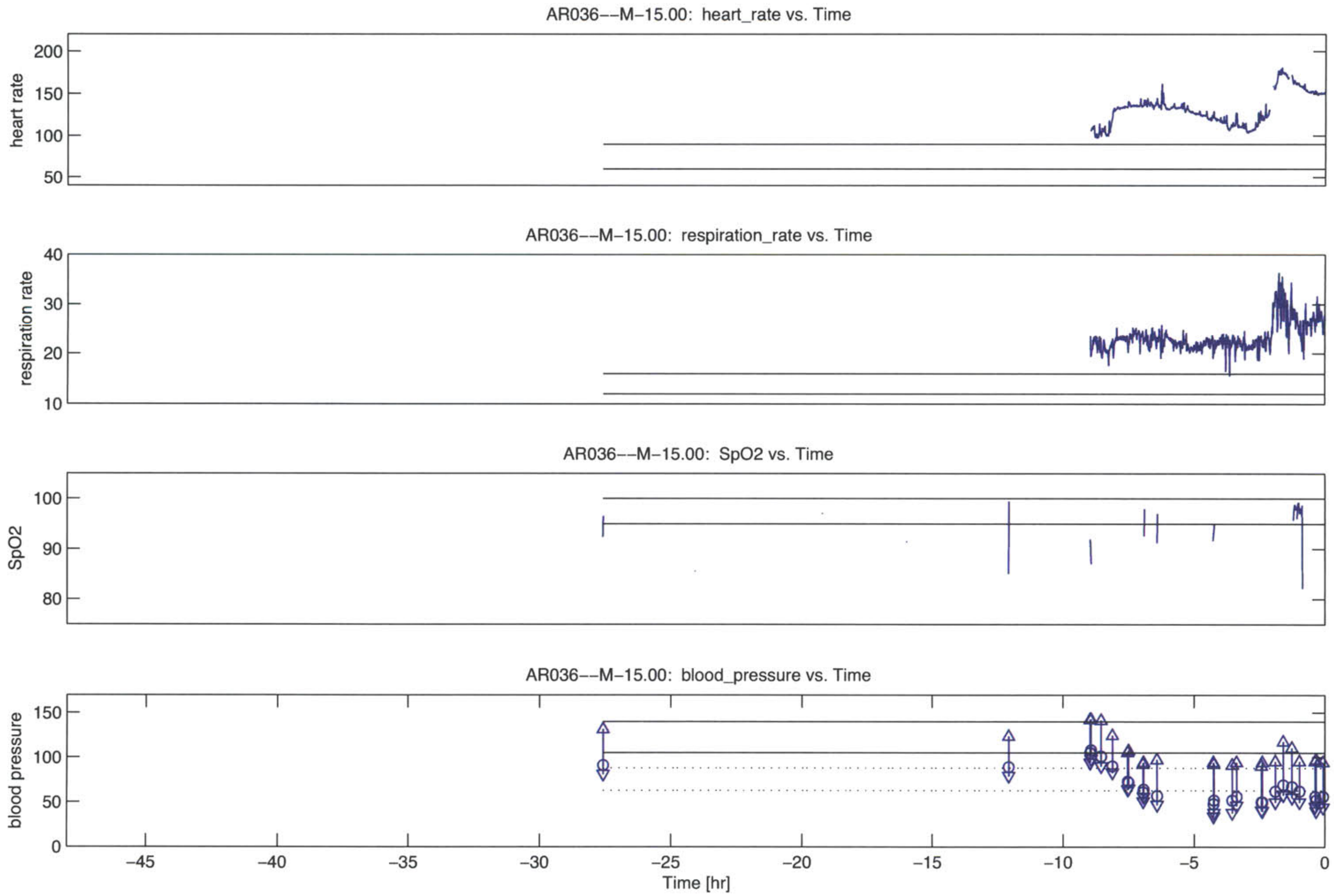
In summary, we have shown that CHEWS is a good distinguishing metric for transfer and control patients. We have not seen equal success with the bedside monitoring trend data, but part of the limitations discussed are insufficient data rather than a discovered and demonstrated limitation of the bedside data itself. The problem of early identification of patients at risk of acute physiological decompensation remains important and is certainly worthy of continuing attention.

Appendix A

Classification Self Test

The following pages are unlabeled patients showing unfiltered machine data. The charts go from [-48, 0] hrs. If the patient is a transfer, then the 0-hour mark is the call time. If the patient is a control, then a random 48 hr (or shorter) stretch of data is plotted. The BCH "normal" limits are the black horizontal lines. The age and gender are in the title. The reader's task is as follows: look through the charts, and for each patient determine whether at time t=0, the patient was transferred or not. The solutions are on the page following the last chart.


AO276--F- 0.92: heart_rate vs. Time
heart rate
200
150
100
50
AO276--F- 0.92: respiration_rate vs. Time
respiration rate
60
40
20
AO276--F- 0.92: SpO2 vs. Time
SpO2
100
90
80
AO276--F- 0.92: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

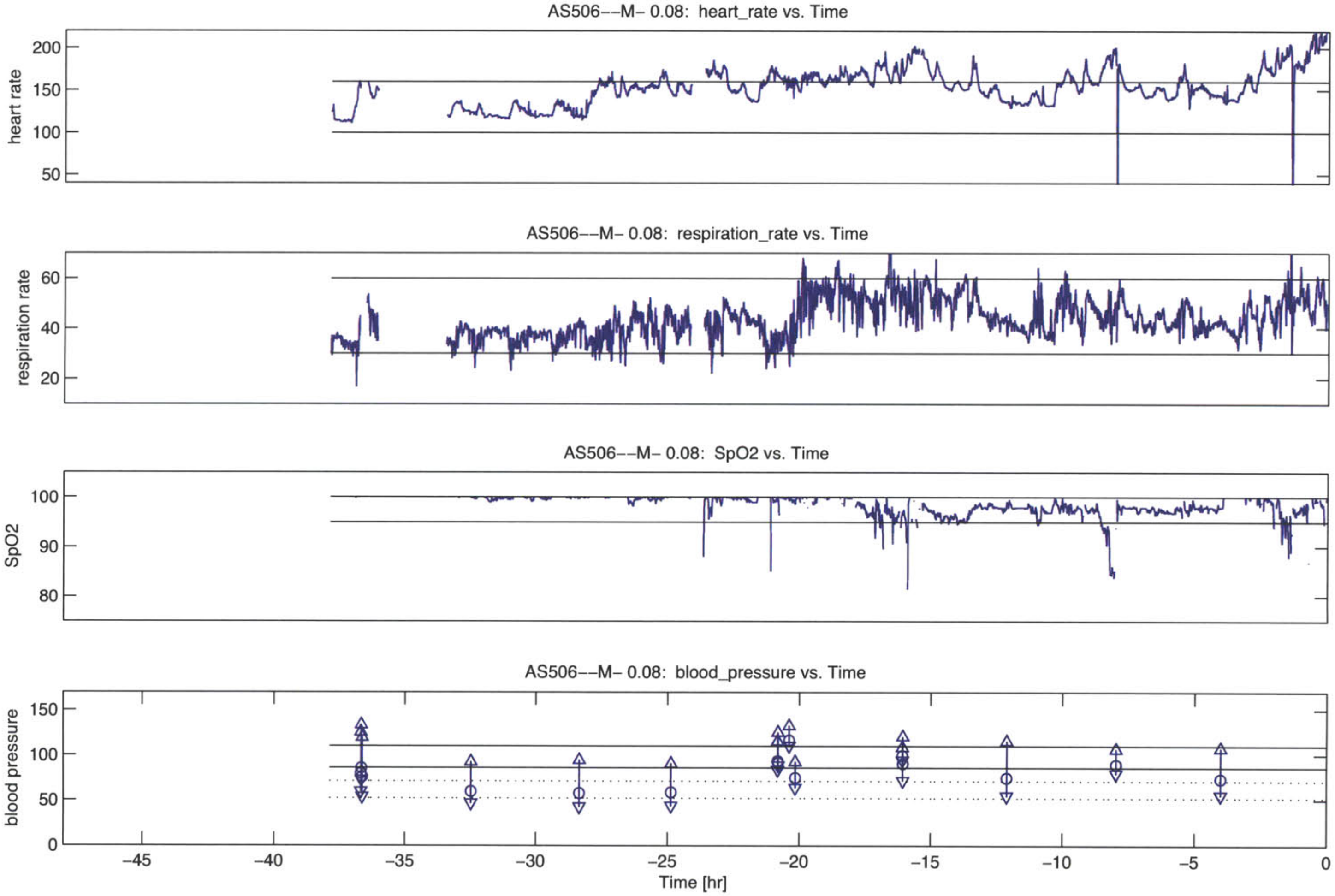


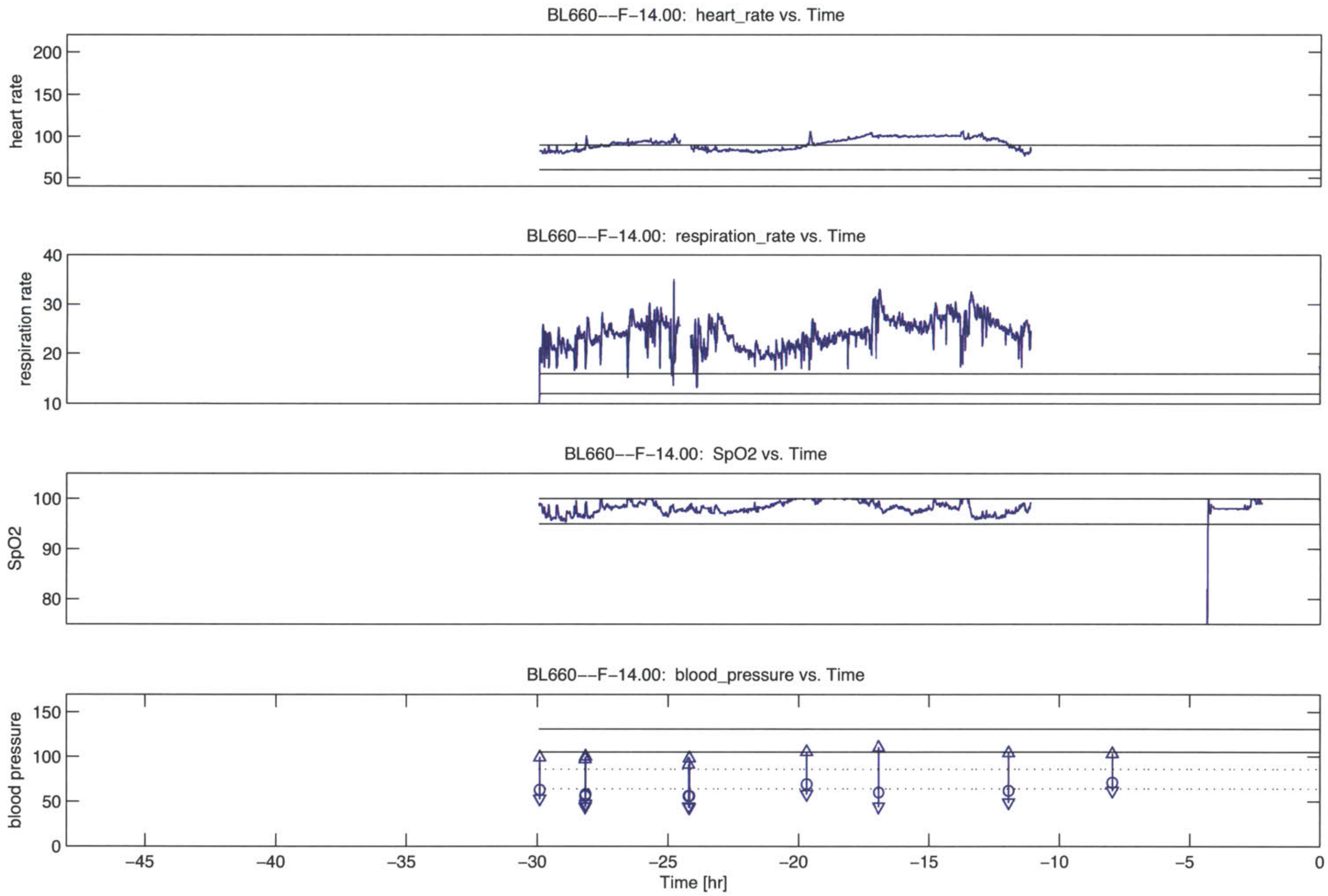
AR036--M-15.00: heart_rate vs. Time

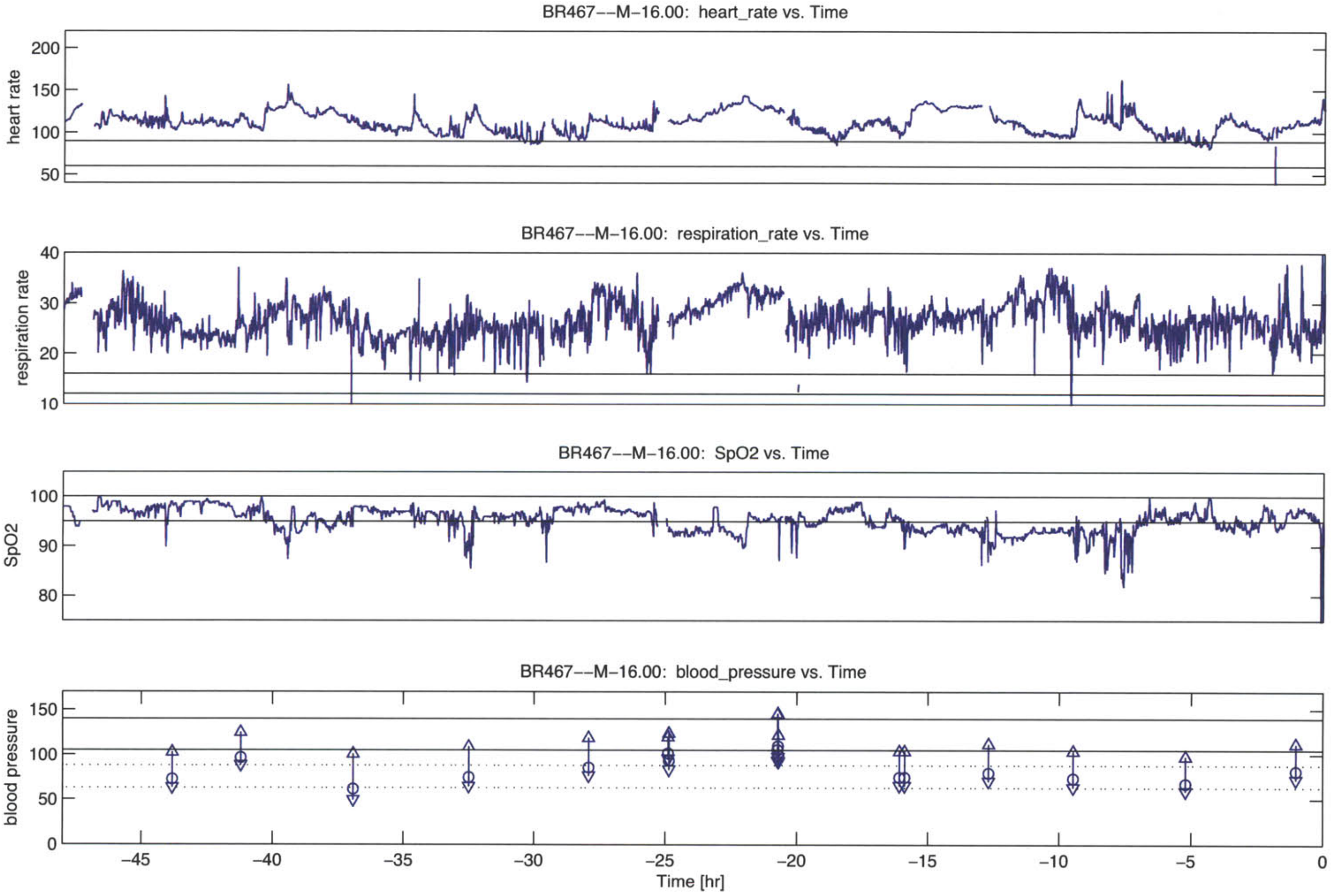
AR036--M-15.00: respiration_rate vs. Time

AR036--M-15.00: SpO2 vs. Time

AR036--M-15.00: blood_pressure vs. Time


AS506--M- 0.08: heart_rate vs. Time
heart rate
200
150
100
50
AS506--M- 0.08: respiration_rate vs. Time
respiration rate
60
40
20
AS506--M- 0.08: SpO2 vs. Time
SpO2
100
90
80
AS506--M- 0.08: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


BL660--F-14.00: heart_rate vs. Time
heart rate
200
150
100
50
BL660--F-14.00: respiration_rate vs. Time
respiration rate
40
30
20
10
BL660--F-14.00: SpO2 vs. Time
SpO2
100
90
80
BL660--F-14.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

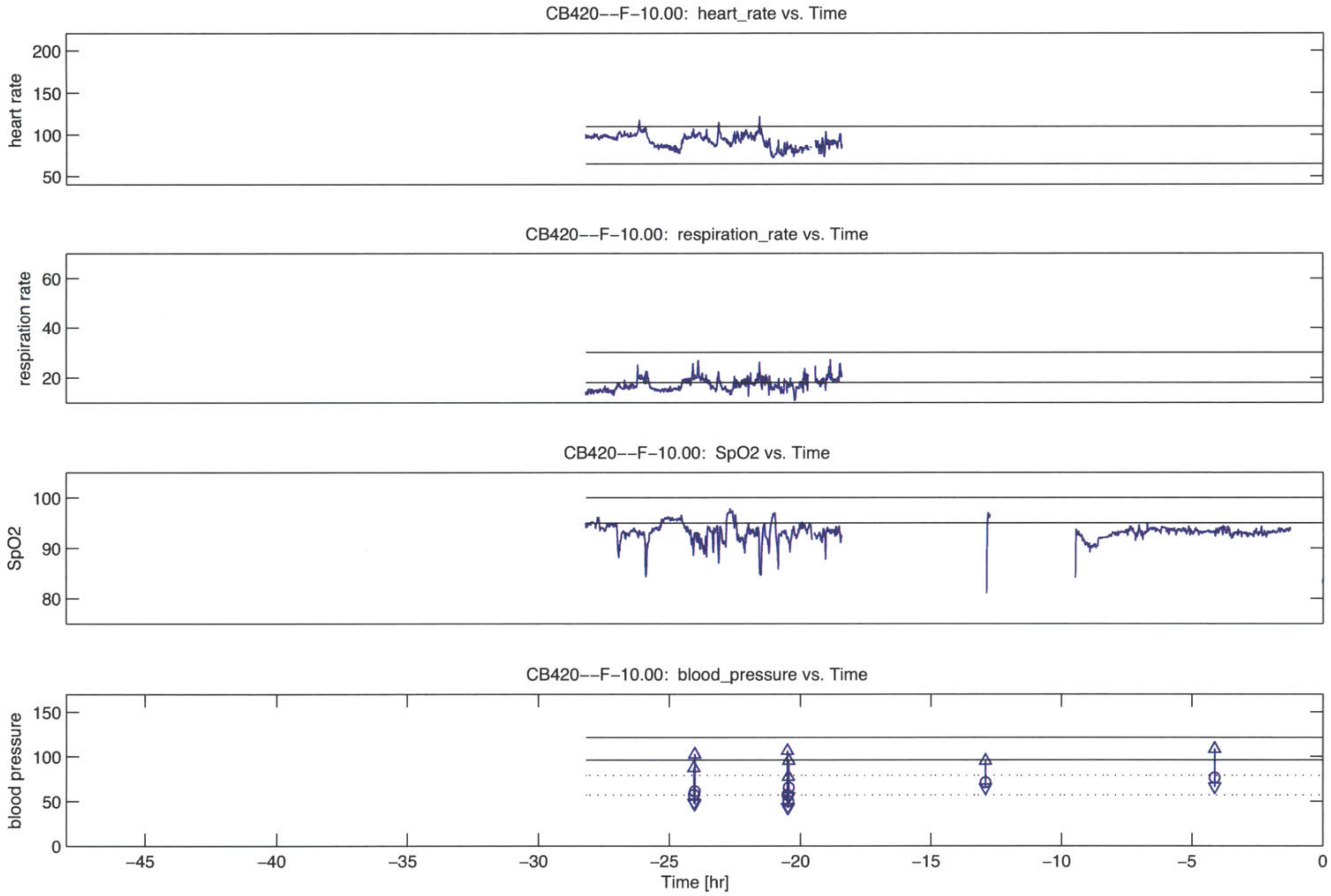


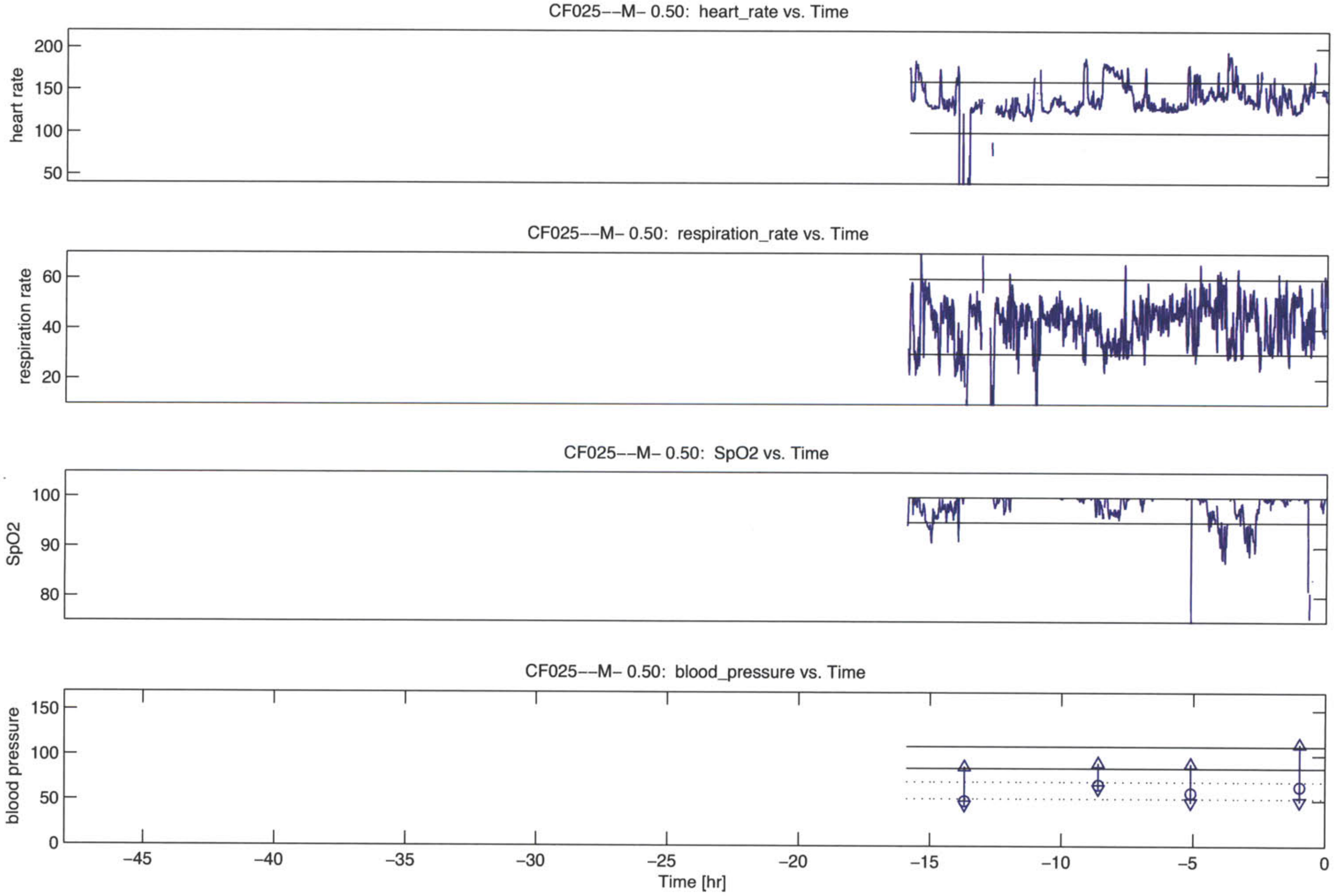
BR467--M-16.00: heart_rate vs. Time

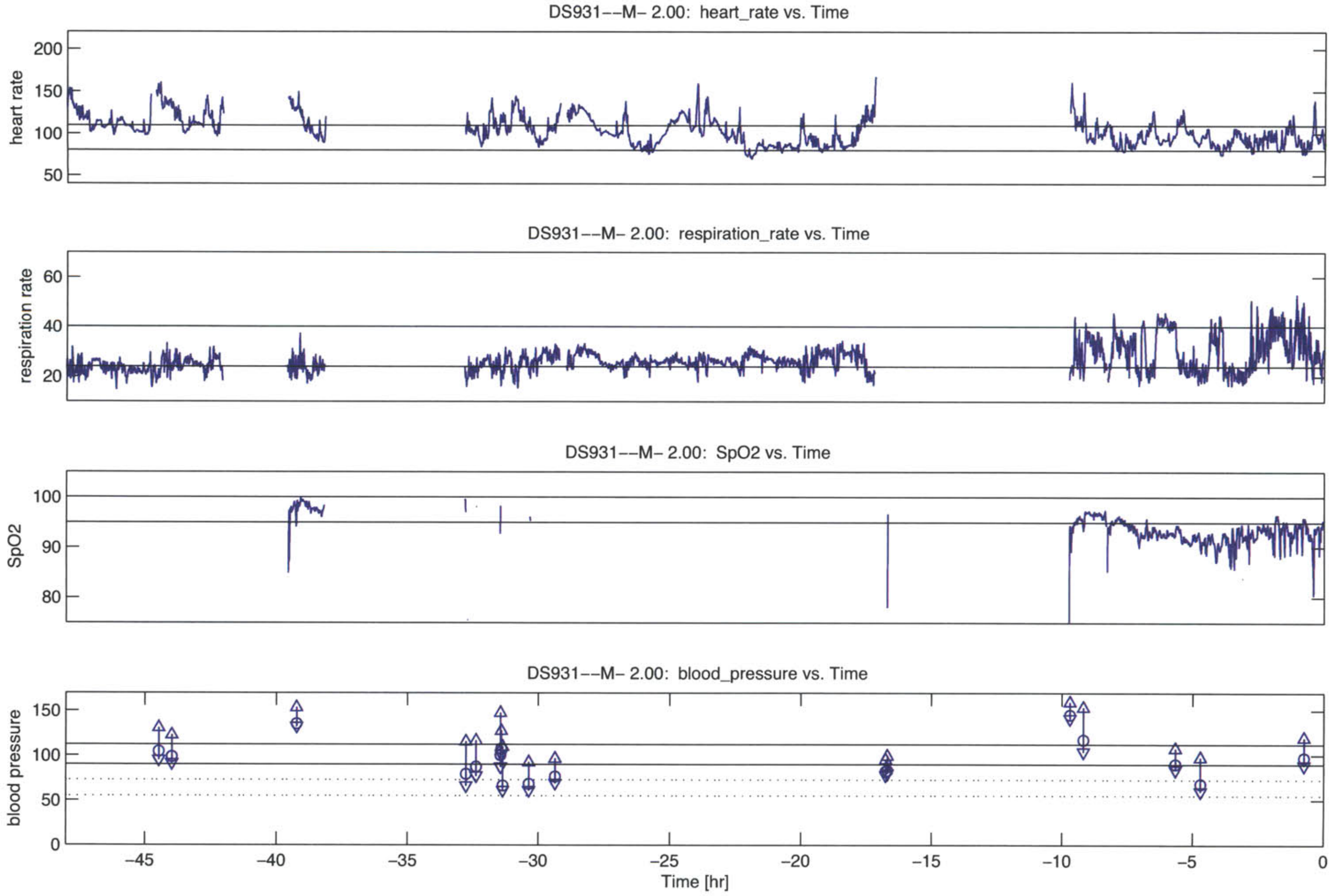
BR467--M-16.00: respiration_rate vs. Time

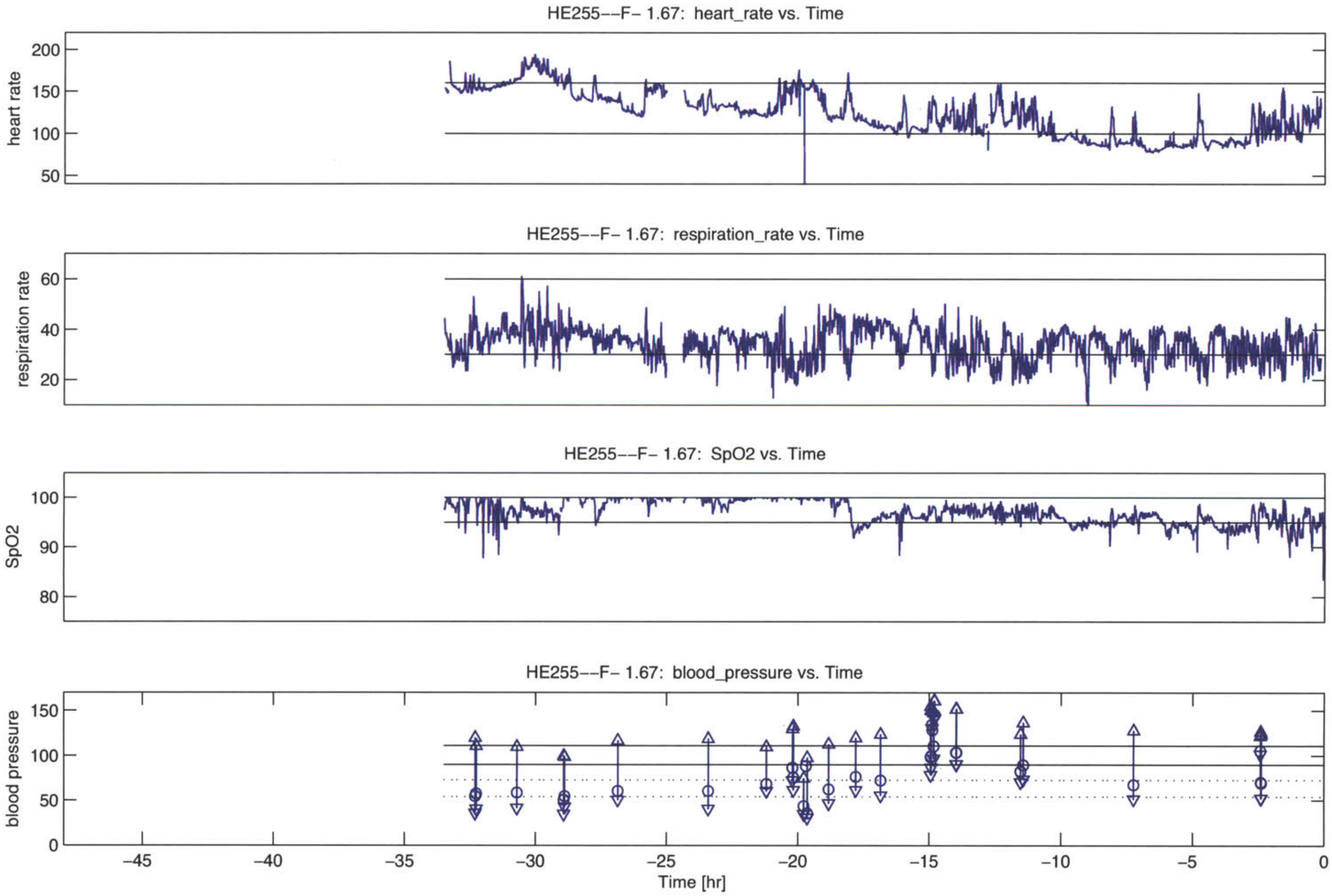
BR467--M-16.00: SpO2 vs. Time

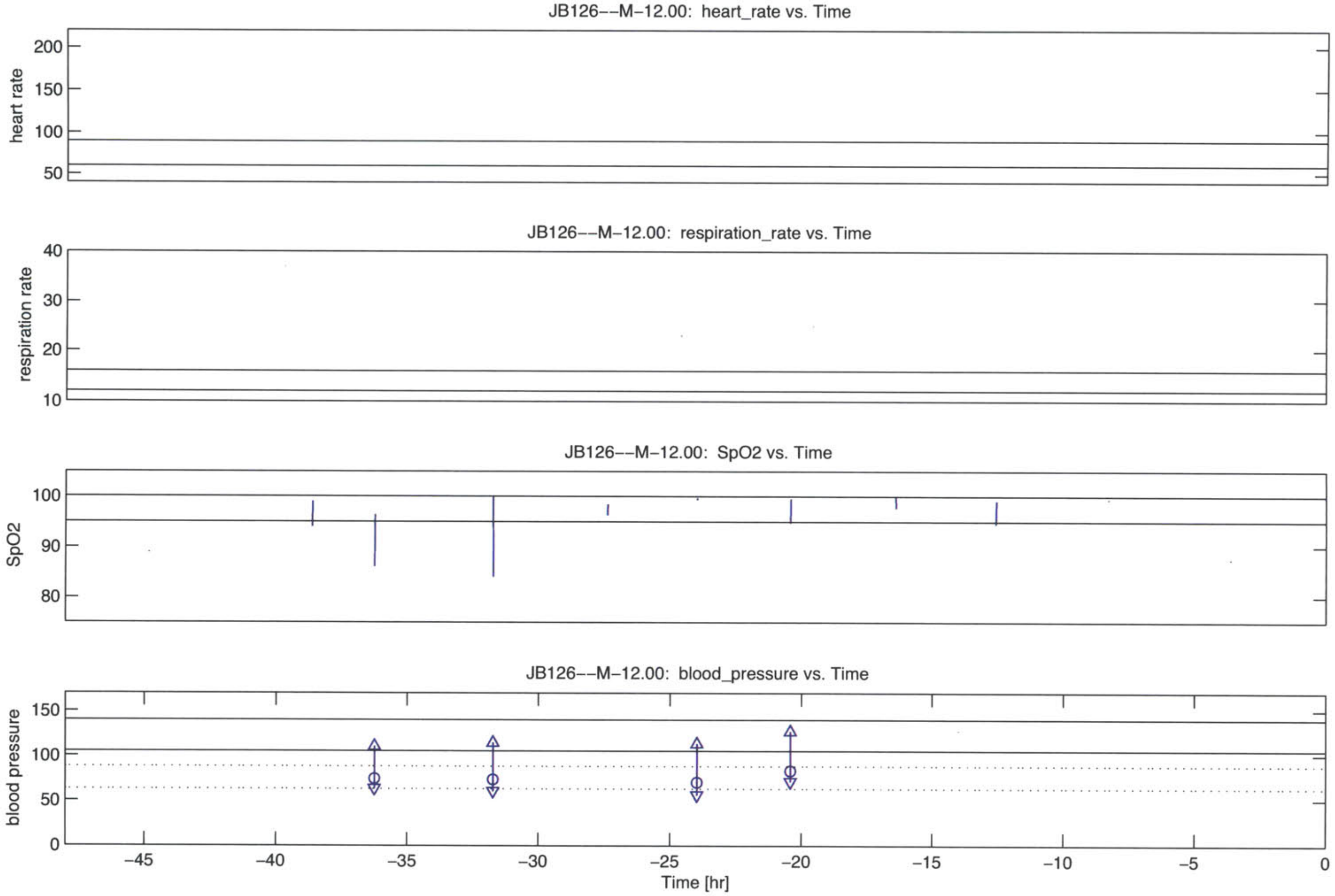
BR467--M-16.00: blood_pressure vs. Time


CB420--F-10.00: heart_rate vs. Time
heart rate
200
150
100
50
CB420--F-10.00: respiration_rate vs. Time
respiration rate
60
40
20
CB420--F-10.00: SpO2 vs. Time
SpO2
100
90
80
CB420--F-10.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


CF025--M- 0.50: heart_rate vs. Time
heart rate
200
150
100
50
CF025--M- 0.50: respiration_rate vs. Time
respiration rate
60
40
20
CF025--M- 0.50: SpO2 vs. Time
SpO2
100
90
80
CF025--M- 0.50: blood_pressure vs. Time
blood pressure
150
100
50
0
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
Time [hr]


DS931--M- 2.00: heart_rate vs. Time
heart rate
200
150
100
50
DS931--M- 2.00: respiration_rate vs. Time
respiration rate
60
40
20
DS931--M- 2.00: SpO2 vs. Time
SpO2
100
90
80
DS931--M- 2.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


HE255--F- 1.67: heart_rate vs. Time
heart rate
200
150
100
50
HE255--F- 1.67: respiration_rate vs. Time
respiration rate
60
40
20
HE255--F- 1.67: SpO2 vs. Time
SpO2
100
90
80
HE255--F- 1.67: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

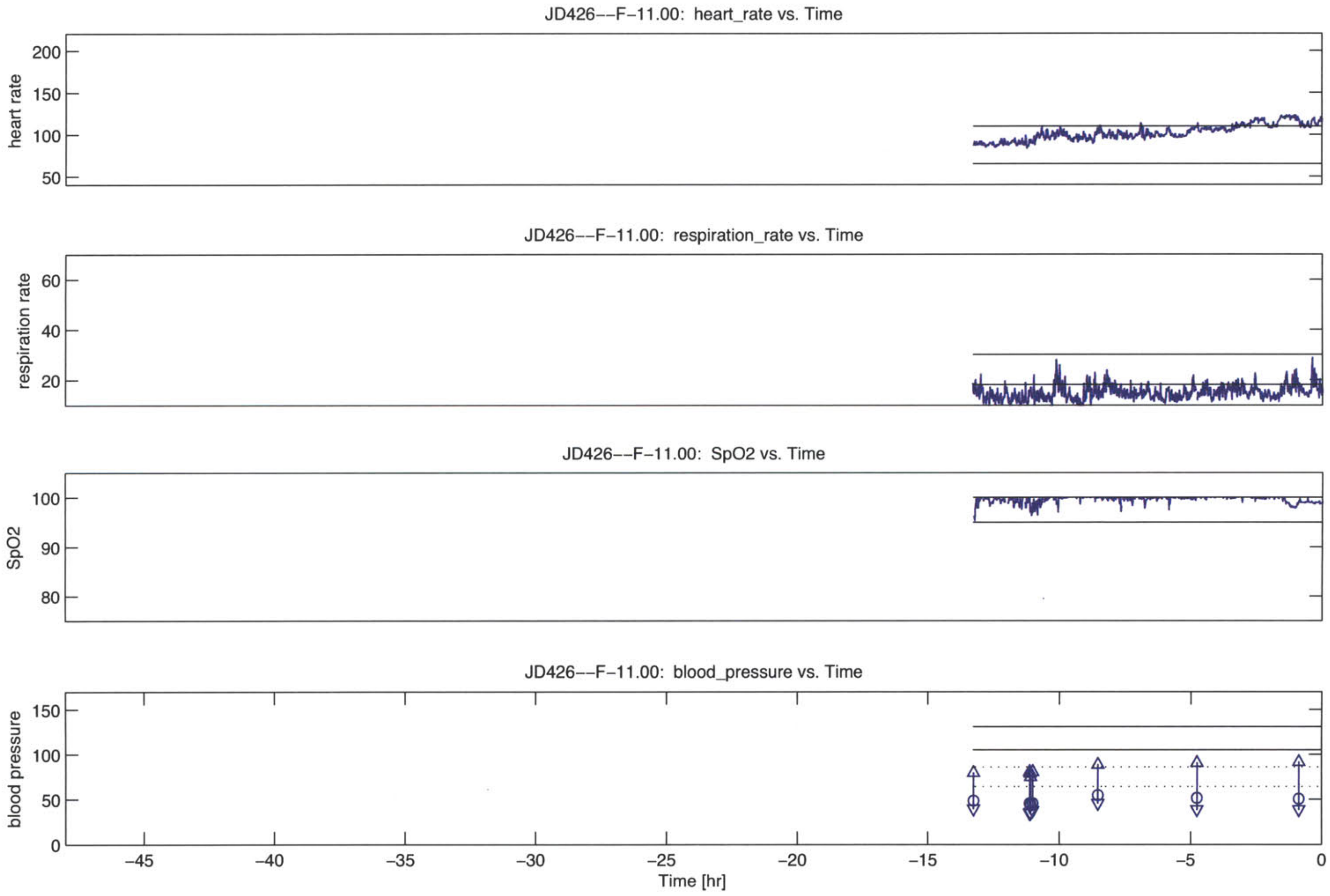


JB126--M-12.00: heart_rate vs. Time

JB126--M-12.00: respiration_rate vs. Time

JB126--M-12.00: SpO2 vs. Time

JB126--M-12.00: blood_pressure vs. Time

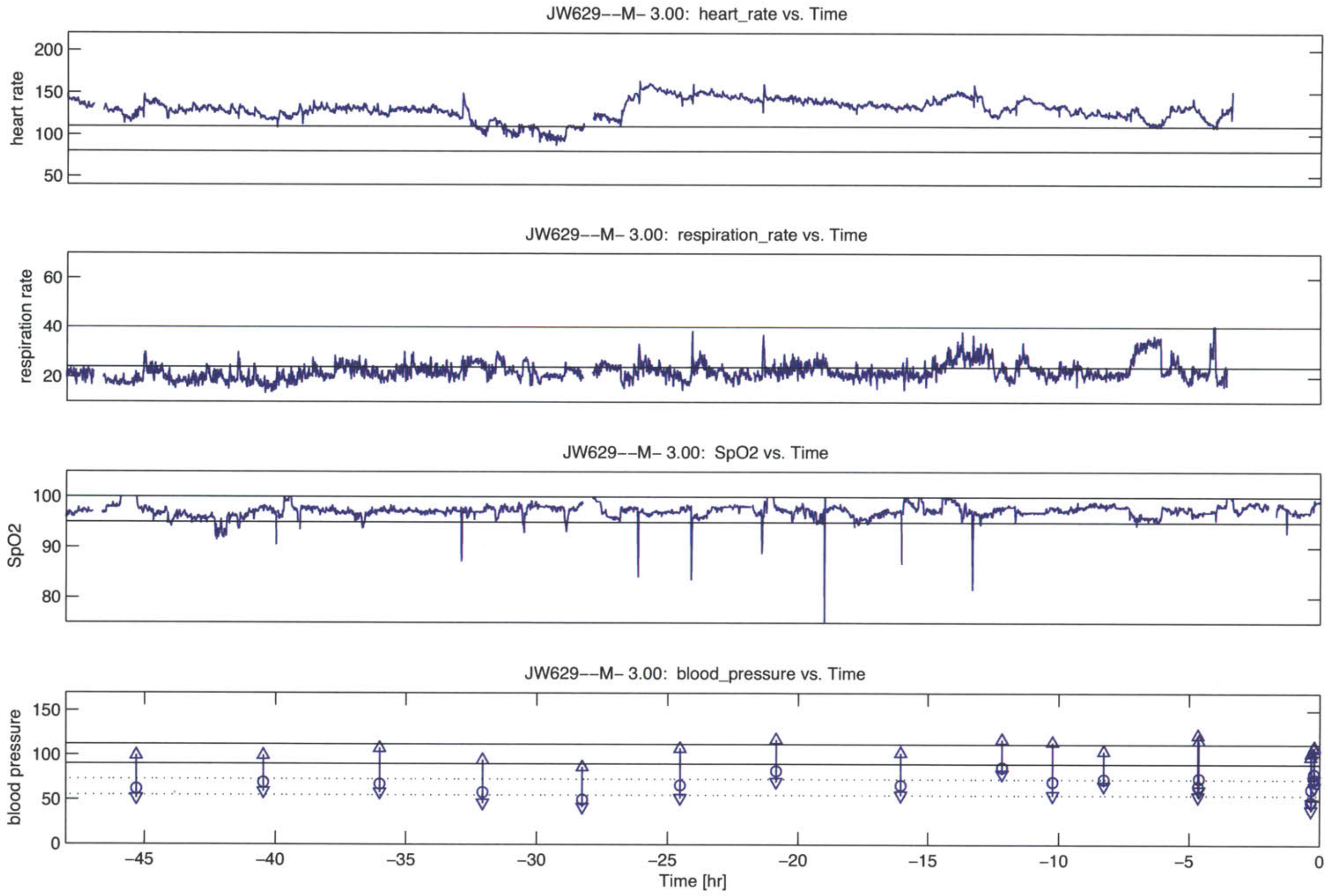


JD426--F-11.00: heart_rate vs. Time

JD426--F-11.00: respiration_rate vs. Time

JD426--F-11.00: SpO2 vs. Time

JD426--F-11.00: blood_pressure vs. Time

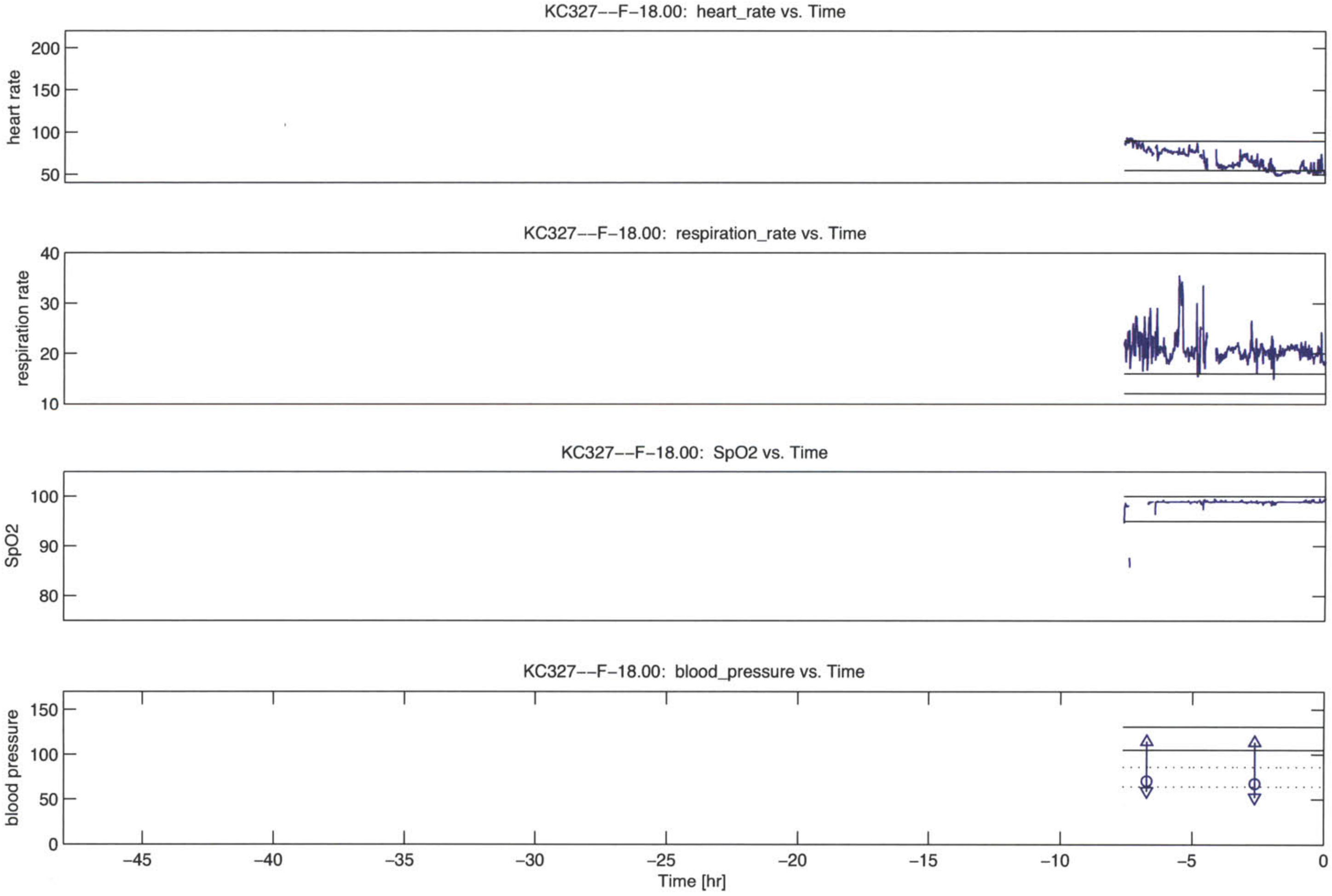


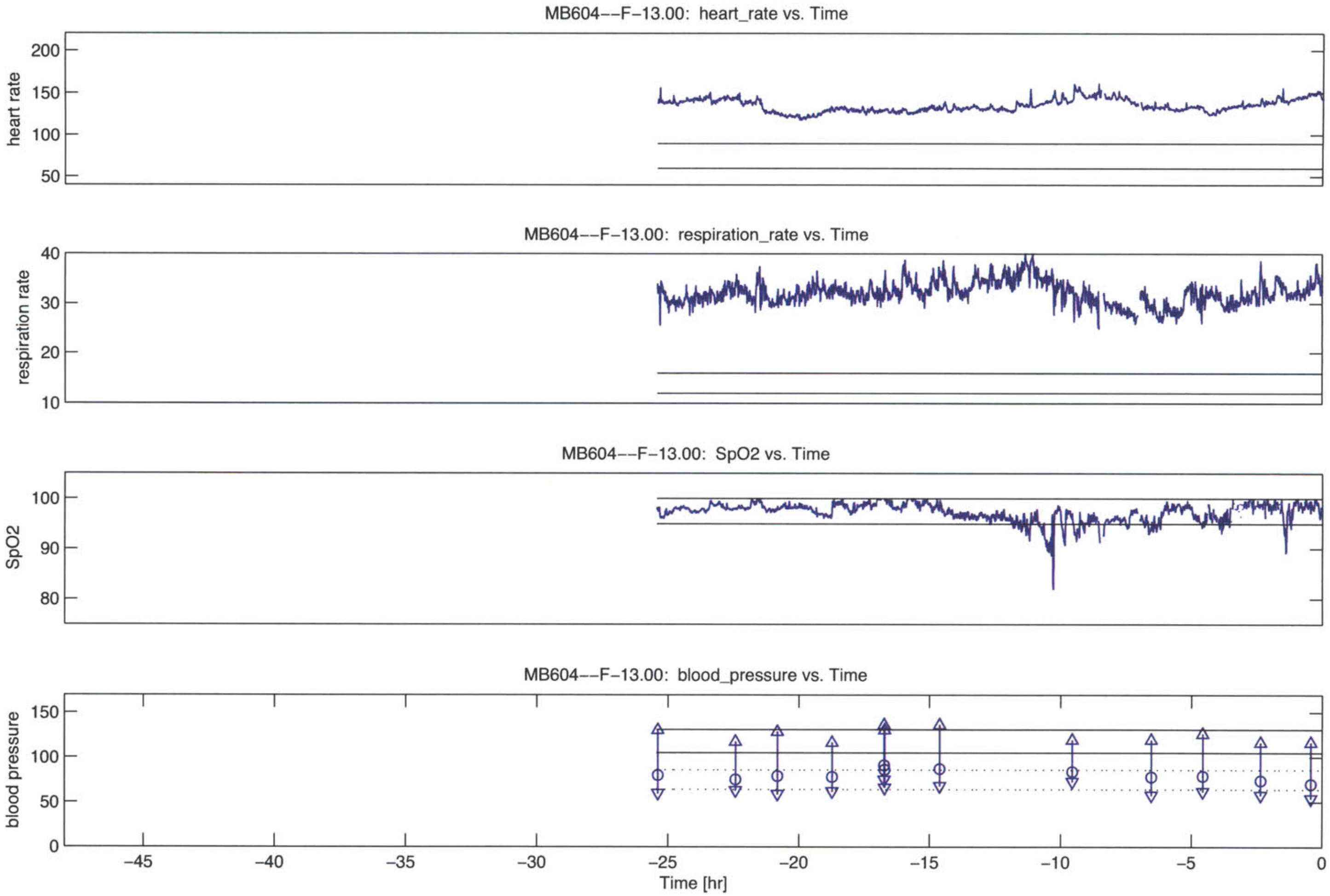
JW629--M- 3.00: heart_rate vs. Time

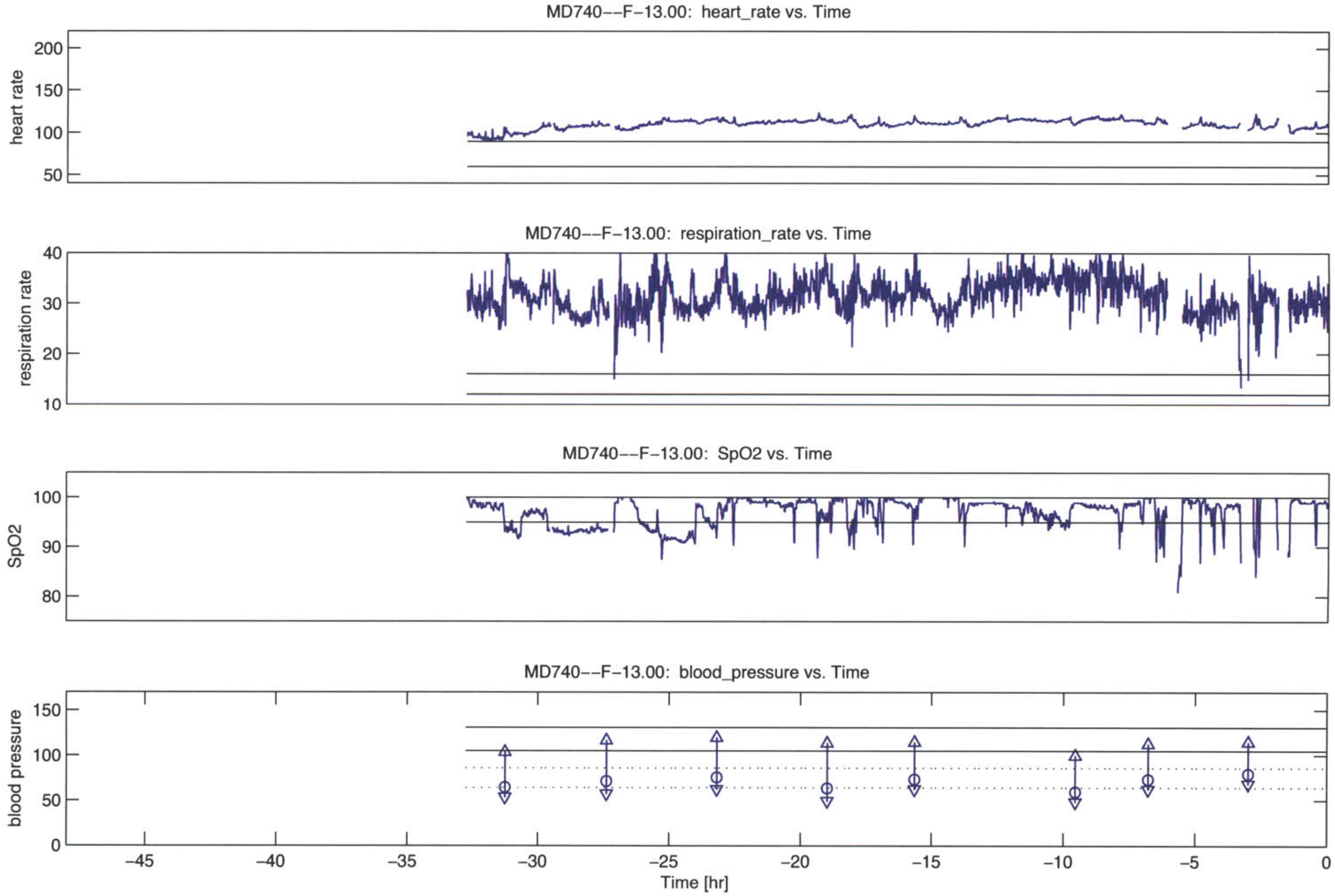
JW629--M- 3.00: respiration_rate vs. Time

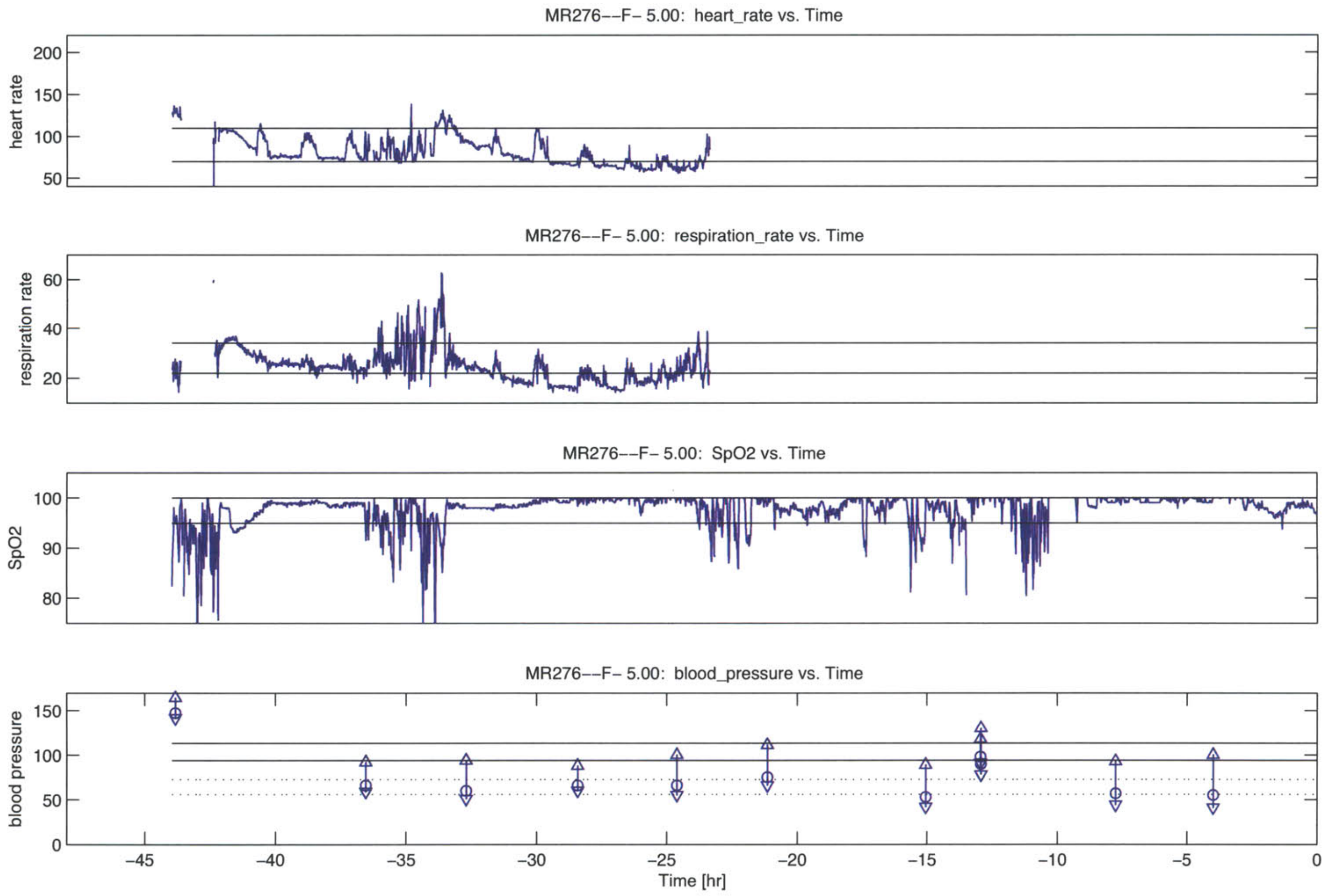
JW629--M- 3.00: SpO2 vs. Time

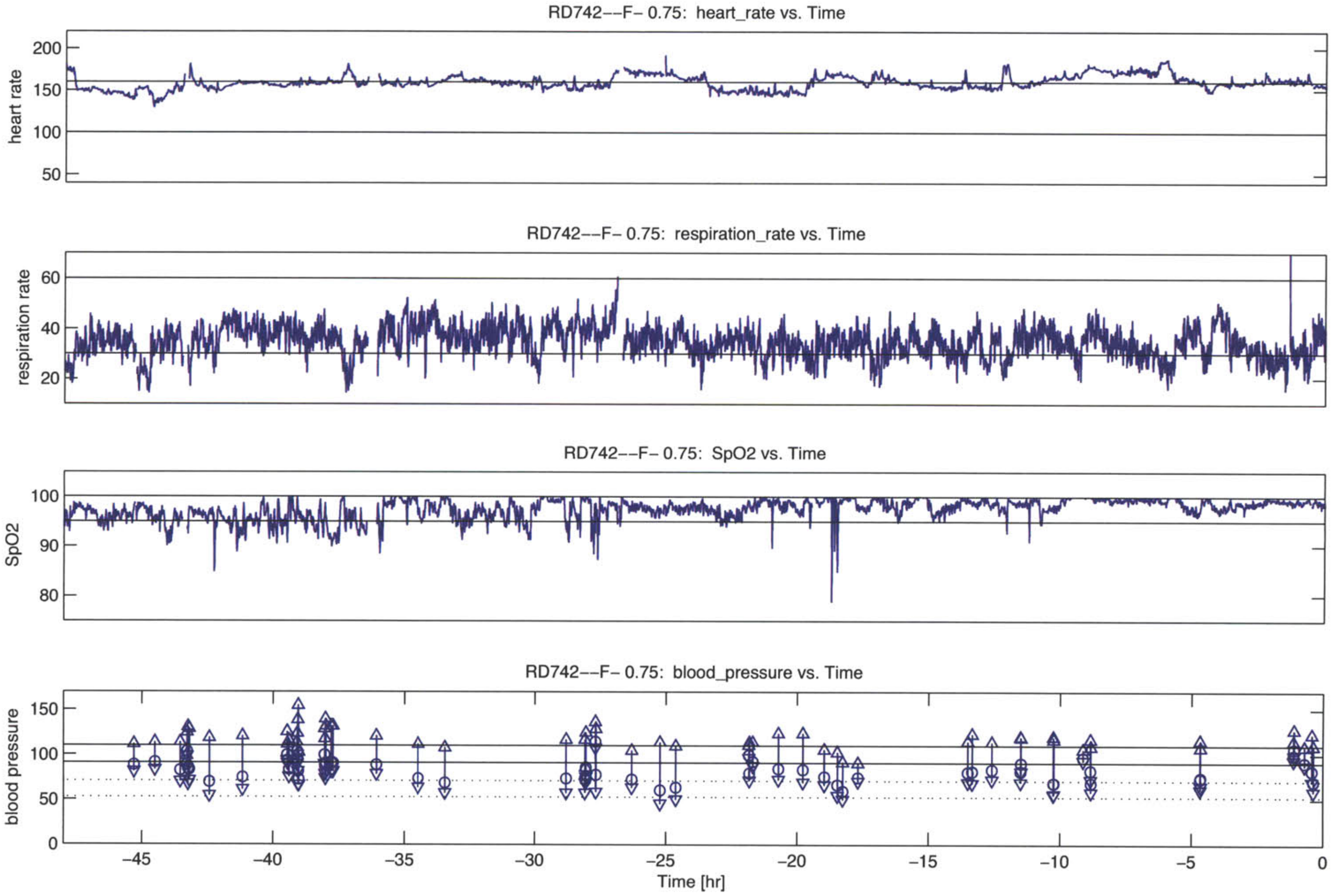
JW629--M- 3.00: blood_pressure vs. Time

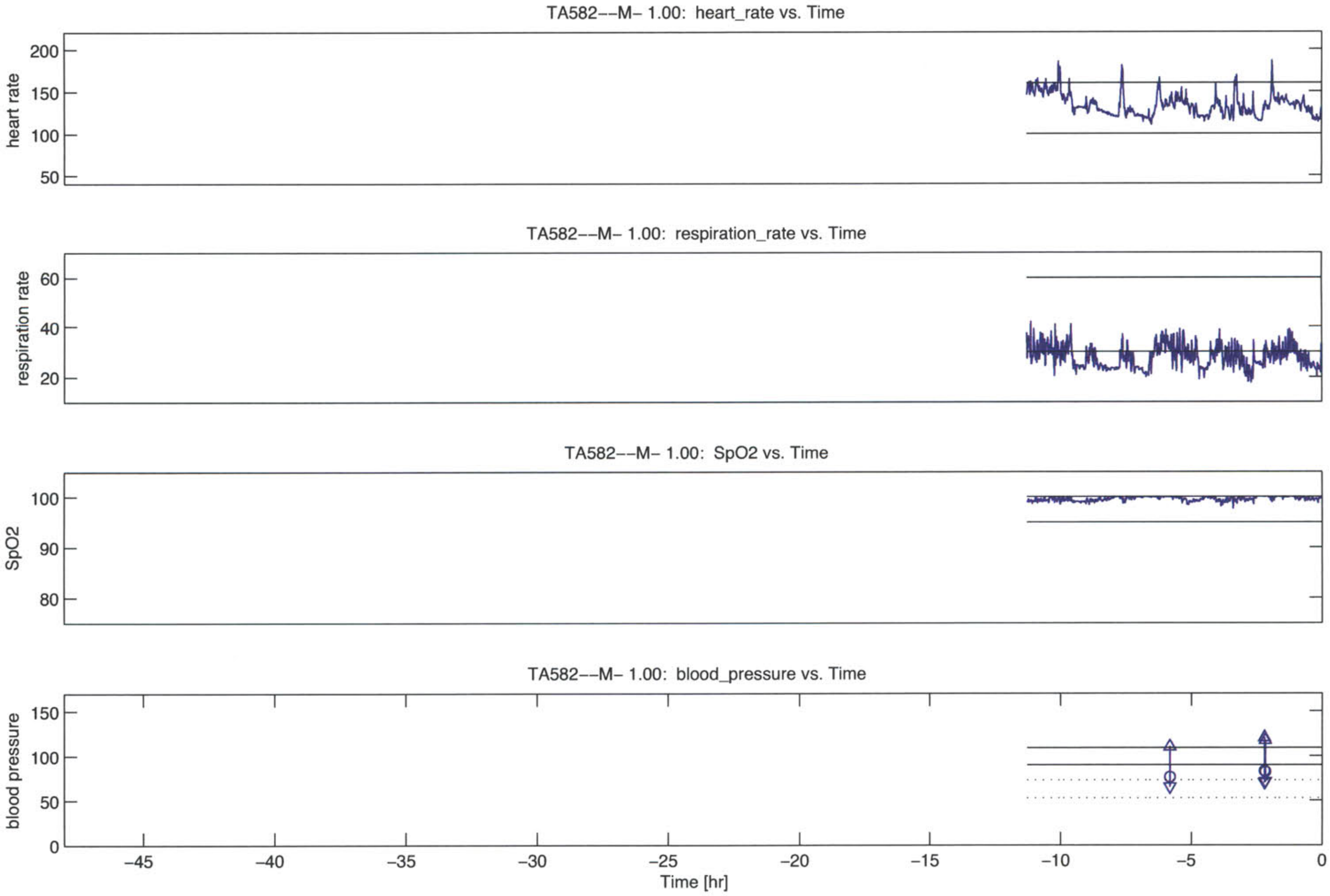

KC327--F-18.00: heart_rate vs. Time
heart rate
200
150
100
50
KC327--F-18.00: respiration_rate vs. Time
respiration rate
40
30
20
10
KC327--F-18.00: SpO2 vs. Time
SpO2
100
90
80
KC327--F-18.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

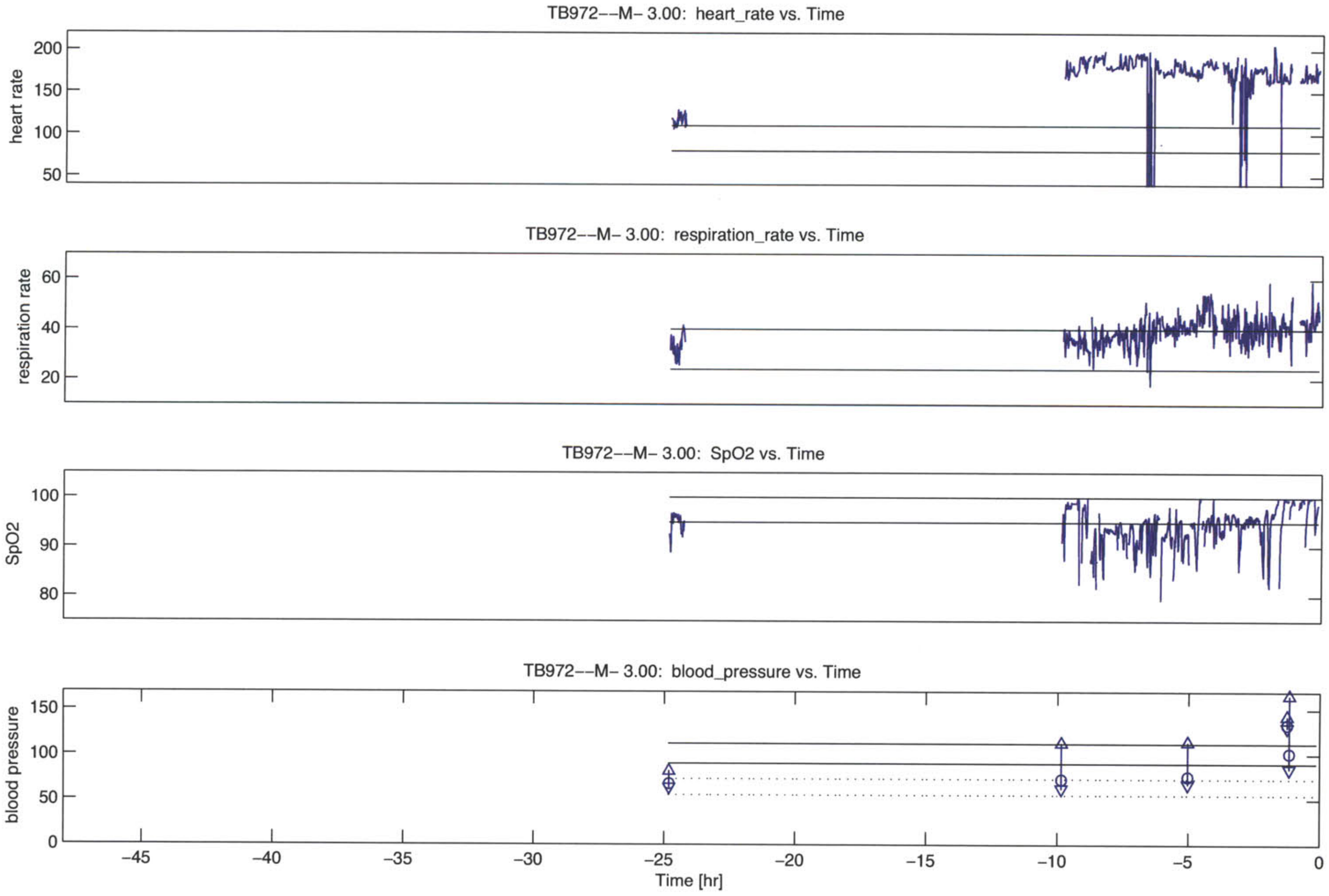

MB604--F-13.00: heart_rate vs. Time
heart rate
200
150
100
50
MB604--F-13.00: respiration_rate vs. Time
respiration rate
40
30
20
10
MB604--F-13.00: SpO2 vs. Time
SpO2
100
90
80
MB604--F-13.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

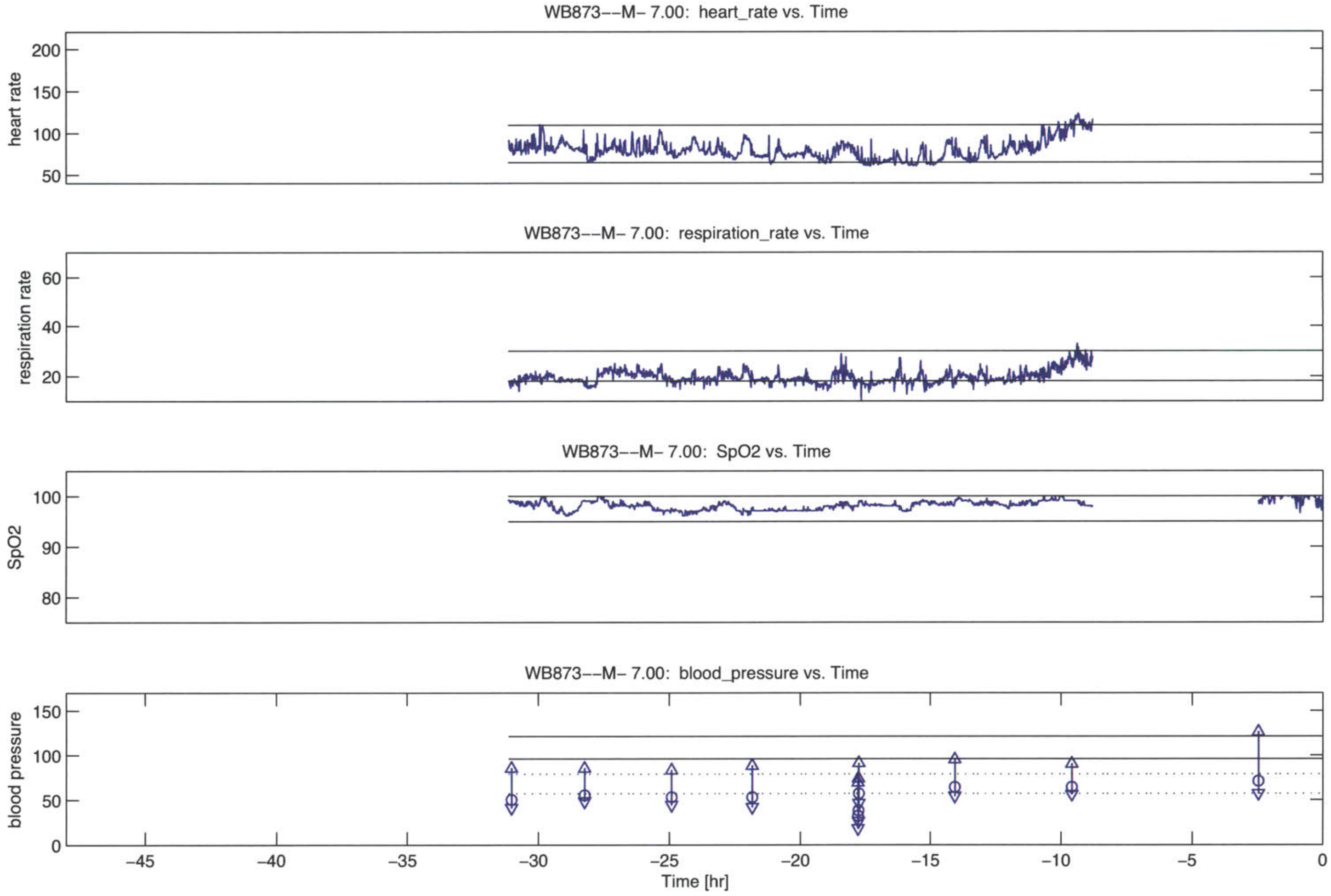

MD740--F-13.00: heart_rate vs. Time
heart rate
200
150
100
50
MD740--F-13.00: respiration_rate vs. Time
respiration rate
40
30
20
10
MD740--F-13.00: SpO2 vs. Time
SpO2
100
90
80
MD740--F-13.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


MR276--F- 5.00: heart_rate vs. Time
heart rate
200
150
100
50
MR276--F- 5.00: respiration_rate vs. Time
respiration rate
60
40
20
MR276--F- 5.00: SpO2 vs. Time
SpO2
100
90
80
MR276--F- 5.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


RD742--F- 0.75: heart_rate vs. Time
heart rate
200
150
100
50
RD742--F- 0.75: respiration_rate vs. Time
respiration rate
60
40
20
RD742--F- 0.75: SpO2 vs. Time
SpO2
100
90
80
RD742--F- 0.75: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


TA582--M- 1.00: heart_rate vs. Time
heart rate
200
150
100
50
TA582--M- 1.00: respiration_rate vs. Time
respiration rate
60
40
20
TA582--M- 1.00: SpO2 vs. Time
SpO2
100
90
80
TA582--M- 1.00: blood_pressure vs. Time
blood pressure
150
100
50
0
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
Time [hr]


TB972--M- 3.00: heart_rate vs. Time
heart rate
200
150
100
50
TB972--M- 3.00: respiration_rate vs. Time
respiration rate
60
40
20
TB972--M- 3.00: SpO2 vs. Time
SpO2
100
90
80
TB972--M- 3.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]


WB873--M- 7.00: heart_rate vs. Time
heart rate
200
150
100
50
WB873--M- 7.00: respiration_rate vs. Time
respiration rate
60
40
20
WB873--M- 7.00: SpO2 vs. Time
SpO2
100
90
80
WB873--M- 7.00: blood_pressure vs. Time
blood pressure
150
100
50
0
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
Time [hr]

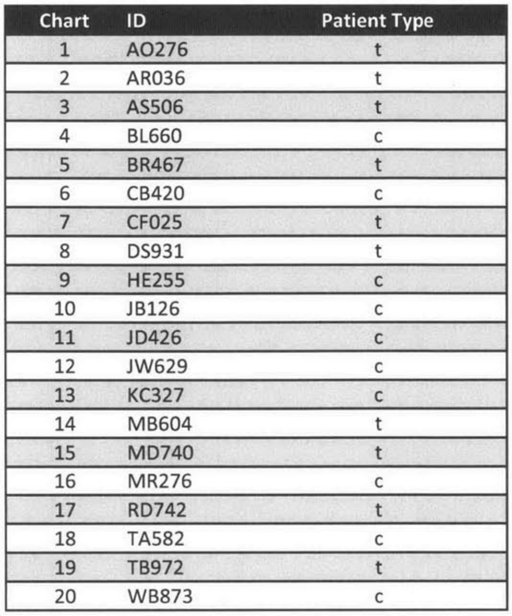


Chart	ID	Patient Type
1	AO276	t
2	AR036	t
3	AS506	t
4	BL660	c
5	BR467	t
6	CB420	c
7	CF025	t
8	DS931	t
9	HE255	c
10	JB126	c
11	JD426	c
12	JW629	c
13	KC327	c
14	MB604	t
15	MD740	t
16	MR276	c
17	RD742	t
18	TA582	c
19	TB972	t
20	WB873	c

Bibliography

[1] W. B. Jonas. *Mosby's Dictionary of Complementary and Alternative Medicine.* Elsevier Mosby, 2005.

[2] F. Sax and M. Charlson. Medical patients at high risk for catastrophic deterioration. *Critical Care Medicine*, 15:510–5, 1987.

[3] C. Franklin and J. Matthew. Developing strategies to prevent in-hospital cardiac arrest: Analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22:244–7, 1994.

[4] D. Goldhill. Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia*, 54:529–34, 1999a.

[5] P. McQuillan, A. Allan, B. Taylor, A. Short, G. Morgan, M. Nielson, D. Barrett, and G. Smith. Confidential inquiry into quality of care before admission to intensive care. *British Medical Journal*, 316:1853–8, 1998.

[6] K. Hillman, P. J. Bristow, T. Chey, K. Daffurn, T. Jacques, S. L. Norman, G. F. Bishop, and G. Simmons. Duration of life-threatening antecedents prior to intensive care admission. *Intensive Care Medicine*, 28:1629–34, 2002.

[7] D. Goldhill. Outcome of intensive care patients in a group of British intensive care units. *Critical Care Medicine*, 26:1337–45, 1998.

[8] L. Grohskopf, R.L. Sinkowitz-Cochran, D.O. Garrett, A.H. Sohn, G.L. Levine, J.D. Siegel, B.H. Stover, and W.R. Jarvis. A national point-prevalence survey of pediatric intensive care unit-acquired infections in the united states source. *The Journal of Pediatrics*, 140:432–8, 2002.

[9] D. Gravel, A. Mallow, M. Ofner-Agostini, M. Loeb, L. Johnston, E. Bryce, M. L. Sample, V. R. Roth, C. Goldman, and G. Taylor. A point prevalence survey of health care-associated infections in pediatric populations in major canadian acute care hospitals source. *American Journal of Infection Control*, 35:157–62, 2007.

[10] A. Valentin, M. Capuzzo, B. Guidet, R. P. Moreno, L. Dolanski, P. Bauer, and P. G.H. Metnitz. Patient safety in intensive care: Results from the multinational sentinel events evaluation study. *Intensive Care Medicine*, 32:1591–8, 2006.

[11] T. Clark. Impact of clinical alarms on patient safety. *ACCE Healthcare Technology Foundation*, 2006.

[12] H. B. Nguyen, E. P. Rivers, S. Havstad, B. Knoblich, J. A. Ressler, A. M. Muzzin, and M. C. Tomlanovich. Critical care in the emergency department: A physiologic assessment and outcome evaluation. *Academic Emergency Medicine*, 7(12):1354–61, 2000.

[13] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich. Early goal directed therapy in the treatment of severe sepsis and septic shock. *The New England Journal of Medicine*, 345(19):1368–77, 2001.

[14] F. McArthur-Rouse. Critical care outreach services and early warning scoring systems: A review of the literature. *Journal of Advanced Nursing*, 36:696–704, 2001.

[15] A. Monaghan. Detecting and managing deterioration in children. *Paediatric Nursing*, 17:32–5, 2005.

[16] H. Duncan. The paediatric early warning score. *Journal of Critical Care*, 21:271–9, 2006.

[17] S. Chapman, M. P. W. Grocott, and L. S. Franck. Systematic review of paediatric alert criteria for identifying hospitalised children at risk of critical deterioration. *Intensive Care Medicine*, 36:600–11, 2010.

[18] M. Akre, M. Finkelstein, M. Erickson, M. Liu, L. Vanderbilt, and G. Billman. Sensitivity of the pediatric early warning score to identify patient deterioration. *Pediatrics*, 125:e763–9, 2010.

[19] J. Tibballs, T. Duke, E. Oakley, and M. Hennessy. Reduction of paediatric inpatient cardiac arrest and death with a medical emergency team: Preliminary results. *Archives of Disease in Childhood*, 90:1148–1152, 2005.

[20] P. Sharek, K. Leong, J. Coombs, K. Earnest, J. Sullivan, L. Frankel, and S. J. Roth. Effect of a rapid response team on hospital-wide mortality and code rates outside the icu in a children's hospital. *JAMA: the Journal of the American Medical Association*, 298:2267–2274, 2007.

[21] R. J. Brilli, R. Gibson, J. W. Luria, T. A. Wheeler, J. Shaw, M. Linam, J. Kheir, P. McLain, T. Lingsch, A. Hall-Haering, and M. McBride. Implementation of a medical emergency team in a large pediatric teaching hospital prevents respiratory

and cardiopulmonary arrests outside the intensive care unit. *Pediatric Critical Care Medicine*, 8:236–46, 2007.

[22] A. Anwar ul Haque, S. Zaidi, and S. Haider. Experience of pediatric rapid response team in a tertiary care hospital in Pakistan. *Indian Journal of Pediatrics*, 77:273–6, 2010.

[23] H. Cushing. On routine determinations of arterial tension in operating room and clinic. *The Boston Medical and Surgical Journal*, 148:250–6, 1903.

[24] A. Goldberger, L. Glass, J. M. Hausdorff, P. Ivanov, R. G. Markand, J. E. Mietus, G. B. Moody, , C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101:e215–e220. **http://www.physionet.org/physiotools/edr/**, 2000.

[25] J. Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28:R1–39, 2007.

[26] D. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of IEEE*, 85:6–23, 1997.

[27] R. Schoenberg, D. Sands, and C. Safran. Making ICU alarms meaningful: A comparison of traditional vs. trend-based algorithms. *Journal of the American Medical Informatics Association*, 1999.

[28] R. Kohli-Seth. The future of bedside monitoring. *Critical Care Clinics.*, 16(4):557–578, 2000.

[29] K. Hillman, J. Chen, M. Cretikos, R. Bellomo, D. Brown, G. Doig, S. Finfer, and A. Flabouris. Introduction of the medical emergency team (MET) system: A cluster-randomised controlled trial. *Lancet*, 365:2091–7, 2005.

[30] L. Tarassenko and D. Young. Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia*, 97:64–8, 2006.

[31] MATLAB. *version 7.13.0.564 (R2011b).* The MathWorks Inc., Natick, Massachusetts, 2011.

[32] L. Costanzo. *Physiology.* Saunders, USA, 2002.

[33] R. Mark and J. Venegas. *HST.542J Quantitative Physiology: Organ Transport Systems.* Massachusetts Institute of Technology: MIT OpenCourseWare, 2004 (Accessed 25 Jan. 2012).

[34] J. Heilman. File:sinus arythmia.jpg. *Wikipedia: The Free Encyclopedia*, **http://en.wikipedia.org/wiki/File:Sinus_arythmia.JPG**, 18 Sept. 2010 (Accessed 15 Jan. 2013).

[35] M. Malik. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation*, 93:1043–65, 1996.

[36] E. Grogan, J. Morris, P. Norris, D. France, A. Ozdas, R. Stiles, P. Harris, B. Dawant, and T. Speroff. Reduced heart rate volatility: An early predictor of death in trauma patients. *Annals of Surgery*, 93:547–56, 2004.

[37] Ratznium. File:oxyhaemoglobin dissociation curve.png. *Wikipedia: The Free Encyclopedia*, **http://en.wikipedia.org/wiki/File:Oxyhaemoglobin_dissociation_curve.png**, 12 March 2007 (Accessed 15 Jan. 2013).

[38] L. Benutzer. File:arterial-blood-pressure-curve.svg. *Wikipedia: The Free Encyclopedia*, **http://en.wikipedia.org/wiki/File:Arterial-blood-pressure-curve.svg**, 12 Nov. 2009 (Accessed 15 Jan. 2013).

[39] P. Davies and I. Maconochle. The relationship between body temperature, heart rate, and respiratory rate in children. *Emergency Medicine Journal*, 26(9):641–3, 2009.

[40] J. Juskewitch, C. Salas, and W. Huskins. Reliability of the identification of the systemic inflammatory response syndrome in critically ill infants and children. *Pediatric Critical Care Medicine*, 13(1):e55–77, 2012.

[41] M. McGregor. Pulsus paradoxus. *New England Journal of Medicine*, 301:480–482, 1979.

[42] G. Natalini, A. Rosano, M. E. Franceschetti, P. Facchetti, and A. Bernardini. Variations in arterial blood pressure and photoplethysmography during mechanical ventilation. *Anesthesia and Analgesia*, 103:1182–8, 2006a.

[43] R. Luten and S. G. Rothrock. Approach to resuscitation and advanced life support for infants and children. In J. M. Baren, S. G. Rothrock, J. Brennan, and L. Brown, editors, *Pediatric Emergency Medicine.*, chapter 1, pages 3–12. Saunders, Philadelphia,PA, 1^{st} edition, 2008.

[44] P. Padlipsky and M. Gausche-Hill. Respiratory distress and respiratory failure. In J. M. Baren, S. G. Rothrock, J. Brennan, and L. Brown, editors, *Pediatric Emergency Medicine.*, chapter 2, pages 13–27. Saunders, Philadelphia,PA, 1^{st} edition, 2008.

[45] J. Turi and I. M. Cheifetz. Acute respiratory failure. In D. S. Wheeler, H. R. Wong, and T. P. Shanley, editors, *Resuscitation and Stabilization of the Critically Ill Child*, chapter 7, pages 115–23. Springer, London, 2009.

[46] A. Donoghue and R. Walls. Intubation, rescue devices, and airway adjuncts. In J. M. Baren, S. G. Rothrock, J. Brennan, and L. Brown, editors, *Pediatric Emergency Medicine.*, chapter 4, pages 37–49. Saunders, Philadelphia,PA, 1st edition, 2008.

[47] J. Arroyo, J. McCarthy, and B. King. Sepsis. In J. M. Baren, S. G. Rothrock, J. Brennan, and L. Brown, editors, *Pediatric Emergency Medicine.*, chapter 13, pages 132–40. Saunders, Philadelphia,PA, 1st edition, 2008.

[48] R. Hotchkiss and I. Karl. The pathophysiology and treatment of sepsis. *The New England Journal of Medicine*, 348(2):641–3, 2003.

[49] R. Otero, H Nguyen, D. Huang, D. Gaieski, M. Goyal, k. Gunnerson, S. Trzeciak, R. Sherwin, C. Holthaus, T. Osborn, and E. Rivers. Early goal-directed therapy in severe sepsis and septic shock revisited. *Chest*, 130:1579–95, 2006.

[50] M. Puskarich, M. Marchick, J. Kline, M. Steuerwald, and A. Jones. One year mortality of patients treated with an emergency department based early goal directed therapy protocol for severe sepsis and septic shock: A before and after study. *Critical Care*, 13:R167–73, 2009.

[51] C. S. Parshuram, H. P. Duncan, A. R. Joffe, C. A. Farrell, J. R. Lacroi, K. L. Middaugh, J. S. Hutchison, D. Wensley, N. Blanchard, J. Beyene, and P. C. Parkin. Multicentre validation of the bedside paediatric early warning system score: a severity of illness score to detect evolving critical illness in hospitalised children. *Critical Care*, 15:R184, 2011.

[52] P. Egdell, L. Finlay, and D. K. Pedley. The PAWS score: validation of an early warning scoring system for the initial assessment of children in the emergency department. *Emergency Medicine*, 25:745–9, 2008.

[53] M. M. Pollack, K. M. Patel, and U. E. Ruttimann. PRISM III: An updated pediatric risk of mortality score. *Critical Care Medicine*, 24(5):743–52, 1996.

[54] R. G. Gallager. *Principles of Digital Communication, Course materials for 6.450 Principles of Digital Communications I.* Massachusetts Institute of Technology: MIT OpenCourseWare, 2008 (Accessed 4 Oct. 2011).

[55] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 2007.

[56] M. M. Pollack, U. E. Ruttimann, and P. Getson. Pediatric risk of mortality (PRISM) score. *Critical Care Medicine*, 16(11):1110–6, 1988.

[57] M. Winterhalder, T. Maiwald, H.U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhagea. The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy & Behavior*, 4:318–25, 2003.

[58] J-R. LeGall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association*, 270(24):2957–63, 1993.

[59] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13(10):818–29, 1985.

[60] W. Zong, M. Saeed, and T. Heldt. A QT interval detection algorithm based on ECG curve length transform. *Computational Cardiology*, 33:377–80, 2006.