Modelling of Vital-Sign Data from Post-operative Patients

Marco A. F. Pimentel

St Cross College

Department of Engineering Science

Supervised by

Prof. Lionel Tarassenko

Dr. David Clifton

Submitted: Hilary Term, 2015

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of

Doctor of Philosophy

For Dad,

you taught me the value of hard work and dedication

For Mum,

without your support and unconditional love,
I wouldn't have got this far

For Filipe,

I'm lucky to have a brother like you

Acknowledgements

This dissertation would not have been possible without the help of so many people, in so many ways. First of all, I would like to express my gratitude to Prof. Lionel Tarassenko and Dr. David Clifton, for the attentive supervision, their enthusiasm for research, and their rigorous scientific approach which are an example to follow. I also acknowledge the tremendous contributions of Dr. Peter Watkinson; I am lucky to have worked with such exemplary people.

I would like to acknowledge my colleagues at the Institute of Biomedical Engineering, both the Biomedical Signal Processing & m-Health Group and the Computational Health Informatics Laboratory, for the most useful and helpful discussions during the last four years of research. Namely, I thank Mauro Santos, Julien Oster, Athansios Tsanas (also for the extensive proof reading), Ahmar Shah, David Wong, Timothy Bonnici, Vitctoria Trubody, Peter Charlton, Dave Springer, Alistair Johnson, Elnaz Gederi, Maxim Osipov, Tingting Zhu, Lise Loerup, Kate Niehaus, and Rebecca Pullon. I would also like to thank Dr. Lei Clifton, for the help in the publications leading up to my thesis, and the clinical and research team involved in the clinical trial and acquisition of the data used for the analysis described in this thesis: Sarah Vollam, Breda Lynch, Deborah Evans, Julie Darbyshire, and Sam Brown. I also acknowledge the contributions of my examiners Prof. Duncan Young (internal) and Prof. Chris Williams (external).

This work would not have been possible without the support of the IT and admin teams at the department, and my funding body: the Research Councils UK Digital Economy Programme, and *Fundação para a Ciência e Tecnologia* (FCT), Portugal.

Finally, to all my friends and family, to the ones I love dearly, a simple and sincere thank you, for supporting me throughout these last years, for inspiring me, for teaching me so much everyday, and for instilling the wonder of learning more and more in me.

Modelling of Vital-Sign Data from Post-operative Patients

Marco A. F. Pimentel

Thesis submitted for the degree of Doctor of Philosophy
St Cross College
Hilary 2015

Abstract

Thousands of in-hospital deaths each year in the UK are potentially preventable, being often preceded by physiological deterioration. The current standard of clinical practice for patient monitoring on general wards is the periodic observation of vital signs by nursing staff. The use of early warning score (EWS) systems should enable a more timely response to, and assessment of, acutely ill patients. The investigations described in this thesis seek to apply principled approaches based on machine learning to the analysis of vital-sign data from patients who are recovering from major surgery.

A dataset comprising observational vital-sign data from 407 post-operative patients taking part in a two-phase clinical trial in the Oxford Cancer Centre is introduced. A second independent dataset collected from clinical data obtained from 24,212 patients admitted to the Medical Assessment Unit of a different hospital is used for validation purposes. When applied to post-operative patients, currently-used EWS systems achieve values of Area Under the Receiver-Operating Characteristic curve (AUROC) that range from 0.717 to 0.841 for predicting a composite outcome of death, emergency admission to the Intensive Care Unit, and cardiac arrest within 24 hours. We also demonstrate that the method of recording vital signs on the ward plays a fundamental role in the design and performance of EWS systems. Using the same set of physiological variables, kernel density estimators and support vector machines give equivalent results to those of EWS systems which have been carefully optimised by trial and error.

A method for describing the physiological trajectories of post-operative patients is developed using machine learning techniques. We further introduce the concept of variability of vital signs over a 24-hour period, and propose a strategy for incorporating this information into the machine learning models studied. The resulting model leads to an improvement in performance (AUROC = 0.856). An approach based on Gaussian processes is then discussed for exploring and representing patterns of vital-sign time-series data. The approach allows different types of normal physiological trends to be identified in patients recovering from surgery.

Knowledge of different patterns among hospitalised patients and their incorporation in monitoring systems improves early-warning scoring systems for the identification of physiological deterioration in specific patient groups.

Related Publications

This thesis aggregates and extends some of the content published throughout the course of my DPhil. Related publications are listed below.

Given my presence as second or third author in [2], [4], [7] and [8], I must elaborate on which portions of the papers are my contributions, and how much of that work is included in the thesis. The code for some of the methods evaluated in [2] and [8] was my original work (included in chapter 5), except the work related to the optimisation of the SVM, which is not included in this thesis. The preliminary analysis performed in the first section in [4] was extended, and is included in chapter 2. The implementation and code for [7] was my original work, and it was extended and included in chapter 3. All work on the extensions and improvements since the papers are a result of my own work.

[**1**] M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, *A Review of Novelty Detection*, Signal Processing 99(1): 215-259, 2014

[**2**] L. Clifton, D.A. Clifton, M.A.F. Pimentel, P.J. Watkinson, L. Tarassenko. *Predictive Monitoring of Mobile Patients by Combining Clinical Observations with Data from Wearable Sensors*, IEEE Journal of Biomedical and Healthcare Informatics 18(3): 722-730, 2014

[**3**] M.A.F. Pimentel, D.A. Clifton, L. Clifton, P.J. Watkinson, L. Tarassenko, *Modelling Physiological Deterioration in Post-operative Patient Vital-Sign Data*, Medical & Biological Engineering & Computing 51(8): 869-877, 2013.

[**4**] L. Clifton, D.A. Clifton, M.A.F. Pimentel, P.J. Watkinson, L. Tarassenko. *Gaussian Processes for Personalised e-Health Monitoring with Wearable Sensors*, IEEE Transactions on Biomedical Engineering 60(1): 193-197, 2013.

[**5**] M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, *Modelling Patient Time-Series Data from Electronic Health Records using Gaussian Processes*, NIPS

Workshop on Machine Learning for Clinical Data Analysis, Lake Tahoe, USA, 2013.

[**6**] M.A.F. Pimentel, D.A. Clifton, L. Tarassenko, *Gaussian Process Clustering for the Functional Characterisation of Vital-Sign Trajectories*, IEEE International Workshop on Machine Learning for Signal Processing, Southampton, UK, 2013.

[**7**] P.J. Watkinson, M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, *Early Warning Scores for monitoring patients on post-operative wards*, 8th International Conference on Rapid Response Systems and Medical Emergency Teams, London, UK, 2013.

[**8**] L. Clifton, D.A. Clifton, M.A.F. Pimentel, P.J. Watkinson, L. Tarassenko. *Gaussian Process Regression in Vital-Sign Early Warning Systems*, EMBC 2012: 6161-6164. 34th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, San Diego, USA, 2012.

[**9**] M.A.F. Pimentel, D.A. Clifton, L. Clifton, P.J. Watkinson, L. Tarassenko. *The Observer Effect when Monitoring Acutely-ill Patients*, MEDSIP 2012. 5th International Conference on Medical Signal and Information Processing, Liverpool, UK, 2012.

[**10**] M.A.F. Pimentel, D.A. Clifton, L. Clifton, P.J. Watkinson, L. Tarassenko. *Vital-Sign Data Fusion Models for Post-operative Patients*, BIOSTEC 5: 410-413, 2012. 5th International Joint Conference (IEEE EMBS) on Biomedical Engineering Systems and Technologies, Algarve, Portugal, 2012.

Contents

List of Figures

List of Tables

Notation

We use different typeface for different objects. We write a scalar as x, a vector as $\mathbf{x}$, a matrix as $\mathbf{X}$, and a set as $\mathcal{X}$. The i^{th} element of a vector is in the typeface of a scalar x_i. The i^{th} row of a matrix is in the typeface of a vector $\mathbf{x}_i$. In the table below, notation that is commonly used in the literature (and in this thesis) is defined. Other notation (and terminology) is introduced throughout the thesis.

Sets / Linear algebra / Probability distributions / Miscellaneous

$\mathbb{R}$	The real numbers
$\|\|\mathbf{x}\|\|_p$	The L_p norm of a vector $\mathbf{x}$, by default it is the L_2 norm
$\|\|\mathbf{X}\|\|_F$	The Frobenius norm (or matrix norm) of matrix $\mathbf{X}$
$\text{diag}(\mathbf{X})$	Column vector containing the diagonal elements of squared matrix $\mathbf{X}$
$\mathbf{1}$	A vector of ones
$\mathbf{I}$	The identity matrix
$\mathcal{N}(\mu, \Sigma)$	A Gaussian distribution with specified mean μ and (co-)variance Σ
$\mathcal{GP}(\mu, k)$	A Gaussian process (GP) with mean function $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$
$x \sim p$	Variable x is sampled from distribution p
$p(\cdot)$	A probability density on a continuous space

Acronyms

Many acronyms are used throughout this thesis. We provide a list of acronyms used widely throughout the literature as well as those defined by the author.

Acronyms in common usage

AUROC	area under the receiver-operating characteristic curve
AVPU	alert-verbal-pain-unresponsive
BIC	Bayesian information criterion
BP	blood pressure
CI	confidence interval
CEWS	centile-based early warning score
DTW	dynamic time warping
ECG	electrocardiogram
EWS	early warning score
FN	false negative
FP	false positive
GCS	Glasgow Coma scale
GMM	Gaussian mixture model
GP	Gaussian process
HDU	high dependency unit
HMM	hidden Markov model
HR	heart rate
ICU	Intensive Care Unit
IQR	interquartile range
LDS	linear dynamical system
MAU	Medical Assessment Unit
MET	Medical Emergency Team
MEWS	modified early warning score
NEWS	National early warning score
NHS	National Health Service
NICE	National Institute for Clinical Excellence
NPV	negative predictive value

PCA	principal component analysis
PDA	personal digital assistant
PPG	photoplethysmogram
PPV	positive predictive value
RBF	radial basis function
RCP	Royal College of Physicians
ROC	receiver-operating characteristic
RR	respiratory rate
SD	standard deviation
SE	standard error
SpO_2	peripheral oxygen saturation
SVM	support vector machine
TP	True Positive
TN	True Negative
UK	United Kingdom
USA	United States of America
ViEWS	VitalPAC early warning score
	"Novel" acronyms
AOM	active outlier method
CALMS-2	Computer Alerting Monitoring System 2
KDE	kernel density estimate
NLML	negative log marginal likelihood

Introduction

Many deaths in hospital are predictable and potentially preventable (McGloin et al. [1999]; Smith et al. [2006a]). Strengthening the in-hospital chain of survival by ensuring the timely recognition of critical illness, and facilitating management by appropriately skilled and experienced personnel, has become a key objective of healthcare providers and policymakers worldwide (Department of Health [2000]; Institute for Healthcare Improvement [2010]; Scottish Executive Health Department [2000]).

Critical illness is often preceded by physiological deterioration (Goldhill [2005]; Hillman et al. [2002]; Kause et al. [2004]; Smith et al. [2006a]). Changes in respiratory rate, pulse, blood pressure, body temperature, oxygenation, and mental function are common (Hillman et al. [2001]), and have been used to assess the well being of patients with the tacit expectation that the observer (a member of the clinical staff) will perceive the significance of deviations from normality and respond accordingly. As treatments have become more complex and life expectancy has increased, it has become apparent that clinical staff responsible for completing routine observations on hospital wards often fail to understand the significance of abnormal recordings (McQuillan et al. [1998]; Smith et al. [2006a]). Early warning scoring systems have been implemented and used in the last two decades in hospitals worldwide (Gao et al. [2007]). These are intended to facilitate objective decision-making and, thus, aid the timely recognition of patients with potential or established critical illness outside critical care areas. The wide interest in the usage of these systems is evident from the increasing number of related scientific publications in the last decade (see Figure 2).

Despite a large and increasing volume of literature, the quality of evidence underpinning the use of early warning scoring systems is poor. Although this

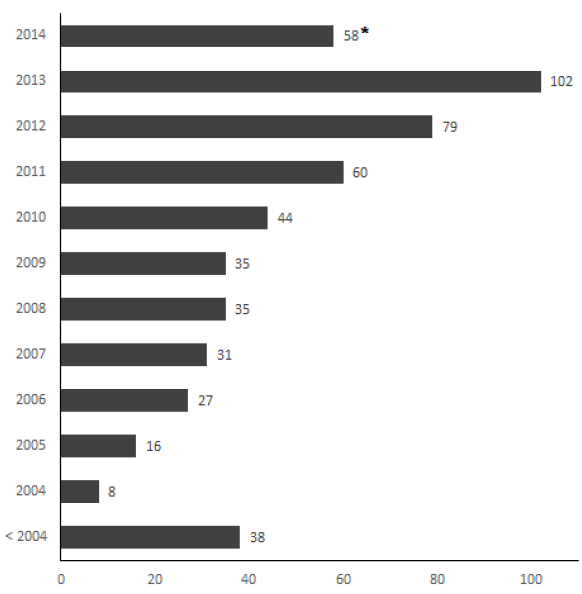


Figure 2: Number of publications on early warning scoring systems in the last decade: the number of studies (per year) reporting the performance of implementation of these systems in the PubMed database is shown. *Search was conducted in June 2014.

analysis is complicated, in part, by the heterogeneity of outcomes, which range from admission to high-dependency units or intensive care units, cardiac arrest, do-not-resuscitate orders, in-hospital length of stay and in-hospital mortality (Bell et al. [2006]; Buist et al. [2004]; Duckitt et al. [2007]; Gao et al. [2007]), recent validation and review studies have reported poor performances of such systems for different patient populations (Alam et al. [2014]; Jansen and Cuthbertson [2010]). This relatively poor performance can be explained at least in part by the following two factors:

i. the modelling techniques used in existing studies are too simple to represent the condition-specific patient data; e.g., the observations of the physiological variables are assumed to be independent and the time-based correlation between them is not taken into account;

ii. early warning scoring systems are used with periodic observations of physiological variables, which may be made as infrequently as once every twelve hours in some wards; patients may deteriorate significantly between observations.

In addition to this, the choice of clinical setting is equally important and problematic, as it is questionable whether the findings regarding a specific pa-

tient population can be extrapolated to other groups; for example, extrapolating from medical to surgical patients; other settings, in particular the general ward environment; and other circumstances, such as follow-up after discharge from intensive care. Consequently, incorporating knowledge and information about specific patients' condition and physiological changes may prove beneficial.

Failure to identify physiological deterioration in a timely manner led to the design of the Computer Alerting Monitoring System 2 (CALMS-2) trial, in which ambulatory post-operative patients with cancer were monitored using wearable sensors, while standard clinical (ward) care was also provided. This trial was designed to assess whether continuous monitoring of vital signs with computer-modelled alerting reduces the patient's length of stay in the hospital by informing clinical staff of patient deterioration more effectively than currently-used systems. CALMS-2 was a two-phase clinical trial that was conducted in the post-operative ward of the Cancer Centre, Oxford University Hospitals National Health Service (NHS) Trust, Oxford, United Kingdom (UK). Two hundred patients were recruited during each phase of the trial. Patient data were organised, reconciled and stored in a local database containing all clinically relevant information. This includes demographic, physiological (acquired both from monitors and observations recorded by nursing staff), and clinical information, such as the occurrence of events, patients' outcomes and the contents of clinical staff notes.

The investigations described in this thesis seek to apply principled monitoring approaches based on machine learning to patients who are recovering from major surgery, and incorporate knowledge and information about the conditions of these patients and their physiological changes. This includes **(1)** the analysis of two independent datasets: one dataset that contains physiological, demographic and clinical data acquired from surgical patients during their in-hospital recovery, and a second dataset collected from clinical data obtained from patients admitted to the Medical Assessment Unit of a different hospital; **(2)** the evaluation of the performance of current systems in these two patient populations; **(3)** the development of new data-driven strategies to overcome or reduce the effects of problems associated with the use of current systems; **(4)** the investigation of new biomarkers derived from signal analysis, and their incorporation in data-fusion models, for the identification of physiological deterioration; and **(5)** the

study of dynamical modelling approaches that can be used to describe different physiological trajectories of patients during their recovery from surgery.

This thesis is organised in eight chapters. In chapter 1, an overview of conventional and current monitoring systems used in hospitals is provided, and their main limitations are described. The study population and data included in the two independent datasets used in this thesis are covered in chapter 2, together with a brief description of the systems used for the methods studied (preliminary results of the work described in this chapter were reported in the related paper [**2**]). In chapter 3, the performance of current early warning scoring systems on both datasets is reported. The importance of the method of recording vital signs on the ward is studied in chapter 4. It also includes the design and evaluation of a modified early warning score system (related work was presented in the conference papers [**7**] and **9**]). Chapter 5 introduces the key concepts of novelty detection, describes the theoretical framework behind novelty detection techniques, and presents the performance of such techniques for identifying deterioration in hospitalised patients (see related publications [**1**], [**4**], and [**8**]). In chapter 6, a method for describing the physiological trajectories of post-operative patients is developed (preliminary results of this work were described in the related papers [**3**] and [**10**]). We further introduce the concept of variability of physiological variables over a 24-hour period, and a strategy for incorporating this information into the data-fusion models studied. Finally, a method for exploratory data analysis and representation of vital-sign time-series data is presented in chapter 7 (related work was presented in the conference papers [**5**] and [**6**]). To conclude, a summary of the work presented in this thesis is given in chapter 8, together with a discussion of the key findings and areas of future research.

Chapter 1

Background: monitoring patients outside intensive care

Patient monitoring systems are nowadays a fundamental component of any developed healthcare infrastructure. They are essential for providing care in operating and emergency rooms, intensive care units, and critical care units. The overall aim of patient monitoring is to give warning of early or dangerous physiological deterioration; this is typically achieved by a compromise involving many clinical, engineering, and economic considerations.

Hudson [1985] defines a patient monitoring system as

> (...) repeated or continuous observations or measurements of the patient, his or her physiological functions, and the function of the life support equipment, for the purpose of guiding management decisions, including when to make therapeutic interventions, and assessment of those interventions.

The development of patient monitoring is closely related to that of resuscitation. As a common goal of both is the welfare and treatment of patients, it is logical to progress from clinical resuscitation to the early detection or prevention of clinical catastrophes (Stewart [1970]). The history of monitoring is traced from ancient times to the invention and development of transducers and computers.

1.1 History of patient monitoring

The earliest written record relevant to the history of patient monitoring is contained in the papyrus discovered by Ebers, 1875 (translated in Bryan and Smith [1974]). This document, which is believed to have been written in 1550 BC, makes it clear that the ancient Egyptian physicians were familiar with the fact that the peripheral pulse could be correlated with the heart beat (Stewart [1970]):

> As to faintness of the heart. It is that the heart does speak or the vessels of the heart are dumb, there being no perception under the fingers.

The next contribution of importance was made some 3000 years later, at the end of the Renaissance period. In 1625, Santorio published his methods for measuring body temperature with the *spirit thermometer* and for timing the pulse rate with a pendulum. The principles for both devices had been established by Galileo, who worked out the uniform periodicity of the pendulum by timing the period of the swinging chandelier in the Cathedral of Pisa, using his own pulse rate as a timer (Graham [1957]). This was perhaps the first example of a clinical measurement. The first scientific report of the pulse rate did not appear until Sir John Floyer published *The Physician's Pulse-Watch* in 1707 (Gibbs [1971]). With subsequent improvements in the clock and the thermometer, and development of the sphygmomanometer (blood pressure cuff), the four vital signs - temperature, heart (pulse) rate, respiratory rate, and blood pressure - became the standard vital signs and, since 1920, have been recorded in standard medical charts.

The medical electronic age began when Waller [1887] recorded the electrical activity of the human heart. Further notable contributions to the science of clinical measurement in the assessment of a patient's condition were made by Mackenzie [1925], a cardiologist from the UK, who emphasised the importance of graphical records of the pulse rate and blood pressure.

The evolution of monitoring equipment was greatly accelerated by the technological resources of the electronics age, with most progress being made in the years after World War II. The development of transducers and electronic instrumentation increased the number of physiological variables that could be monitored. In

the 1950s, the concept of the Intensive Care Unit (ICU) was created, initially as post-operative recovery rooms; then more variations came about, including coronary care units monitoring cardiac rhythmicity from the 1960s. At this time, the continuous oscilloscope display of the electrocardiogram began to be widely used during and after cardiac surgery, and alarms for warning of high and low values were soon incorporated into these monitors. Later, the vital signs were monitored (Geddes [1965]), computers were used to analyse the data (Freiman and Steinberg [1964]), and facilities for "on-line" computing were developed (Jensen et al. [1966]). From the purely technological viewpoint, clear and precise requirements for patient monitoring have been described by Fisher [1968]. These are quoted, with minor paraphrasing, as follows:

> Physiological signals must still be evaluated and correlated by the observer in order to assess the immediate condition of any patient and the deduction of impending crises in many acute situations relies heavily on the intuitive skill and experience of the clinician.
>
> Early detection and identification of significant deterioration in a patient's condition must be achieved "on-line to the computer" so that data may be presented to the observer in an immediate and intelligible form requiring minimal visual interpretation if they are to be of real benefit in emergency situations.

Stewart [1970] described the development of the *Lifeline* patient monitor, which makes a diagnosis and meets the requirements for "on-line" monitoring of automatic analysis, computing, and the generation of alarms for cardiac arrest, hypoxia, or shock, with early warning of tachycardia (elevated heart rate), hypotension (low blood pressure) and other conditions, and with warning of sensor or power failure. Today's monitoring systems include not only basic signal conversion and processing, but also database functions, report generation, and some decision-making capabilities.

Concurrent with advances in patient monitoring, major changes in therapy for life-threatening disorders were also occurring. In summary, prompt quantitative evaluation of measured physiological (and biochemical) variables became

essential in the recognition of patient deterioration, prevention of adverse events, and the overall decision-making process, as physicians applied new therapeutic interventions.

1.2 Preventable and avoidable adverse events in hospitals

This thesis defines an adverse event to be an instance of deterioration of the health of the patient, which may occur due to a variety of major complications. Although deterioration can occur at any time, certain patients are most at risk (Beaumont et al. [2008]), including patients who have recently had a surgical or medical ICU admission or patients who are recovering from a critical illness.

Preventable or avoidable adverse events are a direct result of failures to identify early signs of patient deterioration and in following recognised, evidence-based best practices or guidelines at the individual and system level, which are caused by factors including misdiagnoses, failure to diagnose, delay in diagnosis and treatment, failure to follow up, or poor monitoring system performance. In the extreme case, an unattended adverse event results in a prolonged hospital stay, cardio-respiratory arrest, unanticipated admission to the ICU, or, ultimately, the death of the patient.

There is a body of evidence that shows that the majority of these adverse events are preceded by physiological abnormalities evident in vital-sign data (Buist et al. [2004]), and it has been estimated that such events could be avoided by early identification of patient deterioration (National Patient Safety Agency [2007]).

1.2.1 Evidence of physiological deterioration before adverse event

It is well-recognised that abnormal physiology is associated with adverse clinical outcomes (see Table 1.1), and a number of studies have shown that acute illness is exacerbated by "failure to act" on recognised changes. A large proportion of

patients who suffer cardio-respiratory arrest in hospital have recognisable changes in routine observations during the preceding twenty-four hours, including changes in vital signs, level of consciousness, and oxygenation (Goldhill et al. [1999]; Hillman et al. [2001]).

A multicentre, prospective, observational study (Kause et al. [2004]) found that 60% of primary adverse events (in-hospital deaths, cardiac arrests and unplanned ICU admissions) were preceded by documented abnormal physiology, the most common being hypotension and a fall in the Glasgow Coma scale (GCS)[1]. Another study (Goldhill and McNarry [2004]) found that mortality increased with the number of physiological abnormalities ($p < 0.001$), being 0.7% with no abnormalities, 4.4% with one, 9.2% with two, and 21.3% with three or more.

In the National Confidential Enquiry into Patient Outcome and Death report of 2005 (NCEPOD [2005]), the majority (66%) of patients who had been in hospital for more than twenty-four hours before ICU admission exhibited physiological instability for more than twelve hours. Furthermore, admission to an ICU was thought to have been avoidable in 21% of cases. Communication failures between teams contribute to delays in referrals and in delivering appropriate essential care, which contributes to increased morbidity and mortality. Detailed analyses of serious patient safety incidents in the National Patient Safety Agency [2007] report have shown that 11% of deaths were related to "deterioration not recognised or not acted upon" ($N = 66$).

Table 1.1 contains a summary of the results of a number of studies that have specifically looked at the association between derangements of a range of physiological parameters and the occurrence of outcomes such as death, cardiac arrest, or admission to ICU. The studies included patients from different hospital settings, used different methodologies and outcome measures, and investigated different predictor variables, which makes it difficult to undertake a detailed comparison. However, this summary shows remarkable consistency of the various results, given the degree of variation that exists in the factors detailed above. Other patient characteristics, risk, and predictor factors of in-hospital death and ICU admission (such as time of admission or discharge from ICU, severity of illness,

[1]The GCS is a scale used to assess the level of consciousness of a person. It is a scale from 3 to 15, in which 15 indicates "alert" consciousness and 3 indicates complete unresponsiveness.

Table 1.1: Association between physiological abnormalities and the occurrence of adverse patient outcomes.

		Physiological variables				
Study	Outcome	Heart rate	Resp. rate	O_2 saturation	Blood pressure	Other
Cuthbertson et al. [2007]	Admission to ICU (surgical patients)	✓	✓		✓	
Cretikos et al. [2007]	Serious events*	✓	✓		✓	Consciousness
Jacques et al. [2006]	Mortality	✓	✓	✓	✓	Arterial blood gas, urine output, level of pain, consciousness
Kause et al. [2004]	Serious events	✓	✓		✓	
Buist et al. [2004]	Mortality	✓	✓	✓	✓	Consciousness
Berlot et al. [2004]	Cardiac arrest	✓	✓			Consciousness, chest pain
Goldhill and McNarry [2004]	Mortality	✓	✓		✓	Age, consciousness
Hodgetts et al. [2002]	Cardiac arrest	✓	✓	✓	✓	Temperature, chest pain, clinician concern
Hillman et al. [2001]	Mortality	✓	✓		✓	

* Events that led to ICU admission.

and ICU quality) have been discussed in a number of other studies (Brown et al. [2012, 2013]; Kramer et al. [2012]; Renton et al. [2011]).

1.2.2 Costs of unattended adverse events

Adverse events, particularly those requiring critical care, represent a considerable burden to the health care system, but also their impact on patients and society

is probably underestimated.

The majority of serious adverse events that occur in hospital result in an unanticipated admission to an ICU. There are large variations in the ICU bed provision between countries (Rhodes et al. [2012]). In the UK, this figure remains low and compares unfavourably with other nations. Currently, the demand for ICU beds far exceeds their availability in many countries, and the shortage of beds in ICUs is an increasingly common phenomenon (Bing-Hua [2014]). Consequently, many critically-ill patients have to wait for ICU beds and be cared for in other hospital areas without specialised staff. Since these patients need early intervention to improve their outcome, a delayed or deferred ICU admission has been suggested to be associated with higher mortality (Liu et al. [2012]; Phua et al. [2010]). A number of other studies have reported similar findings. Data on primary outcomes of delayed admission to ICU due to the unavailability of ICU beds are presented in Table 1.2. Primary outcome data include length of stay in ICU, and hospital and ICU mortality.

In total, seven studies have reported on the primary outcomes for patients who had a delayed or non-delayed admission to the ICU. There is considerable heterogeneity between the studies, which precludes a deeper analysis and investigation of other characteristics of the cohort of patients considered in each study and other outcomes. Nevertheless, a significant increase in mortality rates with a delay in ICU admission has been found (Bing-Hua [2014]; Cardoso et al. [2011]; Chalfin et al. [2007]; Schnegelsberg et al. [2014]). Admission delay was associated with longer stays in ICU (Phua et al. [2010]). O'Callaghan et al. [2012] reported that ICU admission delay was associated with both an increased requirement for advanced respiratory support (92.3% delay vs. 76.4% no-delay) and a longer time spent ventilated (median four days delay vs. three days no-delay). In addition, overloading the capacity of an ICU to care for critically-ill patients may affect physician decision-making, resulting in premature discharge from the ICU, which is associated with an increased risk of early death or ICU readmission (Chrusch et al. [2009]).

Young et al. [2003] showed that the occurrence of four or more hours of delay to treatment after physiological deterioration is associated with a 3.5 times higher mortality. A recent systematic review (Vlayen et al. [2012]) on the incidence

Table 1.2: Cost of delayed admission of the ICU: hospital and ICU mortality rates are shown for different patient populations who had a delayed or non-delayed admission to ICU.

			Non-delayed ICU admission				Delayed ICU admission			
Authors	N	Population	N (%)	ICU LoS[a]	ICU Mortality	Hospital Mortality	N (%)	ICU LoS[a]	ICU Mortality	Hospital Mortality
Bing-Hua [2014]	2279	Medical patients	2094 (92)	1.8 (2.2)	140 (7)	-	185 (8)	1.9 (2.0)	16 (9)	-
Schnegelsberg et al. [2014]	277	Medical patients (100% from ED)	186 (67)	-	-	39 (21)	91 (33)	-	-	29 (32)
Robert et al. [2012]	1332	Medical patients	1139 (86)	-	-	276 (24)[b]	193 (14)	-	-	58 (30)[b]
O'Callaghan et al. [2012]	1609	Surgical and medical patients	1460 (91)	4.5 (1.8)	353 (24)	479 (33)	149 (9)	5.1 (1.9)	40 (27)	54 (36)
Cardoso et al. [2011]	401	Medical patients (70% from ED)	276 (69)	4.0 (8.0)	47 (38)	-	125 (31)	5.0 (8.5)	138 (50)	-
Phua et al. [2010]	103	Medical patients (100% from ED)	54 (52)	4.0 (5.3)	-	11 (20)	49 (48)	6.0 (10.0)	-	25 (51)
Chalfin et al. [2007]	50322	Medical patients (100% from ED)	49286 (98)	1.8 (-)	4140 (8)	6358 (13)	1036 (2)	1.9 (-)	111 (11)	180 (17)

LoS, length of stay; ED, emergency department; N, number of patients; "-", not reported.
[a]Length of stay is shown in days, median (interquartile range, IQR).
[b]These values correspond to 28-day mortality rates, i.e., occurence of death within 28 days after admission to the ICU.

and preventability of adverse events requiring intensive care admission concluded that the consequences of adverse events include a mean length of ICU stay that ranged from 1.5 days to 10.4 days for the patient's first stay in ICU and mortality percentages between 0% and 58%. The review also concluded that adverse events are an important reason for (re)admission to ICU and a considerable proportion of these are preventable. This is in agreement with findings of other recent studies (Forster et al. [2008]; Stelfox et al. [2012]).

We shall conclude with a brief discussion of the clinical implications of the presented data. Unattended events or delayed intervention after physiological deterioration prolong the patient's length of stay in the hospital, consume greater health care resources, lead to poor management decisions, and may lead to unanticipated death. Since delayed intervention is an independent predictor of hospital mortality and other poor outcomes, early recognition of patient deterioration is crucial (Smith et al. [2006a]). Action taken during early stages can prevent deterioration progressing to cardio-respiratory arrest and unanticipated ICU admission.

1.3 Conventional patient monitoring

All hospitals classify patients according to their needs. This helps them to provide the most appropriate clinical resources and tailor their services to the most critically-ill patients. In the UK, NHS patients are generally categorised as follows:

- **Level 3.** Patients requiring advanced respiratory support alone, or basic respiratory support together with support of at least two organ systems. This level includes all critical (ICU) patients requiring support for multi-organ failure.

- **Level 2.** Patients requiring close observation or intervention including support for a single failing organ system or post-operative care and those "stepping down" from higher levels of care.

- **Level 1.** Patients at risk of their condition deteriorating, or those recently relocated from higher levels of care, whose needs can be met on an acute

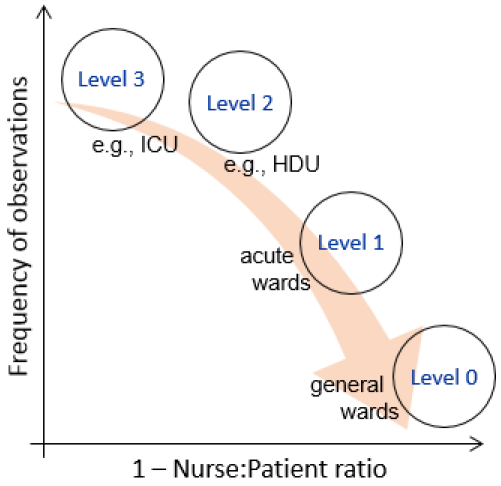


Figure 1.1: Levels of care in NHS hospitals: representation of the frequency of observation and nurse:patient ratio for the different levels of care. Level 3 includes ICUs; level 2 includes high dependency units (HDUs); level 1 corresponds to acute ward; and level 0 includes general wards.

ward with additional advice and support from critical care teams.

- **Level 0.** Patients whose needs can be met through "normal ward care" in an acute hospital.

Most patients who require critical care services fit into either levels 2 or 3. Conventional methods for monitoring the condition of patients in these levels involve the continuous acquisition of vital-sign data, which includes heart rate, respiratory rate, blood pressure, body temperature, and levels of oxygen in the blood. These are then used to trigger alarms if any single parameter exceeds a fixed threshold. Conventional patient monitoring systems are used only for those patients confined to beds (in levels 2 and 3); i.e., patients in levels 0 and 1 are often ambulatory, and so no continuous monitoring of physiological data is available using conventional patient monitors.

Patient vital signs are observed by clinical staff at periodic intervals in all hospital wards (see Figure 1.1). In level 3 wards, the typical nurse to patient ratio is 1:1, but this declines to 1:4 in level 2 wards, and to 1:10 in level 0 wards. Outside the ICU, observations of patient vital signs are typically made every four to eight hours, but the frequency of observation varies according to the patient status (NCEPOD [2012]).

There is a consistent body of evidence that shows that patients who become, or who are at risk of becoming, acutely unwell on general hospital wards receive

suboptimal care (McQuillan et al. [1998]; NCEPOD [2005]; Seward et al. [2003]). Early detection of the warning signs of deterioration may provide an opportunity for the prevention of cardio-respiratory arrest and its attendant mortality, and so the use of rapid response systems has been promoted by the UK National Institute for Clinical Excellence (NICE) as a means of reduction of in-hospital mortality (NICE [2007]).

1.3.1 Introduction of rapid response systems

Rapid response systems aim to identify deteriorating hospitalised patients prospectively and seek to alter their clinical trajectory through increasing the clinical resources directed to them. As hospitalised patients may exhibit warning signs prior to deterioration, rapid response systems have the potential to prevent adverse clinical outcomes (Devita et al. [2006]; Winters et al. [2013]). These are programs that are designed to improve the safety of hospitalised patients whose condition is deteriorating quickly. They are based on prospective identification of high-risk patients, early notification of a team of responders who have been preselected and trained, rapid intervention by the response team, and ongoing evaluation of the system's performance.

Several terms are used to refer to rapid response systems. These terms include critical care outreach, medical emergency teams, medical response teams, and rapid response teams. There are subtle differences between these terms, but all maintain two key features.

(a) *An afferent limb that includes criteria and a system for notifying the response team (i.e., how the team is activated).* Activation criteria usually include vital signs (either single-trigger criteria or early warning scoring aggregated over several vital signs) or general concern expressed by a clinical staff or family member. It defines the variables that indicate deterioration, democratises that knowledge to clinical staff, and empowers bedside clinicians to trigger the response team when the clinician has a suspicion that a patient is deteriorating (Figure 1.2).

(b) *An efferent limb that includes the response of the team.* The response team

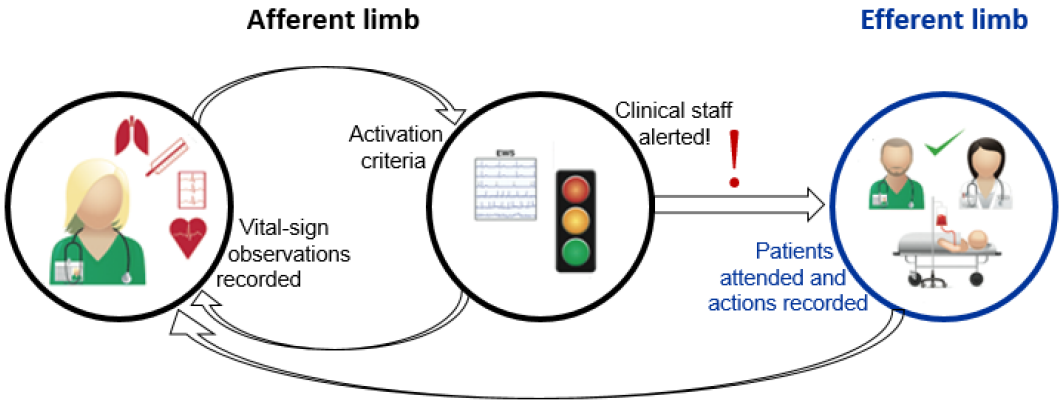


Figure 1.2: Schematic representation of the components of a rapid response system: the afferent limb and efferent limb.

most frequently comprises ICU-trained personnel and equipment. Team composition varies on the basis of local needs and resources but generally uses one of the following models: medical emergency teams (more commonly used in Australia), which typically refer to physician-led teams that have the ability to manage complex airway issues, establish central access, and initiate ICU-level care at the bedside (Hillman et al. [2005]); rapid response teams (more common in the United States of America, USA) that are generally nurse-led teams (Jones et al. [2011]); and critical care outreach (more common in the UK). The latter is slightly different from the other models in that critical care outreach also focuses on educating non-critical care staff and improving transfers between ICUs and the general hospital wards (McGaughey et al. [2007]).

Most rapid response systems also include an administrative and quality improvement team that collects and analyses event data and provides feedback, coordinates resources, and ensures improvement or maintenance over time (Winters et al. [2013]). Independent of the primary physicians who care for the patient, members of the rapid response system team can order critical laboratory and imaging studies and medication, transfer patients to higher levels of monitoring and care, and discuss end-of-life care with patients and their families.

The development of rapid response systems has grown in parallel with an increasing interest in improving hospital quality and outcomes. As a result, hundreds of hospitals have implemented these systems and teams as part of their quality improvement initiatives.

1.3.2 Track-and-trigger systems

For a rapid response system to be effective, it must have a reliable afferent limb whereby early detection and recognition of deteriorating patients occurs. This must be used in tandem with an efferent limb that provides timely, effective clinical assessment and management by a dedicated response team.

Failure of the afferent limb, that is failure to activate a response team, represents the single most significant contributor of failing to rescue a patient (NCEPOD [2005, 2012]), and therefore has a major impact on ICU admission rate and hospital mortality. Chrysochoou and Gunn [2006] reported that the response team was called in only 30% of cases when criteria for activating the response team were fulfilled, thus highlighting the challenge of afferent limb activation.

To improve early recognition of unexpected deterioration and timely attendance by appropriately skilled staff, track-and-trigger warning systems have been proposed according to the clinical guidelines issued by NICE [2007], most commonly using a scoring system based on the evaluation of vital signs known as the early warning score (EWS), first proposed by Morgan et al. [1997]. Track-and-trigger systems rely on periodic measurement of selected vital signs (the "tracking") with predetermined calling or response criteria when a certain threshold is reached (the "trigger"). These systems are drawn from routine observations of vital signs carried out by ward staff using either paper charts or, more recently, electronic devices. There has been considerable research regarding the development and use of track-and-trigger systems, which has resulted in a number of physiological scoring systems being described in the literature (Gao et al. [2007]).

Four main types of track-and-trigger systems have been identified by NICE [2007]. *Single-parameter systems*, as used by Medical Emergency Team (MET) systems, compare selected vital signs with a simple set of criteria with predefined thresholds, and then activate a response algorithm when any criterion is met

(Buist et al. [2004, 2002]). *Multiple-parameter systems* require the presence of two or more abnormal physiological variables (Goldhill et al. [1999]). This can be seen as a variant of the MET calling criteria that requires abnormalities for two different physiological parameters. *Aggregate scoring systems* assign weighted scores to physiological values that are compared with predefined trigger thresholds, and which are then summed to produce a combined score (Morgan et al. [1997]). Finally, a *combination system* includes an aggregate weighted score, but also triggers a response if any individual parameter is scored at the highest level (Sharpley and Holden [2004]).

Single-parameter systems trigger a single response strategy, while multiple parameter and aggregate warning systems allow for the on-going monitoring of a patient's condition and for a graded response strategy to be triggered, depending on the score. According to a review of these systems (Gao et al. [2007]), aggregate weighted scoring systems are the most common systems in use in the UK.

There is a variation in the type and number of physiological measures included in track-and-trigger systems. While there is a core set of parameters that are used in the majority of systems, some systems are much more complex and include a number of parameters that are not routinely measured on general wards, such as base excess, urine output, creatinine, blood sugar level, and blood oxygen and carbon dioxide partial pressures (Hodgetts et al. [2002]). There is also variation between track-and-trigger systems in the thresholds used to trigger a response, and, for aggregate scoring systems (such as early warning scores), differences in the weighting of measurements and scoring algorithms. Response algorithms also vary considerably. Many of the systems use a graded response incorporating different responses at different thresholds, typically increasing the frequency of observations at a relatively low threshold, informing the nurse in charge or junior doctor at an intermediate threshold, and informing the response team or senior doctor at a higher threshold.

1.3.3 Early warning scores

The most commonly used scoring system to evaluate vital signs is the EWS (Morgan et al. [1997]), which is an aggregate weighted scoring system. Subse-

quent modifications have led to multiple variations of the score including: modified EWS (MEWS) by Subbe et al. [2001, 2003], VitalPAC EWS (ViEWS) by Prytherch et al. [2010], centiled-based EWS (CEWS) by Tarassenko et al. [2011], National EWS (NEWS) by McGinley and Pearse [2012], HOTEL score (Hypotension, Oxygen saturation, Temperature, Electrocardiogram abnormality, Loss of independence) by Kellett et al. [2008], among many others.

Physiological track-and-trigger systems, which employ EWS-based scoring systems, have been examined in a variety of settings to determine their ability to identify patients at risk of deterioration. Considerable variation exists between the types of systems evaluated, physiological variables included, choice of trigger and the patient outcomes considered. No physiological track-and-trigger system has been identified that has been validated in a variety of populations and settings. Different systematic reviews of track-and-trigger systems conclude that the performance of most systems is poor and that they lack evidence of reliability, validity, and utility (Gao et al. [2007]; Kyriacos et al. [2011]; McGaughey et al. [2007]; Smith et al. [2008a,b]). Even considering the recently proposed evidence-based EWS systems, the latter have been found to have poor performance and to miss patients requiring assistance if used alone (Romero-Brufau et al. [2014]).

1.3.4 Cost-effectiveness of rapid response systems

Although the use of rapid response systems has broad appeal, previous studies have been limited and have reported mixed results, and individual studies have often not been adequately powered to examine the clinically meaningful outcome of hospital mortality (Calzavacca et al. [2010]; Cuthbertson [2007]; Harrison et al. [2010]; Romero-Brufau et al. [2014]; Teplick and Anderson [2006]; Winters et al. [2013]). A primary action of these systems is to triage sick patients to the ICU, and so it is critical to demonstrate that these interventions not only reduce rates of intermediate outcomes (such as cardiopulmonary arrest outside the ICU) but also reduce hospital-wide mortality, before their widespread adoption.

Track-and-trigger systems that are used in rapid response systems can be viewed as diagnostic technologies. The clinical effectiveness of a diagnostic technology is determined by the extent to which incorporating it into clinical practice

improves health outcomes. So, in most instances, the effectiveness of the technology will depend on whether the overall accuracy of identification is improved by its inclusion, its impact on therapeutic decisions, and the effectiveness of the treatments subsequently chosen (in this case, the response strategies). Ideally, randomised controlled trials (such as cluster randomised controlled trials, randomised by hospital or ward) of a diagnostic technology's ability to improve outcomes should be conducted. If such direct evidence is unavailable, it may be possible to link together separate pieces of evidence from the clinical pathway.

To date, only two randomised controlled trials have been conducted. One of the trials (Priestley et al. [2004]) used a single-hospital, ward-randomised configuration and was set in an acute hospital in England using a nurse-led critical care outreach team with a multiple-parameter track-and-trigger system. Education and training were introduced to staff sequentially, based on ward acuity level, before the implementation of the system. This study investigated two outcomes: in-hospital mortality and hospital length-of-stay. The main finding of this study was that the intervention was associated with a reduced risk of in-hospital mortality. However, two main limitations were reported: the first issue was that the clinical staff was aware of the study and therefore unusually motivated during the study, which may have biased the results. The second problem was the limited sample size of the control and intervention groups due to implementation in a single hospital. The second trial (Hillman et al. [2005]), which is known as the Medical Emergency Response, Interventions and Therapy (MERIT) study, was a cluster-randomised controlled trial that included 23 hospitals in Australia and used a single-parameter track-and-trigger system (the MET criteria). Outcomes (which included cardiac arrests, emergency ICU admissions and unexpected deaths) in 12 randomly selected hospitals using MET were compared with 11 control hospitals where no MET was in place. It was found that the intervention was associated with a higher rate of emergency calls but not associated with better outcomes.

Other studies comprise "before-and-after" clinical trials, in which the outcomes in a "before" phase prior to the introduction of the intervention are compared to the outcomes in the "after" phase in which the intervention is implemented. Among recent publications, Moon et al. [2011] conducted an eight-year before-and-after study in a teaching hospital in the UK. A critical care outreach

team with a modified EWS as the activation criteria was used. A total of 448,633 hospital admissions was reported (213,117 and 235,516 in the "before" and "after" phases, respectively). Significant reductions (as a proportion of all adult admissions) in cardiac arrests (0.4% to 0.2%), emergency ICU admissions (3% to 2%), and mortality rates (52% to 42%) were observed from the "before" phase to the "after" phase. A limitation of this study was not being able to control changes in disease patterns and admission patterns, such as the increase in overall admission rate (10.5%) and the decrease in overall emergency hospital admissions (from 31% to 28%). Studies conducted in paediatric hospitals (Sharek et al. [2007]; Tibballs and Kinney [2009]) also reported that the introduction of a MET system was associated with reduction of total hospital death and reduction of preventable cardiac arrest and death with increased survival on wards.

A positive impact of the implementation of rapid response systems was also reported in other studies conducted in Australia (Laurens and Dwyer [2011]), USA (Ott et al. [2012]) and Europe (Bunkenborg et al. [2014]; De Meester et al. [2013]; Wilson et al. [2013]). The recent review by Winters et al. [2013] concluded that the updated literature since 2008 includes low- to moderate-quality studies, and that several studies have inconsistent findings across outcomes. The elements of the response team, response team activation criteria, sample size, and reporting of outcomes vary among the studies reviewed by the authors. All of the most recent studies have used a before-and-after (historically controlled) design, which needs to be considered carefully because a recent evaluation of a multifaceted patient safety program in the UK found statistical improvements in the before-and-after comparison but not in the concurrent cohort controlled comparison (Benning et al. [2011]), as the MERIT study (Hillman et al. [2005]) did.

Furthermore, as mentioned above, many studies suggest that objective criteria for identifying deterioration are needed to help trigger the activation of a response team, since calling criteria are a crucial part of the rapid response system function. EWS systems add allocated points for each deteriorating vital sign to obtain a global score of risk. The value of this score subsequently determines whether the response team should be activated. Unfortunately, in many cases, vital signs are not reliably measured and scores are, therefore, not correctly calculated (Oliver et al. [2010]). This is where technology comes into play. In the last couple of

years, the charting of physiological variables in hospital patients has been shifting from paper-based to electronic-based systems. The use of electronic records to capture patients' vital signs is still in its infancy in the UK (Nwulu et al. [2012]). Nevertheless, it can help facilitate the measurement of vital signs, avoid transcription errors, derive an EWS automatically, prompt the clinical staff to take action, provide better integration with the hospitals' electronic record systems (Bellomo et al. [2012]), and improve the overall effectiveness of rapid response systems.

In terms of the costs of implementing and running a rapid response system (or critical care outreach system), considerable financial investment has been made in the development of such systems. NICE guidelines suggest that 90% of all in-hospital patients should be receiving (at least) 12-hourly measurements of physiological variables, excluding well patients, those receiving palliative care, and those already in critical care (NICE [2007]). The cost of performing these daily observations was estimated at an extra £3 million p/a, which was justified by NICE as being an opportunistic cost from diverting staff from other activities rather than as an additional cost. Ideally, an economic evaluation would link the effectiveness of track-and-trigger systems with the appropriate response and estimate incremental costs per quality-adjusted life-year (QALY) gained. Many track-and-trigger systems allow for graded responses, which typically result in increasing the frequency of observations (for data with lower scores) and informing more senior staff or a response team (for data with higher scores). It is therefore important to incorporate this aspect of response into any such common evaluation. Key parameters in this evaluation should include length-of-stay in the hospital, the risk of cardiac arrest and death, and quality-of-life. However, the data to perform such and evaluation are largely absent, at least in the published literature. Of the studies into effectiveness of EWS systems that were reviewed above, the overwhelming majority considered the impact of introducing some form of response or outreach service. Outreach services are complex interventions with no apparently consistent typology. The ability of such systems to generalise to new hospitals is therefore a significant problem based on the available data. Any data on the effectiveness of such a service are likely to be specific to the particular characteristics of the intervention in an individual study. Rapid response systems

as assessed in the studies have multiple components (that is, a track-and-trigger system, educational elements, and the response team itself), and so it is unclear how these individual components might separately influence outcomes.

In the present international climate of financial difficulty, the need to rationalise services and obtain quality and value for money is evermore important. The practice of evidence-based medicine is a gold standard, and so far, it appears that the provision of rapid response systems is lacking in this area. However, medical and nursing teams have in part become reliant on rapid access to experienced staff in critical care and dissolution of the service is likely to be unacceptable to many.

To conclude, there is no doubt that further research into the impact of rapid response systems and critical care outreach teams is needed. The weight of evidence is equivocal with respect to the effectiveness of rapid response systems on patient outcomes such as mortality, although subcomponents of such systems may be very important. Interpreting the evidence is further complicated by the diversity of response system configurations. On this basis, the overall cost-effectiveness of these systems compared with conventional care in its absence remains unknown.

1.4 Continuous monitoring: a clinical need, a clinical problem

The implementation of rapid response systems in acute care hospitals has focused primarily on building the efferent limb of the system - the response team. The mixed results reported on the effectiveness of these systems in reducing the occurrence of adverse outcomes have shifted the emphasis to strengthening the afferent limb of rapid response systems - the ability to detect patients at risk for these outcomes. Taenzer et al. [2011] have recently reviewed current and emerging approaches to address "failure-to-rescue"[1], and identified two main problems that may be contributing to this phenomenon.

On the one hand, intermittent vital-sign observation every 4-6 hours may

[1]Failure-to-rescue is defined by Taenzer et al. [2011] as hospital deaths after adverse events, and is used as a measure of patient safety and hospital quality.

not have enough temporal resolution to identify deterioration early, and may not be of sufficient frequency to allow trend analysis of the patient's physiological condition. In fact, adult patients hospitalised in acute care facilities who are critically ill (requiring ventilation, haemodynamic support, or cardiac monitoring) are usually admitted to either an ICU or HDU (levels 2 and 3). These patients especially benefit from frequent monitoring of vital signs and pulse oximetry, continuous electrocardiography monitoring, and higher nurse-to-patient ratios. Non-ICU/HDU beds usually constitute the majority of available beds in acute care hospitals and academic medical centres, and so most adult acutely-ill patients are admitted to "medical-surgical" units (equivalent to level 1) for which continuous monitoring is not available.

On the other hand, although track-and-trigger systems have been able to use observational data at low temporal resolution and identify deterioration with some success, vital-sign observations are by their nature interruptive; clinical staff checking vital signs will in most cases rouse the patient, either intentionally or not. As patients who were asleep or resting become roused due to the data collection and by the clinical staff, many of their vital signs change in magnitude: respiratory rate will naturally increase, subsequently increasing the heart rate and levels of oxygen in the blood, and the blood pressure may also rise. This so-called "white-coat effect" means that assessment of vital signs is not necessarily an accurate measure of patient physiology over time; clinical staff may affect the values of the measurements simply by obtaining them.

Therefore, continuous monitoring of low- to average-risk patients outside of ICUs is a challenge. Frequent vital-sign measurements performed by nursing staff are labour-intensive, may not be accurate, and can be distressing to patients, especially when those patients are trying to sleep. However, several solutions to the problem of having to provide minimally-intrusive continuous monitoring of vital signs on the general ward have emerged in recent years.

1.4.1 Monitoring systems

Continuous monitoring systems represent a proactive approach to identifying patient deterioration, based on the premise that physiological changes can indicate,

and perhaps predict, episodes of physiological deterioration. Some of the most commonly-used systems are summarised below.

Electrocardiograph monitoring has long been used with cardiac patients and has been extended to other groups of patients at risk of developing cardiac arrhythmia. Recent years have seen potential overuse of this type of monitoring (Larson and Brady [2008]), with most patients gaining little in terms of decreasing the risk of having, for example, cardiac arrest; i.e., such monitoring is not necessarily reliable in identifying patients even as they are experiencing cardiac arrest (Schull and Redelmeier [2000]).

Pulse oximetry is regarded as one of the most important technological advances in monitoring patients, especially those patients under anaesthesia. Although designed to provide instantaneous assessment of blood oxygen levels, it has also been used for continuous monitoring, specifically in the ICU and perioperative settings (Pedersen et al. [2014]). In a recent study, the authors noted that, for post-operative patients, continuous patient surveillance based on pulse oximetry with wireless nursing notification resulted in a reduced need-to-rescue events and ICU transfers (Taenzer et al. [2010]). Single-parameter scoring criteria for both blood oxygen saturation and heart rate (i.e., the outputs of pulse oximeters) were used. However, this modality carries important limitations (that are discussed below) that might prevent early detection of respiratory failure (Lynn and Curry [2011]).

A third monitoring system in frequent use is continuous end-tidal CO_2 monitoring, known as capnography. The use of this method, which up until recently has been used nearly exclusively for monitoring patients' ventilatory status during general anaesthesia, is now expanding to include procedural sedation and analgesia (Waugh et al. [2011]). It has been suggested as a tool for continuous monitoring of post-operative patients.

All three of the above-mentioned monitoring methods require sensors to be connected to the patient, which is a disadvantage when regarding monitoring for low- and average-risk patients on non-ICU hospital floors. Nevertheless, in recent years, a number of technologies that enable continuous non-contact monitoring of vital signs have emerged. To name a couple of emerging technologies, Tarassenko et al. [2014] have used remote sensing of the reflectance photoplethysmogram us-

ing a video camera which can be positioned one metre away from the patient's face. The system is able to measure heart rate and respiratory rate, and track changes in blood oxygen saturation, and is being validated in double-monitored patients undergoing haemodialysis in the Oxford Kidney Unit (Oxford, UK). Brown et al. [2014] evaluated the *EarlySense* system (EarlySense Inc., Waltham, Mass) in a medical-surgical unit of a community hospital, which consists of a piezoelectric motion-sensing device embedded in a flat sensor plate under the patient's mattress and is able to measure heart rate and respiratory rate. The sensor plate is connected to a bedside processing and display unit. The authors reported that continuous monitoring with this system was associated with a significant decrease in total length of stay in the hospital and in ICU days for transferred patients.

In the systems mentioned above, alarms are traditionally triggered using single-parameter scoring criteria. Another approach to continuous monitoring is based on the multi-parameter model proposed by Tarassenko et al. [2006]. The system fuses measurements of five vital signs (heart rate, respiratory rate, mean blood pressure, blood oxygen saturation, and temperature) into a single indicator of the patient status. The scoring system is based on a model of physiological normality derived from a large training dataset of continuously-acquired vital signs from high-risk medical and surgical patients not in the ICU. Deviations in the measured physiological values from this learnt model of normality cause the patient status indicator to rise, and sufficiently large deviations trigger automated alerts. The performance of the multi-parameter patient status model (formerly *BioSign*, now *Visensia*, OBS Medical, Abingdon, UK) was assessed in a study with 168 patients in Oxford, UK (Watkinson et al. [2006]). Although a "true alarm" rate of 94.5% was achieved, no effect on outcome measures between monitored and non-monitored patient groups has been shown. Nevertheless, more recently, the introduction of this system in a 24-bed step-down unit in Pittsburgh, USA, has been associated with reduced periods of cardio-respiratory instability in step-down unit patients (Hravnak et al. [2011]). Also, Clifton et al. [2011b] has used a system based on this data-fusion approach in patients who were admitted to the emergency department in a teaching hospital in Oxford, UK. A more detailed description of these works, as well as other related works, is provided in

the following chapters.

1.4.2 The nuisance caused by false alarms

Continuous monitoring of patients has been used outside of the ICU and operating room environments in care areas that are usually unmonitored, e.g., patients with sleep apnoea. However, continuous monitoring of selected patients in general care settings is more similar to monitoring patients in the ICU, in that alarms have a high probability of being "actionable" alarms (alarms that trigger an intervention) than the monitoring of patients not perceived to be at risk using the same alarm methods.

In general care settings, a large number of patients are monitored, which makes nuisance alarms a predominant problem. Nuisance alarms consist of false-positive alarms, usually caused by transient or motion artefacts or by sensor disconnection. Frequent nuisance alarms lead to the desensitisation of the reaction time of personnel and the subsequent disregarding of alarms. Also, in the general ward setting, the nurse-to-patient ratio is lower than in more acute wards, and physicians are typically less readily available, making the immediate attention to alarms more difficult. DeVita et al. [2010] stated "there was concern that current technology is clinically inadequate due to a potential for high false-positive or -negative rates". Thus the implementation of a continuous monitoring system requires different approaches to alarming than the use of simple univariate trigger thresholds.

The model proposed by Tarassenko et al. [2006], mentioned above, attempts to address this issue by combining information from five vital signs and introducing a notification delay. Appropriate delay eliminates many transient and motion artefact-generated false alarms by stipulating that an alarm condition must persist for a certain amount of time before that alarm is annunciated. This is increasingly used in ICU settings to help cope with the problem of alarm fatigue (Graham and Cvach [2010]). An alarm delayed by a matter of minutes may still represent a major improvement over hourly vital-sign observations, and such a delay can have a large effect on nuisance alarm rates. Mechanisms such as median filtering are another method for reducing spurious alarms (Tarassenko

et al. [2006]).

This degree of post-processing is critical in maintaining staff acceptance of a continuous monitoring system while keeping sensitivity to abnormal patient physiology at sufficiently high enough levels to affect patient safety. However, as current studies show, the promise of electronic physiological monitoring for continuous detection and prediction of deterioration has not yet been fully realised. As pointed out by Taenzer et al. [2011], one has to be aware that continuous monitoring may be prompting unnecessary responses, disturbing patients, distracting and interrupting nursing staff in their work, and leading to harmful interventions.

Summary and conclusion

The clinical need for intelligent patient monitoring systems for patients outside intensive or critical care has been reviewed. The key findings are summarised below:

- Physiological abnormalities can be a marker for clinical deterioration;
- Many hospitals have implemented rapid response systems (and physiological track-and-trigger systems) over the past 15 years to improve recognition of and response to deteriorating patients on general wards;
- Some evidence suggests that rapid-response systems are associated with reduced rates of both cardio-respiratory arrests and mortality;
- Important components of successful rapid-response systems include a system for notifying and activating a response team;
- The gradual introduction in hospitals of electronic patient records makes the use of more complex scoring systems, which are based on computerised algorithms, feasible.

Rapid response systems were introduced to identify deteriorating patients on general wards early and to respond rapidly to such deterioration with the aim of preventing serious adverse events. Different ways of identifying those at-risk

patients have been developed, including single-variable systems and EWS systems. Many studies have recently attempted to validate the accuracy of different versions of these scores. Despite the increase in research on the topic, there is still much room for improvement required by studies, particularly in clearly stating the aims and rationale of the work, and for adopting better design and analysis techniques than are currently used. To understand the implications of different studies in the field, it is important to reflect upon the underlying assumptions and rationale for the development of these prediction systems.

As argued by Smith et al. [2013], "there is an inherent assumption, indeed a clinical acceptance, that a high or rising EWS value is 'predictive' of an increased risk of an adverse event outcome, which should, in an ideal world, assure an intervention." The focus of these scoring systems (or single-variable calling criteria) was intended to identify those patients who may be at risk of significant adverse events that are potentially preventable or salvageable. The aim cannot simply be the prediction of in-hospital mortality, because most of in-hospital deaths are expected to be "unavoidable" (Cuthbertson et al. [2007]). The frequently-described adverse outcomes in studies surrounding rapid response systems include unplanned ICU admissions, cardiac arrests, and unexpected deaths. One flaw of studies examining the accuracy of different scoring systems is that the proposed clinical end-points may not be closely related to the purpose of these systems. For example, it is arguable that mortality is not the most sensitive and appropriate end-point with which a scoring system should be developed and validated.

The metrics used are an equally important and often underrated factor in evaluating the performance of physiological track-and-trigger systems. All such systems attempt to strike the right balance between providing early warning of patient deterioration and generating false alerts, which are known to lead to failure to respond to true alerts. However, in any real-life application domain, it is very difficult (if not impossible) to keep a functioning system in which the false alert rate is above 40%. This may be even more critical for systems deployed in the hospital. The popular choice of the area under the receiver-operating characteristic curve (Metz [1978]), may thus lead to misleading results and incorrect interpretations of the performance of such systems. Furthermore, given the inherent goal of track-and-trigger systems, it is important to consider other metrics

that evaluate how long before the adverse event can the system identify deterioration and alert the clinical staff, so timely attendance at the patient bedside may occur. The latter may be difficult to compute in practice depending on the design of the study.

There is also a great need for studies to use a representative patient sample. Medical and surgical patient populations may each have different case mixes and different underlying risk profiles. Results obtained from a particular subpopulation should not be generalised to other patient subpopulations. Validation studies are preferably performed by researchers independent of the original study in a different setting. There is also a need to carefully design the study, such as choosing the most sensitive time-frame (8 hours vs. 24 hours, for example). The clinical practice of vital-sign measurement and documentation, and the methods used to handle missing vital-sign data and other predictive variables should be described. Furthermore, given the multiple sets of vital signs per patient (e.g., one patient may have five sets of vital-sign data before suffering a cardiac arrest within a 24-hour period, and another patient may have eight sets), the exact method for how such data should be handled needs to be carefully considered. It is often unjustifiable to assume the independence of multiple observations from the same patient. It is quite likely that sicker patients may have more observations recorded than stable patients before they suffer an adverse event. Without taking this factor into account, studies may produce artificially-inflated performances or misleading results.

Finally, technology is advancing in the detection of the critically ill. The gradual introduction in hospitals of electronic patient records, means that the use of scoring systems which are based on computerised algorithms and data-fusion methods, rather than simple manually-computable scores, is becoming feasible. Incorporating information from continuous monitoring devices with the periodic observations made by clinical staff may also lead to monitoring systems that have the potential to track the physiological condition of patients better and alert staff to abnormalities earlier and with greater accuracy, when compared with existing systems.

Chapter 2

Study populations and data extracted

For the investigation described in this thesis, a dataset acquired from patients who are recovering from surgery for the removal of upper gastrointestinal cancer was used. Gastrointestinal cancer refers to malignant conditions of the gastrointestinal tract and accessory organs of digestion, including the oesophagus, stomach, biliary system, pancreas, small and large intestine, rectum, and anus. It is the most common type of cancer in Europe (Hellier and Williams [2007]; Williams et al. [2007]). Out of 2.1 million new cancers in Europe in 2000, gastrointestinal cancers accounted for over 579,000 or about 28% of the total. In the UK, there are about 60,000 new cases of gastrointestinal cancer each year (Williams et al. [2007]). The management of upper gastrointestinal cancers depends on the stage of disease and on patient fitness. Patients with disease that has not spread are considered for surgery or radiotherapy, and chemotherapy can also be of benefit.

Any surgery for gastrointestinal cancer is a major operation and it takes some time to recover; typically, about a week (Pearse et al. [2006]; Wilkinson [2011]). Post-operative patient populations are likely to suffer adverse clinical events. While not everyday occurrences, these events occur in most hospitals and are associated with high mortality and morbidity because the workflow of post-operative wards is typically not designed to allow clinical staff to detect these events early (Buist et al. [2002]; Calzavacca et al. [2008]). Even when rapid

response systems and in-house intensivists are summoned and do address these problems, often the opportunity for achieving an optimal outcome is missed. More commonly, these patients are transferred under emergency conditions to ICUs, leaving nursing staff on the general ward with missed opportunities to learn from such complications. Longitudinal studies have demonstrated that any post-operative complication occurring within 30 days of surgery, no matter how trivial, is significantly more important than both pre-operative patient risk and intra-operative factors combined, with respect to long-term reductions in survival regardless of whether or not there was an initial full recovery (Hamilton et al. [2011]; Khuri et al. [2005]).

In the post-operative period, patients are confined to a hospital bed for the first day of their recovery, equivalent to the level of care in a level 2 ward (a high-dependency unit). Some patients may be electively admitted to an ICU for the first 24 hours, depending on the operation and level of anaesthesia. For the remainder of their recovery period, patients are encouraged to walk further on the ward each day; i.e., they are ambulatory, equivalent to those patients in a level 1 ward. During this period, the standard (or "normal ward") level of care is provided to patients, which involves routine observations of the patient. These routine observations comprise spot checks that include the measurement of physiological variables, with the use of a track-and-trigger system enabling the activation of a response team if patient deterioration is identified. The brief observations that come from an array of clinical and non-clinical visits, as well as the regular observations are separated by significant time spans during which no monitoring occurs. The regular observations are typically performed every 4 hours.

2.1 The Computer Alerting Monitoring System 2 trial

The CALMS-2 study was carried out at the Oxford Cancer Hospital in the Oxford University Hospitals NHS Trust (Oxford, UK), and was approved by the Oxford Research Ethics Committee OxREC No.: 08/H0607/79, EudraCT No: 2011-

000928-15. This "before-and-after" design prospective trial was designed to assess whether the introduction of continuous monitoring of vital signs paired with a computer-modelled alerting of patient deterioration, "Visensia" (based on the work by Tarassenko et al. [2006]), reduces patient length-of-stay in the hospital after upper gastrointestinal surgery in comparison with the current paper-based track-and-trigger system.

Patients in this study were connected to conventional bedside monitors during the first day on the post-operative ward. However, as is common in most hospital wards, patients are mobilised after the first day, to gain exercise by walking around the ward, as previously described. This demonstrates the difficulty of monitoring the majority of patients in hospital, because they are ambulatory, hence the motivation for the use of wearable and portable monitors to perform predictive monitoring.

Continuous wearable monitoring devices are widely available, despite the disadvantages of high false-alarm rates caused by noise and motion artefacts due to patient movement. The system deployed in this study used mobile pulse oximeters manufactured by Nonin Medical, Inc. (for the acquisition of the photoplethysmogram, PPG, from which the peripheral oxygen saturation (SpO_2) and heart rate (HR) may be derived)[1]. These wearable devices were configured to communicate via Bluetooth to a patient-worn personal digital assistant (PDA) device, which collected the PPG data at 75 Hz. These waveforms, along with derived estimates of HR and SpO_2, were transmitted to a central server via the hospital wi-fi network. The central station stored the data, along with other patient information that will be introduced below. Manual measurements of blood pressure (BP) and respiratory rate (RR) made by ward staff were entered into the patient PDA, and were then automatically transmitted to the central station, where they were associated with the continuous data described above. Figure 2.1 shows the monitoring infrastructure for the clinical trial, during which bed-confined and ambulatory patients were monitored in parallel using continuous monitors.

A pre-and-post intervention design was used, as a randomised controlled trial was not possible due to the unblinded intervention being assessed. Were patients

[1]Note that the alarm functions of these wearable monitors were deactivated; the devices were used only for continuous data acquisition

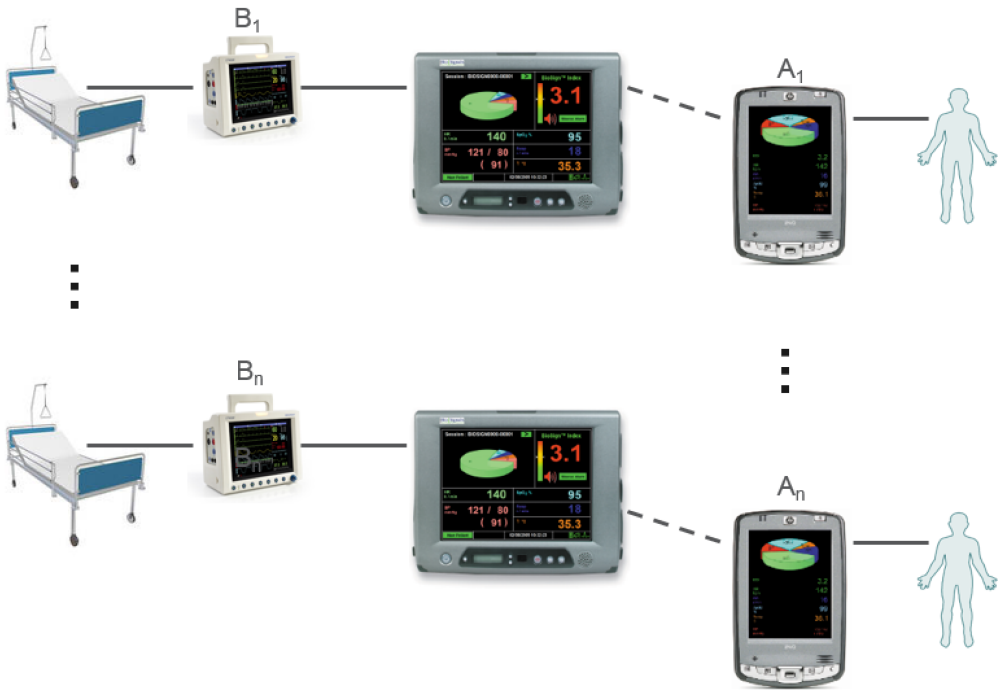


Figure 2.1: Schematic representation of the infrastructure for the CALMS-2 clinical trial. Patients confined to beds are connected to conventional bedside vital-sign monitors (denoted B_n), which in turn are wired to a patient monitor on which the computer-modelled alerting system is implemented. When patients are ambulatory, they have their vital signs measured using mobile sensors (denoted A_n), which are connected wirelessly (using the hospital network) to the patient monitors. Solid lines denote wired connections and dashed lines denote wireless connections. Figure adapted from Clifton et al. [2009].

on the same ward to be monitored with the two different systems, as would be required in a randomised controlled trial in which patients randomly receive the new intervention or not, the alerting system might cause a "learning effect" (recognition between deteriorating vital signs and the computerised score), and treatment of the control group would be altered.

During the study, all elective upper-gastrointestinal surgical patients were managed on a single ward. Three to five patients were admitted each week for major surgery and up to twelve patients were deemed suitable for monitoring at any one time. The study was divided in two phases, as follows:

Figure 2.2: The computer-modelled alerting system scores for all bed-confined and ambulatory patients are reported on a centrally-located display during Phase II of the trial. Figure adapted from Clifton et al. [2009].

- **Phase I** (from May 2009 to June 2011): consented patients were monitored using conventional track-and-trigger scoring systems (for the activation of a critical care outreach team) until deemed fit for discharge from the hospital by the surgical team (as is the normal process on the ward). Simultaneously, all consented patients were monitored from first admission to the ward after surgery with a normal bedside monitor (with the study system described in Figure 2.1). Patients were then transferred to a telemetry monitoring system (designed to be portable) which allowed continuous monitoring of a subset of vital signs combined with intermittent recordings of the other vital signs (during routine observations by nursing staff) until the patient was deemed fit for discharge by the surgical team. The results of the alerting system (i.e., the automatically-computed vital signs score) were not available to the attending staff; that is, the displays of the systems were blanked such that clinical staff could not see the alerting scores associated with each patient during this phase.

- **Phase II** (from June 2011 to December 2012): consented patients were monitored as in Phase I, but the computed scores were displayed to the

clinical staff, as shown in Figure 2.2. The computer-modelled alerting system (Visensia) was used overtly in clinical decision-making, with staff using a clinical response procedure that involved a critical care outreach team. Throughout this phase of the trial, the standard track-and trigger system remained in use, in accordance with normal care on the ward.

This two-phase design allowed a "baseline" period of data to be acquired during Phase I, against which the outcomes from Phase II might be compared to determine the effect of introducing the computerised alerting system. In the month that preceded each phase, research nurses and clinical staff received training in the processes associated with the trial and in the use of the technology. Data collected from the paper-based charts had to be double-entered into an electronic database, reconciled with patient notes and "cleaned" for evaluation of the results of the trial. Due to the long duration of these processes (18 months), the results of assessment of the CE-marked (and FDA approved) alerting system (Visensia) were not available at the time of completion of this thesis (early 2015).

2.1.1 Study population

All patients undergoing routine major upper-gastrointestinal surgery in the Oxford Radcliffe Hospitals NHS Trust were eligible for inclusion in the study. Each patient was provided with verbal information about the study, and was then required to sign a consent form before data could be stored and used. Figure 2.3 shows the steps involved in creating the cohort of patients included in the analysis. About 70% of all patients assessed for eligibility for the study consented and met the inclusion criteria for each phase of the trial. Patients who were either under the age of sixteen years; or unable to give informed consent; or who were pregnant; or whose anatomy precluded the use of the monitoring equipment were excluded from the study. Finally, eighty patients were not included in the study because either they were identified as needing palliative management during their operation; or were missed at admission; or had their operation cancelled. This left 407 patients (200 from Phase I plus 207 from Phase II) for final analysis.

The cohort described above included patients undergoing the following procedures: oesophagectomy, gastrectomy, hepatectomy, pancreatectomy, splenectomy

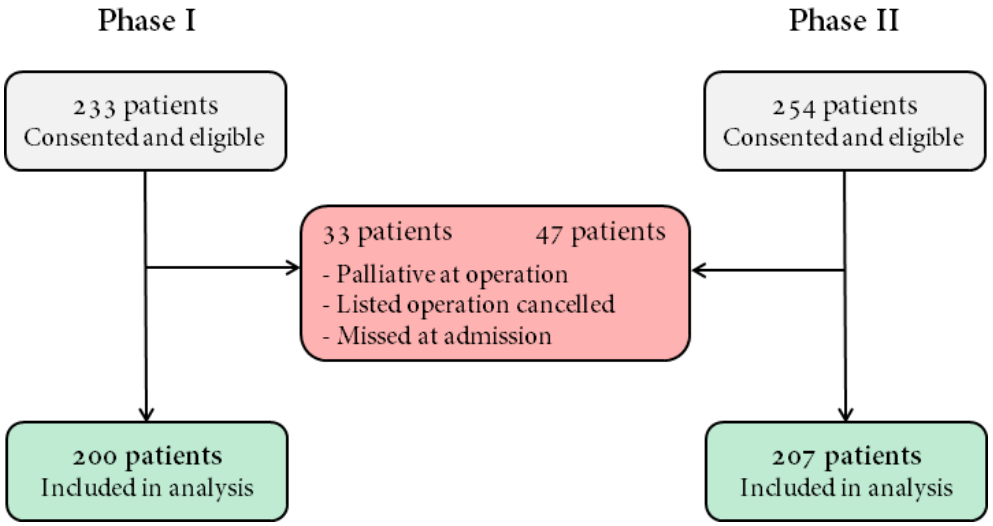


Figure 2.3: Flow chart for patient selection: 487 patients in both phases who were assessed for eligibility consented to participate in the study; 80 patients who were identified for palliative management at the time of operation, missed at admission or had the operation cancelled were excluded leaving (in total) 407 patients for analysis.

and gastric bypass. An approximately equal proportion of patients underwent the same type of procedure in both phases (Table 2.1); e.g., approximately the same proportion of patients, 17% in Phase I and 23% in Phase II, had a hepactectomy ($p = 0.150$). It may be concluded that there was no significant difference in the type of operation between the two phases of the trial.

2.1.2 Data collection

The data collection process provided four types/sources of data: (i) the electronic record system, in which demographic and clinical information from each patient was recorded; (ii) the paper-based observation charts, which contain a summary of the patient assessment conducted by clinical staff in their routine observations, including physiological variables, the corresponding track-and-trigger scores, and any other relevant information that was considered appropriate during the observation of the patient; (iii) the continuous vital-sign data acquired from the bedside and telemetry monitors that were saved to a hospital data server; and (iv) patient notes that contain annotations of clinically-relevant events (such as

the occurrence of cardiac arrest), along with documentation concerning the time of connection and disconnection of the continuous monitors.

The data reconciliation process involved the following stages. Firstly, the physiological data from each patient's observation chart were double-entered (subsequent to a patient's discharge) into an electronic database by two research nurses who were blind to the other entry. Discrepancies were resolved by a third researcher with recourse to the original charts. Secondly, patient and operation data descriptors along with clinical events were prospectively recorded from the notes using a standardised research *pro forma*. Clinical events were subsequently re-entered by a separate research nurse blind to the original data entry procedure, to ensure that all events had been captured. Discrepancies were resolved by recourse to the patient notes. The third stage of the reconciliation process was to assign the continuous vital-sign data recorded by each bedside and telemetry monitor (and stored in the hospital server) to the correct study patient. This involved two research assistants who matched the start and end of each record with each study patient. Discrepancies were also resolved by a research assistant by recourse to the patient notes.

During the transcription and reconciliation processes, data for each study patient were de-identified to conform to ethical requirements. This was achieved by replacing the patient's name and hospital ID with a 3-digit study ID, which was generated automatically and sequentially, so that the first patient consented to the study had the ID "001", the second patient "002", and so on. The study ID was also written on the paper observation charts and patient notes by the research nurses so that they could easily be recovered if required. The mapping between hospital ID and study ID was kept in a separate file that was accessible only by the research nurses and the senior clinical team. Other patient-sensitive information (such as date-of-birth) was also removed from the reconciled database.

2.2 Description of the data extracted

The variables stored in the final database include all the information available and approved by the Ethics Committee. The amount and variety of available variables constitutes a typical challenge of clinical data-mining as it will be described in the

next chapters. Information collected and stored in the final database includes: demographics, vital signs, free text notes, and outcome data.

2.2.1 Definition of outcome

The main aim of this work is to improve the identification of patient deterioration after major surgery. In order to do so, definition of the *outcome* is necessary. In general, an outcome can be a disease, sign, or symptom, but it is often chosen to be mortality for this type of study. Also, depending on the typical time-scale of the problem studied, different types of mortality can be considered to highlight better differences between surviving and non-surviving populations. In this work, however, we are interested in identifying those patients who may be at risk of significant adverse events that are potentially preventable or salvageable. We consider death on the ward, cardiac arrest, and unplanned (i.e., emergency) admission to ICU following the original operation as primary outcomes used for analysis.

Secondary outcomes have also been considered to evaluate the effectiveness of the intervention phase of the trial (Table 2.1). These include 30-day mortality, which denotes whether or not the patient died within 30 days after the operation, the length of stay in the ICU after an emergency ICU admission (for those patients who had an emergency ICU admission), the APACHE II score[1] for each emergency ICU admission, and the length of stay on the post-operative ward.

2.2.2 Physiological data

Physiological variables have been the main variables used for the monitoring of patients in the hospital because they can be acquired non-invasively and (relatively) easily, and have also been shown to be useful to identify those who are clinically deteriorating (Smith et al. [2008a]; Tarassenko et al. [2006]). The physiological variables acquired comprise heart rate, blood pressure, arterial oxygen

[1]The Acute Physiology and Chronic Health Evaluation II, or simply APACHE II (Knaus et al. [1985]), is a severity-of-disease classification system; one of several ICU scoring systems. It is applied 24 hours after admission of a patient to an ICU: an integer score from 0 to 71 is computed based on several measurements and co-morbidities; higher scores correspond to more severe disease and a higher risk of death.

Table 2.1: Characteristics of population studied (N = 407) showing demographics, outcome data and other clinical information for Phase I and Phase II groups of patients.

	N (%) Patients		
Characteristic	Phase I (N = 200)	Phase II (N = 207)	$p-$value[1]
Gender (Male)	111 (55.5)	121 (58.5)	0.547
Elective ICU	83 (41.5)	75 (36.2)	0.276
Epidural Use	168 (84.0)	156 (75.4)	0.031
Surgery type:			
Pancreatectomy	85 (42.5)	70 (33.8)	0.071
Oesophagectomy	39 (19.5)	40 (19.3)	0.964
Hepatectomy	34 (17.0)	47 (22.7)	0.150
Gastrectomy	23 (11.5)	23 (11.1)	0.901
Splenectomy	12 (6.0)	9 (4.3)	0.451
Gastric-bypass	7 (3.5)	13 (6.3)	0.195
Others	0 (0.0)	5 (2.4)	0.027
Outcome:			
Emergency ICU	20 (10.0)	26 (12.6)	0.415
Cardiac Arrest	3 (1.5)	3 (1.4)	0.966
ICU Mortality	2 (1.0)	0 (0.0)	0.149
Hosp. Mortality	6 (3.0)	1 (0.5)	0.051
30-day Mortality	7 (3.5)	3 (1.4)	0.182
	Median (IQR)		
Age, yr	63 (54-70)	63 (51-69)	0.270
Elective ICU, hours	24.0 (20.1-41.3)	22.0 (20.0-28.0)	0.159
ASA grade	2 (2-2)	2 (2-2)	0.925
Outcome:			
APACHE II	13 (12-17)	13 (12-19)	0.883
Emergency ICU, hours	83.6 (45.5-227.4)	79.9 (55.7-139.0)	0.797
Post-operative ward, days	8.90 (6.71-14.01)	8.97 (6.00-13.84)	0.516

[1]For the first part of the table, the p-value shows the result of a Chi-squared test for equal proportion (binomial variable); for the second part of the table, the p-value shows the result of the non-parametric Mann-Whitney U test with the null hypothesis that two groups have similar values.

saturation, temperature, and respiratory rate. Vital-sign data were collected from patients using standard monitors: the Philips *M3046A/Intellivue MP50*, the Nonin *4100 Bluetooth Enabled* Digital Pulse Oximeter, the GE *DINAMAP Pro 400V2*, the Welch Allyn *Spot Vital Signs 4200b-E4*, and the Covidien *Genius 2* Thermometer to measure temperature.

Body Temperature. The body's temperature represents the balance between heat produced and heat lost, otherwise known as thermoregulation. In the clinical environment, body temperature may be affected by factors such as underlying pathophysiology, skin exposure, or age. Other factors may not affect the body's core temperature but can contribute to inaccurate measurements, such as the consumption of hot or cold fluids before oral temperature measurement. In clinical practice, core body temperature is measured using a tympanic, rectal, or oral thermometer. It is not possible to obtain continuous measurements of core body temperature with these techniques, and the continuous measurement of body temperature using thermistors has not proven to be sufficiently accurate, consistent and reliable enough to be used in the CALMS-2 trial (as reported by Wong [2011]). Therefore, temperature measurements were performed using a tympanic thermometer during the routine observations of patients by the nursing staff.

Blood pressure. BP refers to the pressure exerted by blood against the arterial wall. It is influenced by cardiac output, peripheral vascular resistance, blood volume, and the viscosity and the elasticity of the vessel wall (Elliott and Coventry [2012]). BP is an important vital sign to measure because it provides an indication of blood flow when the heart is contracting (systole) and relaxing (diastole). It is also one of many indicators of cellular oxygen delivery. Sudden changes or long-term trends in BP may reflect underlying pathophysiology or the body's attempts to maintain homeostasis. A drop in BP, for example, has been found to be a common sign in patients before cardiac arrest (Rich [1999]). Non-invasive automated blood pressure monitors (or manual sphygmomanometers) are typically used in clinical settings to measure BP. They use an inflatable cuff placed around the upper arm: the cuff is inflated until the blood flow through the brachial artery is completely occluded; the air in the cuff is then released and the oscillations in cuff pressure caused by the in-rushing blood are measured. The mean arterial

pressure, defined as being the average arterial pressure during a single cardiac cycle, is then determined from the pressure signal when the oscillations are of maximum amplitude, and the systolic and diastolic BP values derived using the oscillometric method (Geddes et al. [1982]).

Peripheral oxygen saturation. Oxygen saturation of the blood is defined as the ratio of oxyhaemoglobin, the molecule from the combination of the haemoglobin molecule and oxygen, to the total haemoglobin present in the blood. SpO_2 is an estimate of the arterial oxygen saturation level and is usually measured non-invasively using pulse oximetry. A pulse oximeter is a sensor that contains two light sources at different wavelengths (usually red, 600 nm, and infrared, 940 nm), and a photodetector, which are placed on opposite sides of the translucent part of the patient's body (such as a fingertip or earlobe). The absorption (or transmission) of light at these wavelengths varies for oxygenated and de-oxygenated blood due to the different absorption coefficients for oxyhaemoglobin and reduced (non-oxygenated) haemoglobin. SpO_2 can therefore be determined by measuring the light reflected (or transmitted) for each of the two wavelengths (Wukitsch et al. [1988]). If the patient is breathing room air, the SpO_2 is typically considered "normal" if it is between 95% and 100%. If the patient receives supplemental oxygen (such as oxygen therapy through the use of an oxygen mask or a cannula), SpO_2 routinely reaches 100%. SpO_2 and information about the use of supplemental oxygen were recorded during the trial.

Heart rate. HR is the number of contractions of the heart in one minute, measured in beats per minute (bpm). It is typically measured by using an electrocardiogram (ECG) or derived from photoplethysmography using a pulse oximeter. HR can also be measured manually by measuring the pulse, which is defined as being the palpable rhythmic expansion of an artery by the increased volume of blood pushed into the vessel by the contraction and relaxation of the heart. It is affected by many factors including age, existing medical conditions, and medications (Elliott and Coventry [2012]).

Respiratory rate. RR, also called breathing rate or respiration rate, is the number of breaths in one minute, measured in breaths (or respirations) per minute (rpm), and has been shown to be one of the most sensitive indicators of critical

illness (Smith et al. [2008a]). It may be measured automatically using electrical impedance pneumography (Folke et al. [2003]), which measures the electrical impedance between two ECG electrodes at a frequency typically between 10 kHz and 100 kHz. The impedance increases as the patient inhales due to (mainly) the increased resistivity of the air-filled lungs but also because of changes in the volume of the chest cavity. However, due to the unreliability of this method (Elliott and Coventry [2012]), the current recommended method for measuring RR is to count the number of breaths (by looking at the chest wall movement) in 60 s using a timer. Occasionally, clinical staff do not count RR for a whole minute, but instead count for only a fraction of 60 s (typically for 30 or 15 s) and then scale up the number of breaths to 60 s. This decreases the accuracy of the measurement by amplifying the counting error and enforcing a quantisation effect (e.g., if the number of breaths counted in 15 s is scaled up to be a count in breaths per minute, then the result is a multiple of 4, such that RR is recoded as 12 or 16 rpm for most patients). In the CALMS-2 trial, the manual method described above was used in the routine observations performed by nursing staff.

A different subset of these physiological variables was collected at different stages of the patient's recovery on the ward (Figure 2.4). Routine observations of all vital signs were performed (approximately every four hours) by the clinical staff during the entire stay of all patients on the ward (as it is part of normal ward care). In addition to track-and-trigger observations, data from the bedside and telemetry monitors were also recorded, providing a limited set of continuous physiological data. Bedside monitors were connected to patients typically during the first day of the patient on the ward (after surgery), while the telemetry (portable) monitoring devices were connected to patients during the remainder of their stay on the ward. With the former, a subset of vital signs comprising HR, RR, SpO_2, and BP, was measured "continuously" at the sampling rates described in Figure 2.4. With the latter, a smaller subset of vital signs (only HR and SpO_2) was recorded at the sampling rate of 1 Hz, together with the PPG waveform.

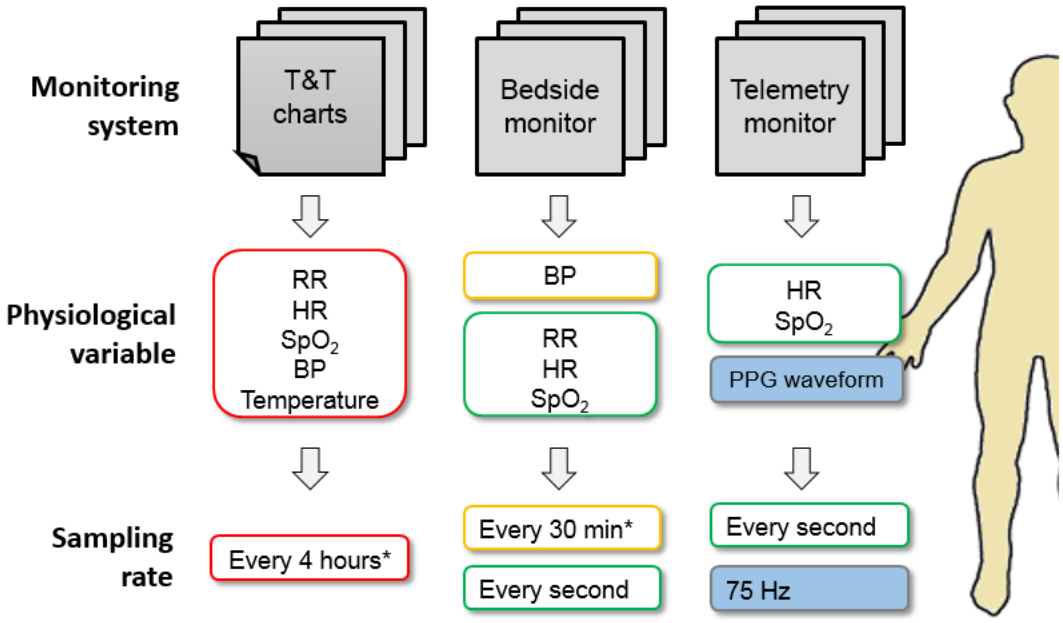


Figure 2.4: Monitoring systems used in the trial, the physiological variables acquired with each system, and the correspondent sampling rate of each data stream. *These are periodic observations performed by ward staff, and may vary according to the patient's condition.

2.2.3 Neurological status

Different metrics have been developed to determine the neurological status or level of consciousness of hospitalised patients. The GCS is a clinical score from 3 to 15 whereby higher values indicate a healthier neurological status. The GCS decomposes into: best verbal, motor, and eye opening responses scoring between 1 and 4, 1 and 6, and 1 and 5 (Teasdale and Jennett [1974]), respectively. The Alert-Verbal-Pain-Unresponsive (AVPU) scale is an alternative metric that uses a 4-point scale to measure the level of consciousness: alert, responds to verbal stimuli, responds to pain, and unresponsive. For this study, neurological status was assessed using the AVPU scale in all routine observations performed by the clinical staff.

2.2.4 Demographic information

Demographic data comprise all high-level information about a patient that is considered as constant for a hospital admission (Table 2.1): age, gender, ethnicity. Additional administrative information was also collected such as admission type (e.g., whether the patient was electively or urgently admitted to the hospital) and source (such as whether the patient was admitted from the emergency department or transferred from another hospital or health institution). In this study, all patients were electively selected for upper-gastrointestinal surgery and were subsequently admitted to the post-operative ward.

2.2.5 Other clinical data

Other clinically-relevant information recorded include the American Society of Anesthesiologists (ASA) grade (Dexter et al. [2002]), which is a physical status classification system for patients undergoing surgical procedures, and the use of patient controlled epidural analgesia for the management of pain (Table 2.1). Post-operative pain is a potential trigger for the body's stress response, it activates the autonomic system, and is thought to be an indirect cause of adverse effects on various organ systems (Liu et al. [1995]). The most common side-effect associated with having an epidural is low blood pressure. The standard procedure of the post-operative ward in our study is to remove the epidural catheter on the third day after surgery. For patients who did not have epidural administration of opioids, patient-controlled analgesia with intravenous opioids was used instead.

2.2.6 Continuous wearable monitoring

Different metrics have been used to determine the effectiveness of introducing our ambulatory continuous monitoring system on the post-operative ward.

Figure 2.5 shows the percentage of the total monitoring time for each patient (defined to be the time for which the wearable sensor was attached to the patient) for which actual data were acquired. For the purposes of this analysis, a data gap was defined to be a period of over 60 minutes without any acquired data, and patients with a length of stay on the ward longer than 60 days were not included

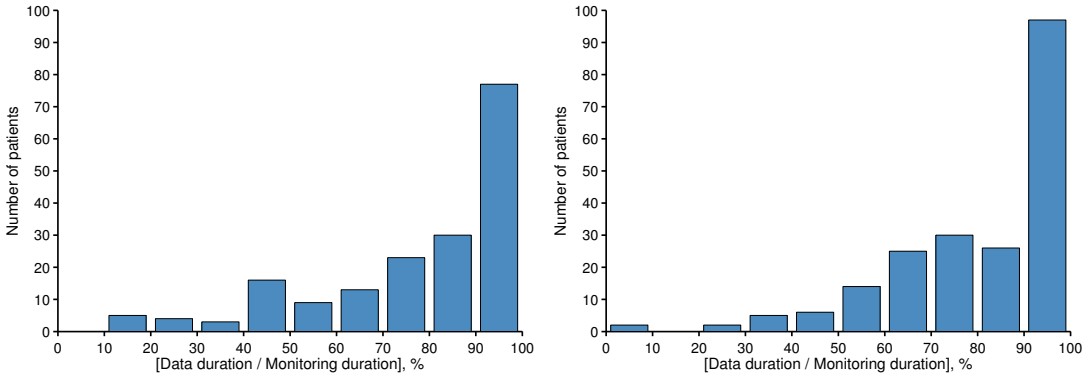


Figure 2.5: Histogram of continuous data completeness for patients in Phase I (left) and Phase II (right) as a percentage of the total time that the patient was equipped with a wearable patient monitor.

(5 in Phase I, and 4 in Phase II). It may be seen that for the majority of patients in both phases, the percentage of actual data acquired during the periods of time for which the patient was equipped with the monitor is close to 100% (Table 2.2). We may see a slight improvement in data acquisition from Phase I to Phase II; that is, more data were acquired for more patients. The major causes of data incompleteness were infrequent malfunction of the wearable sensors and PDAs, failures in the hospital wi-fi network, occasional crashes of the central server, and loss of battery power in the wearable sensors and PDAs. A team of research nurses was responsible for ensuring that patient compliance and device readiness were kept as high as possible during both phases.

The total monitoring time for patients in each phase of the trial using both bedside and telemetry monitoring systems is shown in Figure 2.6, where the length of stay on the ward and the actual duration of acquired data are also shown. The comparison with Figure 2.5 shows that patients were typically connected to the wearable patient monitors for a small proportion of their stay on the ward, with a median monitoring time of approximately 2 days in Phase I, and 4 days in Phase II (compared with a median length of stay of approximately 9 days in both phases). The total time of actual acquired data (with both bedside and telemetry systems) was approximately 375 days in Phase I and 784 days in Phase II (the total time spent on the ward was 3152 days in Phase I and 2536 days in Phase

Table 2.2: Continuous monitoring data completeness for Phase I and Phase II groups of patients.

	Median (IQR)	
	Phase I	Phase II
Bedside monitor:		
Monitoring time, days	0.17 (0.00-0.99)	1.20 (0.56-1.93)
Data time, days	0.05 (0.00-0.78)	0.78 (0.45-1.53)
Data time, %	86.4 (64.3-96.5)	82.4 (63.1-95.3)
Telemetry monitor:		
Monitoring time, days	1.32 (0.18-2.70)	2.47 (0.99-4.30)
Data time, days	1.03 (0.17-2.00)	2.28 (0.81-3.79)
Data time, %	100 (72.6-100)	100 (84.3-100)
Total data time, days	375.08	784.03
Average data time per patient, days	1.88	3.79

II).

Much of the difference between total stay on the ward and total monitoring time is due to patients' compliance. ECG sensors were particularly unpopular with patients, despite their small size, probably due to their positioning on the chest following upper-gastrointestinal surgery (as also discussed in Clifton et al. [2014]). Due to the perceived discomfort of the ECG sensors, they were discontinued from use after 52 patients had been continuously monitored in Phase I of the trial[1]. The pulse oximeters attached to the fingertip were tolerated much better by this cohort of patients. However, patients typically removed the pulse oximeters before eating or washing and often failed to replace the device afterwards. This was particularly evident during weekends, when research nurses were unavailable to check the connectivity of each patient.

There was no statistically-significant difference between the groups of patients from both phases of the trial regarding their demographic and clinical information (Table 2.1), as observed earlier. More importantly, no statistically-significant differences were observed in the primary and secondary outcomes considered. In

[1]For this reason, ECG waveforms from the small subset of patients were not taken into account in the analysis presented in this thesis.

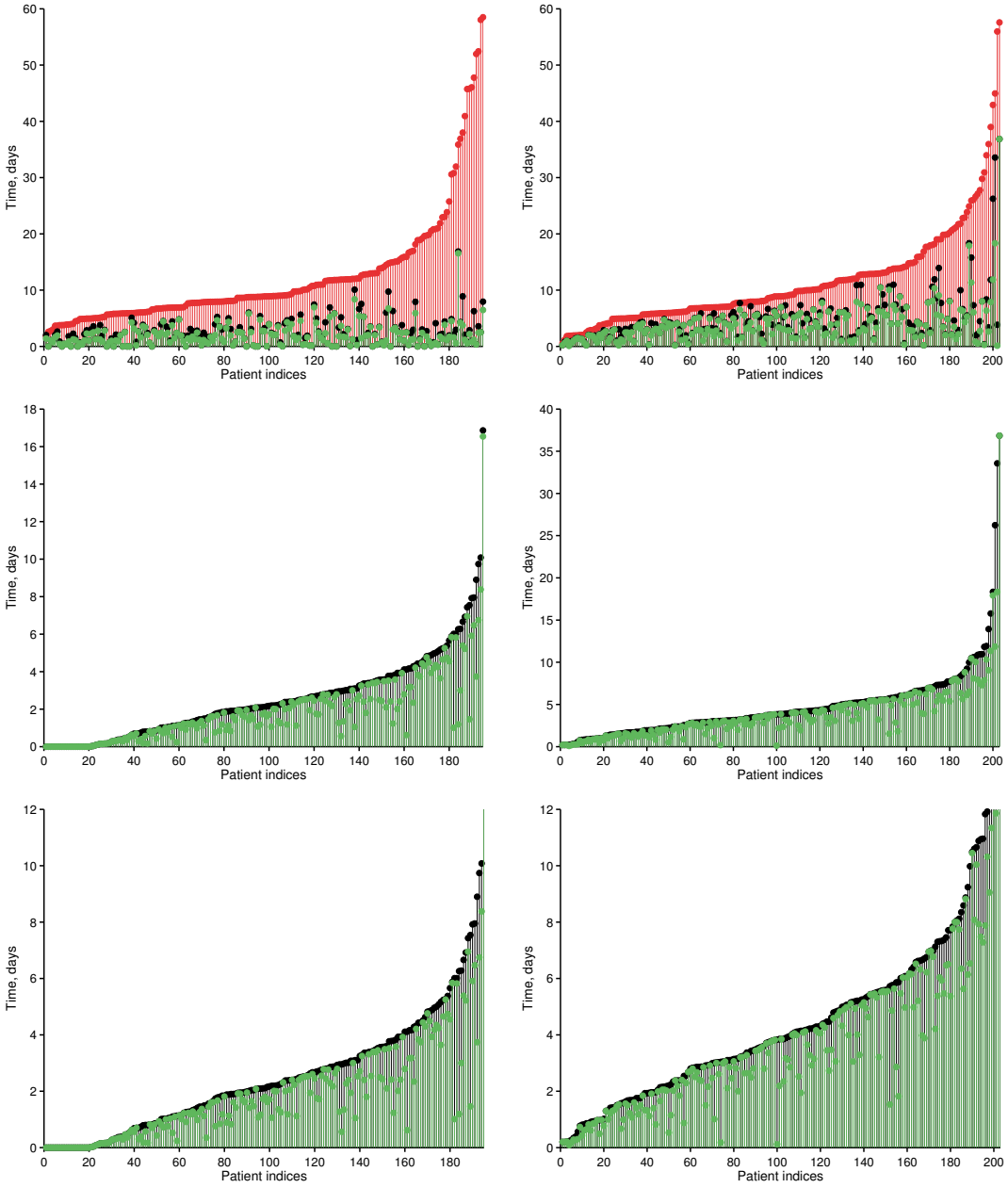


Figure 2.6: Representation of continuous monitoring time for patients in Phase I (left) and Phase II (right). Top row shows the length of stay of the patient on the ward (sorted in ascending order and shown in red) with the time that patients were equipped with the continuous monitoring systems (black) and the actual time of acquired data (green). Middle row shows the time that patients were equipped with the continuous monitoring systems (sorted in ascending order and shown in black) with the actual time of acquired data (green). Note the different y-axis scales used. Bottom row highlights the plots from the middle row for a maximum number of 12 days (same y-scale is used).

Phase I, 20 patients had at least one emergency admission to the ICU, in which three of them were preceded by a cardiac arrest. In Phase II, 26 patients had at least one emergency admission to the ICU, and three of them were preceded by a cardiac arrest. Hence, more patients in Phase II were admitted to the ICU, and their length of stay in the ICU was 80 hours, compared with 84 hours for those in Phase I ($p = 0.797$, which is not statistically-significant). Six patients died during their recovery in the hospital in Phase I of the trial, compared to only one in-hospital death in Phase II. The combination of these results suggests that the wearable monitoring systems had a moderate impact on the clinical outcomes of post-operative patients: in general, patients who deteriorated were admitted earlier to the ICU in Phase II.

The continuous, observational, and clinical data acquired for all 407 patients was approximately 22 GB. These data are used in this thesis for investigating methods for improved predictive monitoring and early identification of deterioration.

2.3 Study populations

In the work described by this thesis, clinical data obtained from post-operative patients who participated in the CALMS-2 trial were used to test the proposed methods. In order to design, train, and evaluate the algorithms proposed, a second, distinct, and independent vital signs database was used, the Portsmouth database. The study populations for each dataset are described in the following sections.

2.3.1 The CALMS-2 dataset

The CALMS-2 dataset includes clinical data from patients who were admitted to the post-operative ward of the Oxford Cancer Hospital, as described above.

The aim of this thesis is to identify patients who are at risk of significant adverse events that are potentially preventable or salvageable. Therefore, as described previously, we consider a composite outcome comprising death on the ward, cardiac arrest and unanticipated admission to ICU following the original

Table 2.3: Characteristics of the CALMS-2 population (N = 407) showing demographics and other clinical information for "normal" and "abnormal" groups of patients.

	N (%) Patients		
	Normal (N = 357)	Abnormal (N = 50)	$p-$value
Gender (Male)	206 (57.7)	26 (52.0)	0.446
Elective ICU	124 (34.7)	34 (68.0)	<0.001
Epidural Use	281 (78.7)	43 (86.0)	0.231
Surgery type:			
Pancreatectomy	145 (40.6)	10 (20.0)	0.005
Oesophagectomy	55 (15.4)	24 (48.0)	<0.001
Hepatectomy	75 (21.0)	6 (12.0)	0.135
Gastrectomy	39 (10.9)	7 (14.0)	0.520
Splenectomy	18 (5.0)	3 (6.0)	0.774
Gastric-bypass	20 (5.6)	0 (0.0)	0.086
Others	5 (1.4)	0 (0.0)	0.400
	Median (IQR)		
Age, years	63 (53-69)	63 (51-72)	0.480
Elective ICU, hours	22.6 (20.0-32.2)	24.5 (20.2-46.7)	0.146
ASA grade	2 (2-2)	2 (2-3)	0.067

[1]For the first part of the table, the p-value shows the result of a Chi-squared test for equal proportion (binomial variable); for the second part of the table, the p-value shows the result of the non-parametric Mann-Whitney U test with the null hypothesis that two groups have similar values against the hypothesis that one population has larger values.

operation. This composite outcome is described as a *major adverse event*. The group of patients who experienced a major adverse event was separated from the group who did not for the purpose of analysis.

Table 2.3 details the main characteristics of the two cohorts of patients considered in the CALMS-2 dataset (N = 407). The "normal" group contains 357 patients (88%) who did not suffer any major adverse event during the course of their stay on the post-operative ward; i.e., patients in this group had a "normal" recovery after surgery. The "abnormal" group comprise patients who had at least one major adverse event during their in-hospital recovery period from

surgery. There is no statistically significant differences between the two groups with respect to demographic information. We observe, however, that the majority of the patients who suffered a major adverse event were electively admitted to the ICU just after surgery, which, in proportion, is statistically different from that of the "normal" group. This is directly correlated with the fact there is also a greater proportion of patients in the "abnormal" group who underwent a oesophagectomy, which is considered to be a higher risk surgical procedure, leading to (elective) admission to ICU for the first 24 hours after surgery. Nevertheless, this appears to not have a substantial impact on the time spent in the ICU: patients from the "normal" group who were electively admitted to the ICU stay approximately 23 hours (median), while the same figure for the "abnormal" group of patients is close to 25 hours.

As will be described in the next chapter, current scoring systems include the measurement of physiological variables, demographic information, and other variables from laboratory tests that are not measured for all patients nor are collected as frequently as vital-sign data. It will be important, however, to consider factors such as surgery type when evaluating the results from our analyses.

2.3.2 The Portsmouth dataset

The second dataset used in this thesis was collected from clinical data obtained from consecutive admissions to beds in the Medical Assessment Unit (MAU) of the Portsmouth Hospital (Portsmouth Hospitals NHS Trust) in the UK between 8 May 2006 and 30 June 2008. It has been described in other publications (Prytherch et al. [2010]; Smith [2013]; Smith et al. [2008a]). The MAU is the common entry point for all general medical emergency patients aged $\geq$ 16 years, with the exception of those transferred directly to critical care areas of the hospital.

As routine part of clinical care, MAU staff entered each patient's vital signs data (during routine observations) into PDAs running the *VitalPAC* software (The Learning Clinic, London, UK, described in Smith et al. [2006b]). The PDAs were linked by a wireless local area network to the hospital's intranet, where physiological data were integrated with patient demographic information (e.g., age and gender). The outcome at hospital discharge (alive/dead) was obtained

Table 2.4: Characteristics of the Portsmouth population (N = 34,060 patient-episodes) showing demographic information for "normal" and "abnormal" groups of patients.

	N (%) Patient-Episodes		
	Normal (N = 31,018)	Abnormal (N = 3,042)	$p-$value
Gender (Male)	14,791 (47.7)	1,435 (47.2)	0.589
	Median (IQR)		
Age, years	72 (55-82)	83 (75-88)	< 0.001

[1]For the first part of the table, the p-value shows the result of a Chi-squared test for equal proportion (binomial variable); for the second part of the table, the p-value shows the result of the non-parametric Mann-Whitney U test with the null hypothesis that two groups have similar values against the hypothesis that one population has larger values.

for all patients from the hospital's patient administration system and added to the database for analysis. We note that in this database, the only outcome measure available is discharge status. Data from patients who were discharged from hospital before midnight on the day of admission were excluded from the database.

As with the previous dataset, the non-survivors at discharge (major adverse event) were separated from the group who was alive for the purpose of analysis. Table 2.4 shows the demographic information for the two cohorts of patients. There were 34,060 patient-episodes, that is, 34,060 admissions to the MAU corresponding to 24,212 patients: 18,438 (76.2%) patients had one admission; 3,783 (15.6%) had two admissions; 1,120 (4.6%) had three admissions; and 871 (3.6%) had four or more admissions to the MAU during the duration of the study. The median (IQR) length-of-stay in the hospital was 5 (2 - 12) days[1]. We observe that in this dataset there is a statistically-significant difference between the age of patients in the two cohorts: non-survivors are older than those who were alive at discharge. No statistically-significant difference between the two cohorts were observed with respect to gender.

Finally, we note that there are significant differences between the patient

[1]Vital signs observation sets were not available when the patient was transferred out of the MAU. The median (IQR) length-of-stay in the MAU (determined as the time between the first and last observation sets in each patient-episode) was 0.7 (0.4-1.1) days.

populations included in each of the databases (the CALMS-2 database and the Portsmouth database), which are important to take into account when developing and testing scoring systems. This problem is related to that of *dataset shift* (and *domain shift*), which is commonly found in the machine learning literature. Typically, the conditions (or environments) in which we use the systems we develop (test of the system) will differ from the conditions in which they were developed (training of the system). Dataset shift and domain shift occur when the joint distribution of inputs and outputs differs between training and test phases, which is present in most practical applications, for reasons ranging from the bias introduced by experimental design to the irreproducibility of the testing conditions at training time (Quionero-Candela et al. [2009] provide an overview of current efforts to deal with dataset and domain shift). More specifically, patients included in the CALMS-2 database are, in general, younger than those included in the Portsmouth database. Also, the proportion of female patients is lower in the CALMS-2 database. More importantly, the hospital settings from which data were obtained are substantially different. In the CALMS-2 database, patients were admitted to the hospital for elective surgery, and were subsequently admitted to the post-operative ward after their operation; while the Portsmouth database was developed from clinical data obtained from patients admitted to a MAU, which represents a much broader range of demographics information. These and other factors that will be highlighted throughout this thesis, will be important factors to consider when evaluating the results from our analyses using these two independent databases in the following chapters.

Chapter 3

Early-warning scores for post-operative patients

We have briefly introduced the role of track-and-trigger systems and their early-warning scores (EWSs) for in-hospital patients (in section 1.3). We then presented a description of the data acquired and extracted for creating the two databases with the clinical information required to compute these scores, and which are used for the analysis described in this thesis. In this chapter, we will investigate various EWS systems that might be applied to both datasets and set a baseline to which some of our results will be compared. Firstly, we will briefly describe various scoring systems in clinical use, and their performance in post-operative patients as reported in the literature. Then, some of these scoring systems will be further explained, implemented, and applied to the observational data.

3.1 Design & evaluation of scoring systems

The majority of EWS systems to date have been developed empirically. The various thresholds of existing EWS systems mostly rely on heuristic methods and best-guess physiological variables ranges, and often lack clinical validation. The evaluation of such systems, when reported, usually relies on metrics such as the area under the receiver-operating characteristic curve (AUROC). The details of these metrics are introduced in this section to help the understanding of previous

work.

3.1.1 Evaluation of EWS system performance

Performance is typically evaluated by relating a clinical outcome to the EWS value calculated for a given observation set[1]. This can be achieved by recognising that the goal of the EWS system is twofold: (1) it must alert clinical staff to the need for clinical intervention; that is, it should correctly identify all events when patients are in a condition of physiological distress (such events in this context will be defined as the *positive* class); and (2) it must correctly recognise when a patient's condition does not give rise to concern (defined to be the *negative* class).

3.1.1.1 Discriminative metrics

The aspect of an EWS system's performance that is most commonly evaluated in the literature is the ability to discriminate between positive and negative outcomes when presented with observational data that precede the outcome by some interval (e.g., 24 hours). A threshold on the EWS has to be selected; if an observation set is assigned an EWS that is above or below the threshold, then the classification is as belonging to the positive or negative class, respectively. These binary classifications then have to be compared with the actual (true) values.

At every candidate value of the threshold on the EWS, it is possible to allocate each observation set to either the positive or negative class, as described above. Each classification may thus be defined as being true positive (TP), false positive (FP), false negative (FN), or true negative (TN) as detailed in Table 3.1. An observation set correctly classified as being positive or negative is TP or TN, respectively. Likewise, incorrect classification as being positive or negative is FP or FN, respectively. Additional metrics may be derived from these numbers (Table 3.1):

- the positive predictive value (PPV) and negative predictive value (NPV), which correspond to the proportions of correctly identified cases within all cases classified positive and negative, respectively;

[1]Each vital-sign observation set contains: date/time of the observation set and the measurements of the vital signs.

Table 3.1: Confusion matrix showing the relation between classified and actual values (deterioration or no deterioration) according to positive, $\oplus$, and negative, $\ominus$, classes: True Positives (TP), False Positives (FP), False negatives (FN), and True Negatives (TN). Additional metrics of discriminatory power are shown on the right and bottom of the table as combinations of these.

		Actual value $\oplus$	Actual value $\ominus$	
Classified value	$\oplus$	TP	FP	Positive Predictive Value $\frac{TP}{TP+FP}$
	$\ominus$	FN	TN	Negative Predictive Value $\frac{TN}{FN+TN}$
		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

- the sensitivity (*Sens*) and specificity (*Spec*) values, which are defined as the proportion of correctly identified cases within all actual positive and negative cases, respectively;
- the accuracy, which is defined as the proportion of correctly identified cases within all cases.

Threshold metrics. Several threshold (also called discrete) metrics can be used to evaluate the discriminative performance of a given EWS system (Jeni et al. [2013]). The accuracy, for example, is a widely used metric for measuring the performance of a classifier; however, when the numbers of examples in each class are substantially different (i.e., when the dataset is highly imbalanced), this is of limited usefulness as an index of discriminatory performance (as discussed Metz [1978]). In screening for a relatively rare event (such as deterioration), one can be very accurate simply by ignoring all evidence and calling all cases negative. If only 5% of patients have events, a system that always assigns the normal class will be right 95% of the time. This limitation is overcome by defining decision performance in terms of the pair of sensitivity and specificity values (Fawcett [2004, 2006]; Metz [1978]).

An alternative choice for a single performance measure is the F_1-score, which can be interpreted as a weighted average of the *precision* (which corresponds to the positive predictive value) and *recall* (which corresponds to the sensitivity or true positive rate) values:

$$F_1\text{-score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3.1}$$

We note, however, that the F_1-score can also be affected by imbalanced datasets (as discussed in Jeni et al. [2013]).

The Matthews correlation coefficient (MCC), introduced by Matthews [1975], is also a single performance measure that is less influenced by imbalanced datasets as it considers simultaneously accuracies and error rates on both classes (Bekkar et al. [2013]; Fawcett [2006]), and incorporates all values of the confusion matrix (Table 3.1):

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.2}$$

MCC ranges from 1 for a perfect classification to -1 for the worst possible classification. A value close to 0 indicates a model that performs randomly.

Rank metrics. When assessing the overall discriminatory performance of a scoring system, many authors use rank metrics; i.e., metrics that do not restrict the analysis to a specific threshold but rather consider all the different thresholds allowed by the system (Fawcett [2006]; Hanley and McNeil [1982]). Rank metrics, therefore, depend on the ordering of the cases, and not on the actual classification. This is typically performed using receiver-operating characteristic (ROC) curve analysis (Metz [1978]). For this, sensitivity and specificity are defined as a function of the threshold (or operating point, o), and the ROC curve is defined as the plot $Sens(o)$ *vs.* $1 - Spec(o)$ for all possible values of o; i.e., the ROC curve indicates all possible combinations of $(Sens, 1 - Spec)$ for a given system[1].

[1]Note that $1 - Spec$ corresponds to the false positive rate.

Integrating the ROC curve gives the AUROC, which indicates the discriminatory power a scoring system can provide over all possible thresholds. The line defined by $Sens(o) = Spec(o)$, which is the diagonal joining the bottom-left corner of this plot to the top-right corner, with an AUROC of 0.5, represents the discriminative power of random guessing. The "optimal" operating point can be found by minimising the loss function g defined by

$$g(o) = v(1 - Sens(o)) + (1 - v)Spec(o) \tag{3.3}$$

where v corresponds to the cost of a false negative. If the relative importance of false positives and false negatives is equal, then $v = 0.5$. This is equivalent to finding the closest point on the ROC curve to the point $(0, 1)$.

An alternative to the area under the ROC curve is the area under the precision-recall curve, which shows the precision as a function of the recall. Nevertheless, as observed by Fawcett [2004] and Jeni et al. [2013], precision-recall curves are dramatically affected by class imbalance.

Prytherch et al. [2010] introduced the "EWS efficiency curve" to evaluate the diagnostic performance of different scoring systems (without, however, introducing a quantitative metric). For each EWS value, the percentage of the total number of observations at, or above, that EWS value is shown against the percentage of the total number of observations for which the outcome was true at, or above, the EWS value. This generates a plot of the cumulative proportion of observations against the cumulative proportion of true outcomes, similar to the Lorenz curve typically used in economics (Lorenz [1905]).

3.1.1.2 Other metrics

To date, all studies published in the literature report the performance of EWS systems using mainly the AUROC metric discussed above. Nevertheless, given the goal of EWS systems, it is important to consider other metrics to evaluate their diagnostic performance. One of the major practical drawbacks of the AUROC as an index of diagnostic performance is that it summarises the entire ROC curve, including regions that frequently are not relevant to practical applications (e.g.,

regions with low levels of specificity). To alleviate this deficiency while benefiting from some of the advantageous properties of the AUROC, one can use a partial area under the ROC curve (pAUROC). When using the pAUROC, originally proposed by McClish [1989], one considers only those regions of the ROC space which correspond to clinically relevant values of test sensitivity or specificity. This therefore summarises the section of the curve over a more meaningful pre-specified range of interest, which may bring clear advantages when evaluating the performance of different diagnostic tests for practical applications (as discussed in Ma et al. [2013]). Other metrics should include how long before an event (within a certain period of time) the system is able to identify deterioration, so that timely attendance at the patient bedside may occur when the EWS system is in real-time use. For these reasons, we define the following two metrics:

Partial AUROC. We define pAUROC over the range of false positive rates $[0, e]$ to be an integral of the ROC function over the given range: $A_e = \int_0^e ROC(f)df$. When $e = 1$, the partial area represents the conventional AUROC. Because the integrand is $\ll 1$, the summary index $A_{0.4}$ cannot attain the maximum value of 1 that is achievable by the AUROC. In order to regain the desirable property that the summary measure varies between 0 and 1, we consider instead the *scaled* partial AUROC. For a ROC curve describing better-than-chance performance, A_e can be shown to vary from ${}^{e^2}/_2$ to e; hence, a natural transformation to the partial AUROC aimed to "standardise" the range of its values is as follows (as suggested by Ma et al. [2013]):

$$\tilde{A}_e \approx \frac{1}{2}\left(1 + \frac{A_e - {}^{e^2}/_2}{e - {}^{e^2}/_2}\right) \tag{3.4}$$

For ROC curves describing better-than-chance performance, $\tilde{A}_e$ varies from 0.5 to 1. In this work, we set $e = 0.4$; i.e., we focus on a range of specificities that range from 0.6 to 1, as lower values of specificity may not be tolerable in clinical settings.

Time-to-event. This is defined as the difference between the time of the observed outcome and the time of the first observation set before the outcome event

(within a given period of time) that generates an alert at a given operating point. That is, this metric evaluates how long before an adverse event the scoring system can identify deterioration (a similar metric has been used by Fawcett and Provost [1999] to evaluate the timeliness of alarms for fraud detection).

We note that the various discriminative metrics described above can only be understood with respect to the event markers used for the test, that is, the criterion for a "positive" case. In a clinical setting, collecting event markers may be difficult when the aim is to have a diagnosis as broad and diffuse as "deterioration". For this reason, many researchers use a clearly defined marker (the outcome) such as death, cardiac arrest or unanticipated admission to ICU, and consider all the observation sets performed within a certain period of time before the outcome as belonging to the "positive" class. That is, all the observations made by nursing staff within, for example, 24 hours before an observed adverse event might be assumed to indicate whether or not the patient is deteriorating (the "diagnostic test"). One drawback of this approach is that different choices of outcome events may lead to different values of sensitivity and specificity. Furthermore, the different outcome measures mean that it is often not possible to compare sensitivities and specificities obtained in different studies.

3.1.1.3 Performance estimation using out-of-sample data

The *generalisation* performance of a scoring system (or any learning classification method or model) relates to its classification capability on independent test data. Assessment of this performance is extremely important in practice, since it guides the choice of the method, model, or scoring system, and gives us a measure of the quality of the model ultimately chosen.

The gold standard for validating a scoring system is to collect new data prospectively and to compute the aforementioned metrics of performance on the predictions made with these data. Unfortunately, prospective data collection in practice is not always possible. Therefore, pseudo-prospective approaches are used instead, which consist of mimicking this behaviour in a retrospective setting: the most recent observations are kept aside during design for model validation. However, due to changes in medical practice and the use of different technologies

through time, this may not be suitable. Probably the simplest and most widely used method for estimating a system's performance is *cross-validation*. To do so, the entire dataset is usually split into independent and complementary datasets with the following goals:

Model selection/design: estimating the performance of different models (or different parameterisations of the same model) in order to choose the most appropriate for the data; and

Model assessment: having chosen a final model, estimating its performance on new data.

If we are in a data-rich situation, the best approach for both problems is to randomly divide the dataset into two parts: a training/validation set, and a test set. The training/validation set is used to build the models and to estimate classification errors (or performance) for model selection. The test set is used for assessment of the general performance of the chosen model. We note that the number of observation sets in each partition may greatly influence the result. Performance will also vary depending on what training and validation sets are drawn from the data. For instance, missing values in the training data may impair parameter estimation. Similarly, different data distributions may affect results. Furthermore, data are often scarce, which may make it impossible to set aside an independent test set.

For selecting the parameters of the model (during the model selection/design), there are various cross-validation techniques that may be used. K-fold cross-validation uses part of the available training/validation data to fit the model, and a different part to evaluate it. Data are split into K independent, complementary groups and roughly equal-sized groups. The model fitting is repeated K times, each time leaving one k^{th} of the data aside for model evaluation. At the end of the procedure, each observation set has been used an equal number of times for training ($K-1$ times) and for evaluation (once). This guarantees that all observations have equally contributed to the computation of K performance metrics. The case $K = N$ is known as *leave-one-out* cross-validation: data are partitioned so that a single observation is left out for validation ([Hastie et al., 2009, Chapter 7]).

Table 3.2: Examples of early warning score systems.

VitalPAC Early Warning Score (ViEWS) - Prytherch et al. [2010]
A score of 5 or more triggers a review of the patient

	Score						
Variable	3	2	1	0	1	2	3
Heart Rate		≤ 40	$41-50$	$51-90$	$91-110$	$111-130$	≥ 131
Resp. Rate	≤ 8		$9-11$	$12-20$		$21-24$	≥ 25
Temperature	≤ 35.0		$35.1-36.0$	$36.1-38.0$	$38.1-39.0$	≥ 39.1	
Systolic BP	≤ 90	$81-100$	$101-110$	$111-249$	≥ 250		
SpO_2	≤ 91	$92-93$	$94-95$	≥ 96			
Inspired O_2				Air			Any O_2
AVPU scale				A			V, P, U

Assessment Score for Sick patient and Step-up in Treatment (ASSIST) - Subbe et al. [2007]
A score of 4 or more triggers a review of the patient

	Score							
Variable	4	3	2	1	0	1	2	3	4
Heart Rate	≤ 49			$50-60$	$61-100$	$101-120$	$121-140$		≥ 141
Resp. Rate	≤ 9				$10-25$	$26-30$	$31-35$		≥ 36
Systolic BP	≤ 84		$85-90$	$91-99$	$100-220$				≥ 221
AVPU scale					A	V	P		U
Age					< 70	≥ 70			

SpO_2, oxygen saturation in peripheral arterial blood; AVPU scale, 4-point scale used to measure level of consciousness: (i) Alert, (ii) responds to Verbal stimuli, (iii) responds to Pain, and (iv) Unresponsive.

In this thesis, we used the Portsmouth dataset for model training/validation. Data from the CALMS-2 dataset were used as an independent set for testing the models selected during the training/validation phase.

3.1.2 Design of scoring systems

While a number of different track-and-trigger systems have been proposed in the literature, all of them have the same fundamental structure. Single-parameter systems compare selected physiological variables with a simple set of criteria with pre-defined thresholds (usually called the Medical Emergency Team, MET, criteria), and trigger a single response algorithm. EWS-based systems, on the other hand, comprise a multi-parameter, aggregate scoring strategy. Two examples of these different scoring systems are shown in Table 3.2. These systems allocate points in a weighted manner, based on the derangement of the patient's

vital signs from a (typically) arbitrarily agreed "normal" range; i.e., univariate scoring criteria are applied to each physiological variable in turn. Typically, seven scoring bands which involve weightings of 3, 2, 1, 0, 1, 2, and 3, are used from abnormally low values (3) to normal values (0), to abnormally high values (3). The sum of the allocated scores results in the final EWS value. This final value is then compared to a pre-determined threshold (which varies for different systems) and used to escalate care; e.g., to increase the frequency of vital-sign observations, to involve more experienced staff in the care of the patient, or, in some hospitals, to call a rapid response team.

Based on previous literature reviews (Gao et al. [2007]; Smith et al. [2008a,b]) and other recent published material (Prytherch et al. [2010]; Tarassenko et al. [2011]), we identified 26 EWS and 7 MET systems that have been used clinically. While there is a core set of variables that is used in the majority of systems, there is variation in the number of physiological variables included (between 3 to 7). The core set of variables that is used in the majority of systems comprises HR (measured in beats per minute), RR (measure in breaths per minute), systolic BP (measured in mmHg), body temperature (measured in °C), SpO_2 (measured as a percentage), and a level of consciousness (typically, the GCS).

The unique 26 EWS and 7 MET systems identified in the literature are listed in Table 3.3 and Table 3.4, respectively. The physiological components of each system are also indicated.

3.2 Performance reported in the literature

A few literature reviews of physiologically based track-and-trigger systems have analysed their predictive ability for serious adverse outcomes in hospitalised patients. In a landmark review study by Gao et al. [2007], 25 distinct physiological track-and-trigger systems were identified (including EWS systems). The authors carried out a study that included 15 datasets representing hospitals in England and Wales to compare the ability of the activation criteria used in these hospitals to diagnose illness. For this study, the outcome was measured as a composite of death, admission to critical care, "do not attempt resuscitation" (DNR) or cardiopulmonary resuscitation. Each hospital used a different set of calling cri-

Table 3.3: Twenty-six aggregate-weighted track-and-trigger systems identified and their physiological components (marked with ✓).

Year	No.	Study	Heart Rate	Resp. Rate	O_2 saturation	Systolic BP	Temperature	O_2 Support	Age	Consciousness
2000	1	Wright et al. [2000]	✓	✓		✓	✓			✓
2001	2	Subbe et al. [2001]	✓	✓		✓	✓			✓
2001	3	Subbe et al. [2001]	✓	✓		✓	✓		✓	✓
2001	4	Riley and Faleiro [2001]	✓	✓		✓	✓			✓
2001	5	Cooper [2001]	✓	✓		✓	✓			✓
2003	6	Subbe et al. [2003]	✓	✓		✓	✓			✓
2004	7	Rees and Mann [2004]	✓	✓			✓			✓
2004	8	Allen [2004]	✓	✓		✓	✓			✓
2005	9	Goldhill et al. [2005]	✓	✓	✓	✓	✓			✓
2005	10	Chatterjee et al. [2005]	✓	✓	✓	✓	✓			✓
2005	11	Andrews and Waterman [2005]	✓	✓		✓	✓			✓
2006	12	Paterson et al. [2006]	✓	✓	✓	✓	✓			✓
2006	13	Smith et al. [2006a]	✓	✓		✓	✓			✓
2006	14	Lam et al. [2006]	✓	✓		✓	✓			✓
2006	15	Gardner-Thorpe et al. [2006]	✓	✓		✓	✓			✓
2007	16	Subbe et al. [2007]	✓	✓			✓		✓	✓
2007	17	Odell [2007]	✓	✓			✓			✓
2007	18	Hancock and Durham [2007]	✓	✓		✓	✓			✓
2007	19	Duckitt et al. [2007]	✓	✓	✓	✓	✓			✓
2007	20	Lilienfeld-Toal et al. [2007]	✓	✓	✓	✓	✓			✓
2007	21	Lilienfeld-Toal et al. [2007]	✓	✓	✓		✓	✓		✓
2008	22	Oxford-based MEWS	✓	✓	✓	✓	✓			✓
2010	23	Prytherch et al. [2010]	✓	✓	✓	✓	✓	✓		✓
2011	24	Tarassenko et al. [2011]	✓	✓	✓	✓	✓			✓
2012	25	RCP [2012]	✓	✓	✓	✓	✓	✓		✓
2014	26	Badriyah et al. [2014]	✓	✓	✓	✓	✓	✓		✓

Table 3.4: Seven single-parameter track-and-trigger systems identified and their physiological components (marked with ✓).

Year	No.	Study	Heart Rate	Resp. Rate	O_2 saturation	Systolic BP	Temperature	O_2 Support	Age	Consciousness
1995	1	Lee et al. [1995]	✓	✓		✓	✓			✓
2003	2	Ball et al. [2003]	✓	✓	✓	✓	✓			✓
2004	3	DeVita et al. [2004]	✓	✓	✓		✓			✓
2006	4	Bell et al. [2006]	✓	✓			✓			✓
2006	5	Bell et al. [2006]	✓	✓			✓			✓
2006	6	Bell et al. [2006]	✓	✓			✓			✓
2007	7	Subbe et al. [2007]	✓	✓						✓

teria and the diagnostic accuracy of the systems varied widely: sensitivities and positive predictive values were low, with mean (± standard deviation) values of 49.2% (±29.6) and 36.6% (±14.8), respectively[1]. Specificities were generally acceptable for such systems, with a mean value of 76.1% (±26.0). An overview of these systems was also performed by Smith et al. [2008a,b], who reviewed 33 unique aggregate weighted track-and-trigger systems (i.e., EWS-based systems) in clinical use (Smith et al. [2008a]), and several single-parameter scoring systems in clinical use (Smith et al. [2008b]). The authors assessed their ability to discriminate between survivors and non-survivors who were admitted to the MAU of a large hospital in the UK (this corresponds to part of the Portsmouth dataset described in the previous chapter). With respect to single-parameter scoring systems, the main conclusion of the study was that, although specificities were high, sensitivities were too low to provide institutions with confidence that single-parameter scoring systems could identify patients at risk of in-hospital death using admission vital signs (Smith et al. [2008b]). With respect to the EWS-based systems, the performance of most systems tested was also poor with

[1]The mean and standard deviation values were calculated using the Electronic Supplementary Material provided by the authors (Gao et al. [2007]).

AUROCs that ranged from 0.657 (95% confidence interval, CI, of 0.634-0.678) to 0.782, (CI of 0.767-0.797) (Smith et al. [2008a]).

The calling criteria of the EWS systems described in the review papers mentioned above were developed empirically. They are based on clinical experience alone, or they have been related to the prevalence of abnormal physiological signs in the ward population. Although these variables seem clinically intuitive and rational, they include best-guess physiological variable ranges and thresholds, and typically lack clinical validation. Some recently-proposed EWS systems do not rely solely on expert opinion.

In the ViEWS system, Prytherch et al. [2010] used the Portsmouth dataset (see section 2.3.2) to set the scoring ranges of values for each physiological variable. ViEWS was developed using an iterative, pragmatic, "trial and error" approach, with the cutoffs for its scoring bands (0, 1, 2, and 3) being adjusted to maximise its ability to predict in-hospital death within 24 hours of a vital-sign observation set. An additional feature was added to the scoring system: if the patient is on oxygen support (via an oxygen mask or cannula), an additional score of 3 is added to the aggregated score (Table 3.2). Using the AUROC as the performance metric, the authors evaluated the comparative performance of other published EWS systems (which have been previously reviewed in Smith et al. [2008a]). ViEWS performed better than any other EWS for the outcomes tested (death or discharged alive from hospital within 6, 8, 12, 18 and 24 hours of the vital-sign observation set), with an AUROC of 0.888 (CI 0.880-0.895). No attempt was made to modify ViEWS with respect to maximising its ability to predict any other outcome. One criticism of this approach is that the range and cutoff values for each vital sign were set such that alerts are meant to predict a binary outcome (mortality/survival) at some later time. The true benefit of any EWS should be its ability to recognise patients who are deteriorating but who can have their outcome changed by a timely intervention, rather than identify those who are very likely to die independent of intervention (Tarassenko et al. [2011]).

Since its publication, ViEWS has been used in different studies to assess its performance in other cohorts of patients and other hospitals (Hands et al. [2013]; Kellett et al. [2013]; Opio et al. [2013]). Members of the RCP of London, who

formed the National Early Warning Score Design and Implementation Group (NEWSDIG), made some adjustments to ViEWS, based on clinical opinion, to develop and establish a standard scoring system to be used across the NHS: the "National EWS" (RCP [2012]). An example of a modification performed was the cutoff value and score assigned to high systolic blood pressure. The ability of NEWS to discriminate patients at risk of cardiac arrest, unanticipated ICU admission, or death within 24 hours was tested using the same database of vital signs that was used to develop its precursor, ViEWS. The AUROC values (95% CI) for NEWS for cardiac arrest, unanticipated ICU admission, death, or any of these outcomes, all within 24 hours, were 0.722 (0.685-0.759), 0.857 (0.847-0.868), 0.894 (0.887-0.902), and 0.873 (0.866-0.879), respectively. NEWS showed the best performance of all the 34 EWSs used in the comparison.

More recently, Badriyah et al. [2014] used the same dataset, as used in the original ViEWS paper and the analysis of NEWS, to generate an EWS entirely algorithmically using decision trees. A combined outcome of cardiac arrest, unanticipated ICU admission or death within 24 hours of a given vital signs observation was used. Despite the different processes behind their development, the structures of NEWS and the decision tree-based EWS are very similar, with the two major differences being the weightings assigned to low RR and high systolic BP values. Unsurprisingly, the performance of the two scoring systems was virtually identical for all three outcomes and the composite outcome described above.

Tarassenko et al. [2011] used a different approach and proposed a "centile-based EWS" system derived from statistics of the distributions of vital signs acquired from at-risk hospitalised patients. A large dataset of continuously-recorded vital-sign data (using bedside monitors) acquired from 863 acutely-ill patients was used to investigate the distributions of each vital sign, and thus to construct a new aggregate centile-based alerting system with seven bands for each vital sign (i.e., scores of 3, 2, 1, 0, 1, 2, and 3, as used in other systems such as ViEWS). Observations were treated as being abnormal if they occur in the extremes of the distributions of the vital signs. Although the average values for each vital sign are remarkably similar to previously published values for hospitalised patients, the values of the score boundaries in CEWS are different, and differ most for RR and systolic BP. To date, no validation study of the CEWS system has been

published, but it is currently undergoing clinical validation on trauma wards in a medium-sized teaching hospital in Oxford, UK.

Most EWS systems have been used in a variety of clinical circumstances, locations, and cohorts of patients, including oncology patients (Cooksley et al. [2012]), and both medical and surgical wards patients (De Meester et al. [2013]; Ludikhuize et al. [2012]; Paterson et al. [2006]). Only a few studies report the performance of such systems in the care of post-operative patients alone. Although generic systems assessed in surgical patients have been found to have reasonable overall sensitivity and specificity for adverse outcomes, positive predictive values have been reported to be very low (Gardner-Thorpe et al. [2006]; Smith et al. [2012]).

Smith et al. [2012] reported the impact of an expanded version of MEWS in predicting clinical deterioration for surgical patients. The system included the basic physiological parameters included in earlier versions of MEWS, but added some new parameters: urinary output, and a subjective parameter that reflects the nurse's level of concern about the patient's condition. The authors looked at 592 patients admitted to the general and trauma surgery wards of a level 1 trauma center in the Netherlands, in which 8% of patients had the composite outcome of death, resuscitation, unexpected ICU admission, emergency operation, or severe complication. The specificity of MEWS, for a score of at least 3, was 82% and the negative predictive value 97%, while the sensitivity was 74%, and the positive predictive value was 26%. The AUROC was 0.87. Similar results were reported by Gardner-Thorpe et al. [2006], who prospectively studied a total of 334 emergency and elective surgical patients.

Cuthbertson et al. [2007] tested the ability of existing EWS systems to predict major deterioration in a patient's specific condition. Two cohorts of general surgical high-dependency patients, a group of 72 patients who did not require ICU admission and another group of 72 patients who did require ICU admission, were used. Existing EWS systems had good discriminatory power between groups at 24 hours before ICU admission, with an AUROC of 0.80. Other small studies (Kyriacos et al. [2014]; Odell et al. [2002]; Peris et al. [2012]) describe the use of EWS-based systems and, although they do not report quantitatively the performance of these systems, they suggest that the use of simple and reproducible

score systems may help in reducing ICU admissions after surgery.

In summary, a striking observation in this short review of the performance of track-and-trigger systems is the remarkable heterogeneity of the studies presented in terms of:

1. the type of analysis performed, which may be retrospective, prospective, observational, or interventional[1];

2. population studied, involving a variety of settings, different inclusion and exclusion criteria, and different types of time periods analysed;

3. patient outcomes considered, which may include different subsets of serious adverse events;

4. methods used, which include different physiological variables and other subjective variables introduced by authors.

In particular, the type of populations investigated in these studies typically includes both medical and surgical patients in different hospital settings. It has been recognised that scoring systems need to be developed taking into account the population for which they will be used (Cuthbertson et al. [2007]). Nevertheless, commonly recommended track-and-trigger systems have not been developed using specific patient populations (Prytherch et al. [2010]; RCP [2012]; Tarassenko et al. [2011]). Another important observation is that, although the sample size is a source of variability between the studies, the absence of previously-unseen and independent test sets, is not. The most recent data-driven systems such as those proposed by Prytherch et al. [2010] and Badriyah et al. [2014], benefit from a very large sample size (data from 35,585 medical admissions) but present *in-sample* analysis (i.e., no previously-unseen data were used to test the scoring system), which may influence the results reported.

[1]Primary research has been categorized in different ways; common categorization schema include, among others, temporal nature of the study design (retrospective or prospective), or role of the investigator (observational or interventional).

Table 3.5: Identification of low and high threshold values (thresh.) for the rejection of physiologically-implausible observations, and values corresponding the 1^{st} and 99^{th} percentiles over both datasets (as an indication of the overall variability of the data).

Variable (units)	1^{st} percentile	99^{th} percentile	Low thresh.	High thresh.
Heart Rate (bpm)	48	135	30	300
Resp. Rate (rpm)	10	32	3	60
O_2 saturation (%)	87	100	60	101
Temperature (^{o}C)	35.2	38.2	32.0	42.0
Systolic BP (mmHg)	78	189	40	300
Diastolic BP (mmHg)	38	110	20	200

3.3 Performance on study populations

This section provides a baseline with which some of the results obtained in the analysis described in this thesis are compared. Observational data from both datasets (see sections 2.3.1 and 2.3.2) were used in this analysis. Each vital-sign observation set in both databases contains: date/time of observation set, HR, RR, SpO_2, systolic and diastolic BP (diastolic BP is not used by any of the systems studied), temperature and consciousness status (using the AVPU scale). Additionally, each observation set also contains the age of the patient at admission to the hospital, and a variable that indicates whether or not the patient was on oxygen support during the observation of the vital signs. The performance of the available scoring systems on the vital-sign databases is discussed in the next sections.

3.3.1 Data pre-processing

Data pre-processing is an integral part of the data analysis and evaluation procedures that are described in this thesis. It is desirable to avoid "over-fitting" the pre-processing algorithm to a particular dataset, to avoid poor generalisation on previously-unseen test data. It is therefore generally good practice to not to select the pre-processing method on the basis of metrics derived from the data

used to evaluate the performance of the algorithm. Given the cohorts of patients involved (all of whom were adult patients), the physiological variables being measured may inform pre-processing procedures that are universally applicable. Our pre-processing involved the following steps: identification of specific artefacts and coding to represent missing data (e.g., HR= -1); conversion of measurements entered with different units (e.g., % instead of unit-ratio for SpO_2); and rejection of physiologically-implausible values. The range of physiologically-implausible values used in this work was that used in previous works by Hann [2008], Wong [2011] and Hugueny [2014] (when applicable). Table 3.5 shows the lower and upper cutoffs for the physiological variables used in the study.

This "data cleaning" procedure removed artefactual observations from the observation sets in both datasets. Observation sets with more than two missing variables (after the pre-processing procedure) were not included in this analysis. For those observation sets with one or two missing variables, the missing value was replaced by the population mean value of the corresponding variable[1].

3.3.2 Experimental setting

Some scoring systems described in the literature include variables that are not available in the database used in this thesis. Thus, we focused on the analysis of the 26 EWS systems detailed above (for which all variables are included in the database). It has been suggested that the system proposed by Prytherch et al. [2010] (and later refined in RCP [2012]) has a better performance when applied to a general hospital population, but no strong evidence exists with respect to the performance of these systems in sub-populations such as post-surgical patients. We implemented 26 unique EWS (Table 3.3) systems, and evaluated their performance on both databases of physiological data. The advantages of EWS systems over the MET criteria, as those presented in Table 3.4, have been shown in previous studies (Gao et al. [2007]; Smith et al. [2008a]). The performance of the latter systems was considerably lower when evaluated in a range of patient populations. For this reason, we did not include the MET criteria in our analysis.

[1]By replacing the missing value with the population's mean value of the corresponding variable, we assume that this variable does not contribute to the final score calculated with each scoring system, as the score assigned to this particular variable is 0.

To assess performance, we followed the methodology developed by Prytherch et al. [2010]. For the Portsmouth dataset, we used the combination of the outcome at discharge (alive or dead), date/time of discharge, and date/time of each observation set to derive a "diagnostic test result" (a derived outcome): the status at 24 hours post-observation set. This derived outcome variable was added to the database for each observation set, and was used as the class membership variable in our analysis. That is, each observation set is labelled according to whether or not the patient was dead within 24 hours of that observation set (i.e., this derived outcome is positive for observation sets occurring within 24 hours of an adverse event, and 0 otherwise). Similarly, for the CALMS-2 dataset, we used the same strategy to define the derived outcome. In this dataset, however, we used a composite outcome of death on the ward, cardiac arrest, and emergency admission to ICU within 24 hours of each observation set. It may be argued that a single outcome (death) should have been used in the CALMS-2 database. Only 7 patients died during the CALMS-2 trial, however. Because of the problems associated with small-scale statistics, it was decided to use a composite outcome that includes any of the three major adverse events (death, cardiac arrest, and emergency admission to ICU). Ideally, the same composite outcome should also be used for the Portsmouth dataset. However, only the in-hospital mortality information was available to us, and so the single outcome (death/alive) was used with the Portsmouth dataset.

The simplicity of the methods and systems analysed is such that no validation procedure was performed to select the "model" and evaluate the overall performance; these scoring systems only involve simple thresholding techniques and do not require calibration coefficients. Therefore, the entire Portsmouth dataset (as in Prytherch et al. [2010]) was then used for evaluating each system using the AUROC and for selecting the optimal operating point (threshold) of each EWS system. The optimal operating point (i.e., the optimal value of the threshold above which an alert is generated by the EWS system) was selected using ROC analysis, based on the minimisation of the loss function introduced in Equation 3.3, with $\upsilon = 0.5$; i.e., the relative importance of false positives is assumed to be equal to that of false negatives. All systems were then tested using the (independent) CALMS-2 dataset and the thresholds determined during the

Table 3.6: Characteristics of the observation sets included in the analysis from the 34,060 patient-episodes (N = 196,598): Obs_{Abn} includes the observations that were followed by death within 24 hours; and Obs_{Nor} includes the observation that were not followed by death within 24 hours.

	Mean (SD)	
Variable	Obs_{Nor} (N = 195,012)	Obs_{Abn} (N = 1,586)
Heart Rate	82 (19)	94 (22)
Respiratory Rate	17 (4)	22 (6)
SpO_2[a]	96 (95-96-98)	93 (91-95-97)
Temperature	36.7 (0.4)	36.6 (0.7)
Systolic BP	127 (22)	112 (25)
Oxygen support[b]	44,988 (23.1)	1,208 (76.2)
Level of consciousness (AVPU scale):[b]		
- A	181,862 (93.3)	1,068 (67.3)
- V	10,226 (5.2)	351 (22.1)
- P	2,236 (1.1)	67 (4.2)
- U	688 (0.4)	100 (6.3)

[a]Values presented as mean (25^{th}-50^{th}-75^{th} quantiles); [b]Values presented as number of observations (%).

"evaluation" step.

3.3.3 Performance on the Portsmouth patient population

A total number of 196,598 completed observation sets from 34,060 patient-episodes were considered for analysis (no observation sets were removed in the pre-processing step). The number of observation sets followed by death within 24 hours (i.e., number of observation sets labelled as positive) was 1,586 (1%). Table 3.6 summarises the characteristics of the observation sets (vital-sign means and standard deviations) included in each class that were used to evaluate the performance of the systems in this dataset: (1) observation sets followed by death within 24 hours, Obs_{Abn}; and (2) observation sets not followed by death within 24 hours, Obs_{Nor}.

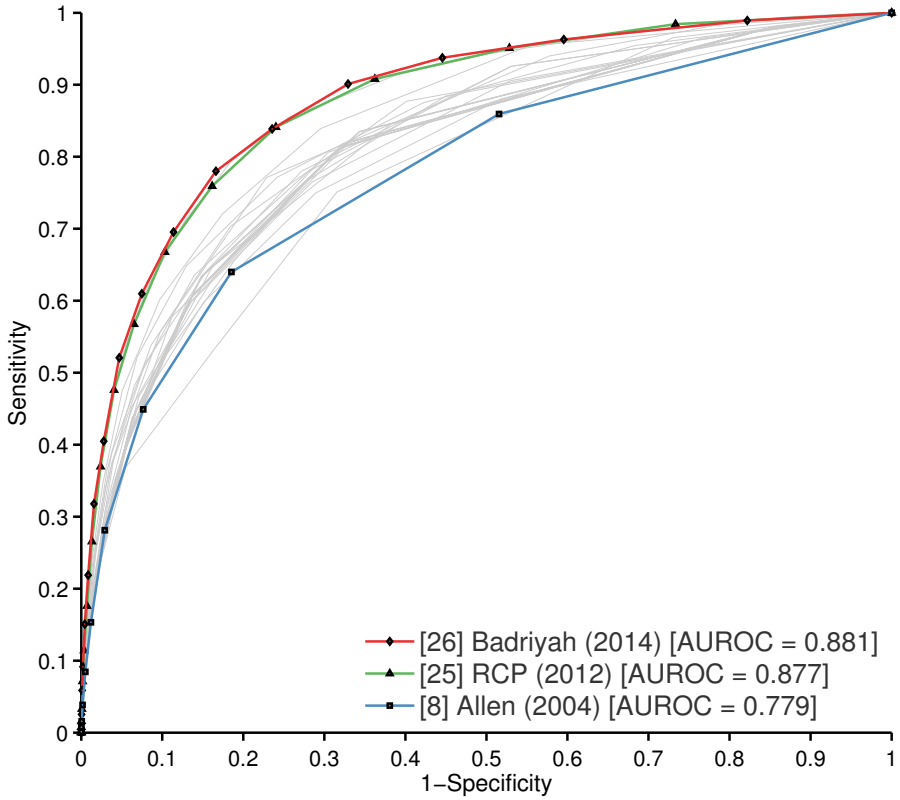


Figure 3.1: The ROC curve for each EWS system evaluated for the 34,060 patient-episodes in the Portsmouth dataset are shown in light gray. The best- and worst-performing EWS systems are shown in red and blue, respectively (the system recommended by the RCP [2012] is shown in green, for reference).

The ROC curves and "efficiency curves" obtained are shown in Figure 3.1 and Figure 3.2, respectively. Confidence intervals for the AUROC were estimated directly from the data using the standard error (SE) of the Wilcoxon statistic, which was determined with the exponential approximation formula of Hanley and McNeil [1982] (further described in Hajian-Tilaki and Hanley [2002]). Table 3.7[1] gives the different performance measures computed for each scoring system on the Portsmouth population of patients.

As reported in recent literature, the performance of the system [23] proposed by Prytherch et al. [2010] (ViEWS) and its successors (systems [25] and [26], pro-

[1]For ease of reading, the standard error associated with the mean values displayed in the table are omitted. The complete version of the table can be found in Appendix A.

Table 3.7: Performance metrics for the 26 track-and-trigger systems evaluated in the 34,060 patient-episodes in the Portsmouth dataset (with the outcome being death within 24 hours). The results are presented in descending order of AUROC, and the best values for each performing metric are underlined.

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[26]	0.881	0.827	0.780	0.834	0.037	0.146
[25]	0.877	0.854	0.841	0.760	0.028	0.125
[23]	0.877	0.862	0.793	0.815	0.034	0.139
[21]	0.854	0.777	0.721	0.826	0.033	0.128
[19]	0.836	0.728	0.772	0.758	0.025	0.110
[3]	0.835	0.712	0.771	0.771	0.027	0.115
[9]	0.827	0.752	0.780	0.719	0.022	0.099
[20]	0.827	0.744	0.780	0.728	0.023	0.102
[12]	0.826	0.680	0.703	0.816	0.030	0.119
[11]	0.815	0.799	0.835	0.656	0.019	0.092
[15]	0.814	0.798	0.835	0.657	0.019	0.092
[14]	0.814	0.798	0.835	0.657	0.019	0.092
[5]	0.812	0.764	0.806	0.695	0.021	0.097
[10]	0.811	0.748	0.765	0.715	0.021	0.095
[22]	0.811	0.739	0.766	0.723	0.022	0.097
[18]	0.810	0.829	0.647	0.838	0.032	0.117
[6]	0.809	0.779	0.818	0.677	0.020	0.094
[13]	0.809	0.778	0.813	0.679	0.020	0.094
[1]	0.809	0.779	0.818	0.677	0.020	0.094
[2]	0.809	0.778	0.818	0.677	0.020	0.094
[4]	0.809	0.785	0.821	0.668	0.020	0.093
[17]	0.808	0.776	0.814	0.680	0.020	0.094
[7]	0.808	0.775	0.814	0.680	0.020	0.094
[24]	0.796	0.746	0.750	0.711	0.021	0.090
[16]	0.780	0.749	0.751	0.685	0.019	0.084
[8]	0.779	0.668	0.640	0.815	0.027	0.104

posed by RCP [2012] and Badriyah et al. [2014], respectively) for predicting death within 24 hours, are superior to the other 23 EWS systems used in this analysis. The AUROC (SE) for system [26] was 0.881 (0.006), which is much higher than

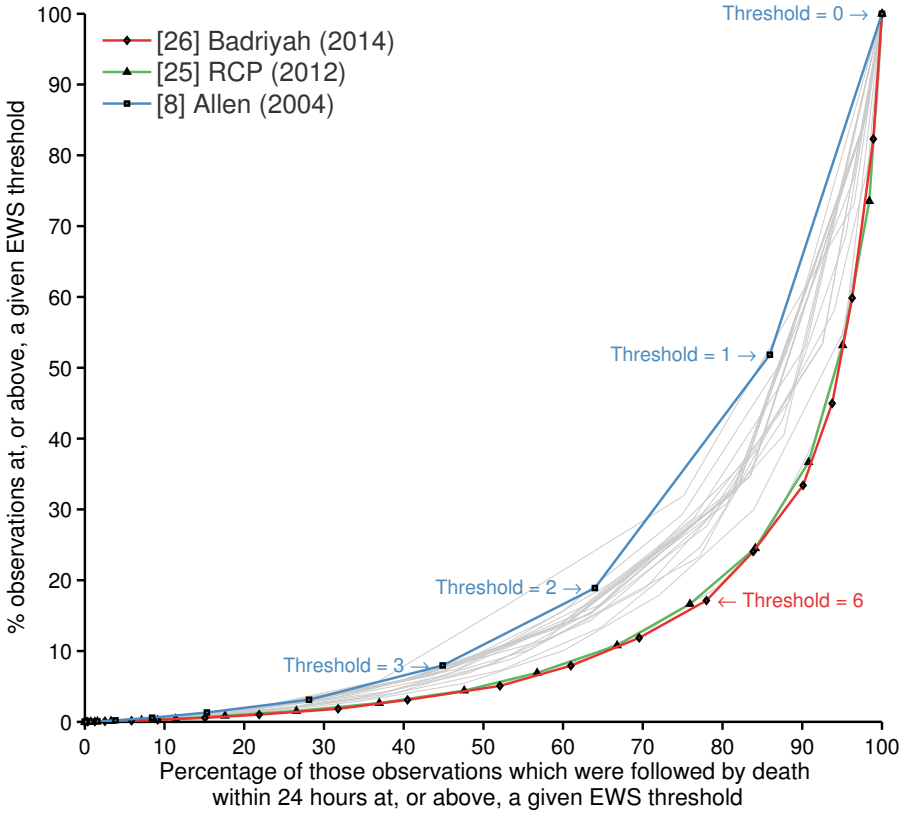


Figure 3.2: "EWS efficiency curve"* comparing different EWS systems for the outcome - death within 24 hours of a given observation set - evaluated in the Portsmouth dataset. The best and worst performing EWS systems are shown in red and blue, respectively (the system recommended by the RCP [2012] is shown in green, for reference). The curves correspondent to the other systems are shown in light gray.
* From the point at (100,100) the EWS values are 0, 1, 2, 3, ... for each EWS system.

that of the worst performing EWS system (system [8]), with an AUROC of 0.779 (0.007), as shown in Table 3.7. Furthermore, the best sensitivity-specificity pair was also obtained by the system [26]. This is not entirely unexpected, as the three best performing systems were developed using this database of vital signs.

A fundamental observation from the results presented in Table 3.7 is the low values of PPV and MCC. As discussed above, these low values reflect the imbalanced dataset considered. This is an important consideration when assessing the performance of these systems on other databases (in which the imbalance ratio may be different). Nevertheless, these measures can be used to compare the

performance of the EWS systems within the same database. Hence, we observe that all performance measures show the superiority of the top-three performing systems on this dataset. Another interesting observation is that, although systems [25] and [23] have virtually the same overall AUROC value, the EWS system [23] has a higher pAUROC. This shows that system [23] (by Prytherch et al. [2010]) has a better performance than the system [25] proposed by the RCP [2012] in the range of specificities that are clinically acceptable. This is confirmed by the values of other threshold metrics such as the PPV and MCC values.

We may also investigate the performance of the different systems using the concept of the "efficiency curves" (Figure 3.2). For each scoring system, this plot provides a relative measure of the number of "triggers" that would be generated for different EWS threshold values and permits the comparison of the estimated "workload" generated by different scoring systems. As an example, a score of 1 for the worst-performing system (Allen [2004]) would generate a trigger for around 50% of observations, and this would "identify" (approximately) 85% of all deaths within 24 hours of the observation set. A score of 4 with the NEWS system (RCP [2012]) would generate a trigger for far fewer observations (around 22%), while being able to "detect" (approximately) the same percentage of events. To detect the same proportion of deaths, the workload required by the latter system would therefore be much lower (assuming that triggers correspond to workload, as clinicians attend to triggering patients).

We note that the results obtained in the study published by Prytherch et al. [2010], which has only considered the analysis of these systems in terms of AUROC, are very similar to those presented here, as the same database of vital signs (and derived outcome variable) is used here to evaluate these systems. Prytherch et al. [2010] also evaluated the systems for subsidiary derived outcome variables that include death or alive at 12, 48, 72, 96 and 120 hours post-observation set. As expected, the performance of all systems was inversely proportional to the period of time considered for detecting deterioration. More significantly, this change in performance according to the different outcomes was fairly consistent over all scoring systems studied (i.e., the performance of all EWS systems decreased in a similar manner as the time interval between the vital sign observation set and the adverse event increased).

Table 3.8: Completeness of the observation sets (obs.) from the 407 post-operative patients (N = 32,961) included in the CALMS-2 dataset after the pre-processing procedure.

	N (%) obs.
Observation sets with 1 variable missing:	
Heart Rate	597 (1.8)
Respiratory Rate	658 (2.0)
SpO_2	300 (0.9)
Temperature	3710 (11.3)
Systolic BP	225 (0.7)
Level of consciousness (AVPU scale)	10200 (30.9)
Observation sets with more than 2 variables missing	316 (1.0)
Observation sets with all variables	19591 (59.4)

As mentioned above, the three best-performing systems were developed using the Portsmouth dataset, which was also used to evaluate their performance. For this reason, it is important to *assess* these systems on an independent test set, using the "optimal" threshold as determined on the Portsmouth dataset.

3.3.4 Performance on the CALMS-2 patient population

A total number of 32,961 observation sets from the 407 post-operative patients have been considered for this analysis. Data were pre-processed in order to remove artefactual observations from the observation sets. Table 3.8 shows the completeness of the observation sets considered for analysis. Following the pre-processing procedure described above, 32,645 (32,961-316) observation sets were included in the final analysis[1].

During the study, there were a total of 56 major adverse events on the post-operative ward: 51 emergency admissions to the ICU, 6 of which were preceded by a cardiac arrest, and 5 deaths (the other two deaths occurred in the ICU). The

[1]We observe that a significant amount of observations do not include level of consciousness. According to research nurses involved in the trial, this value was often omitted for patients who were Alert (which yields a score of 0).

Table 3.9: Characteristics of the observation sets included in the analysis from the 407 patients (N = 32,645): Obs_{Abn} includes the observations that were followed by a major adverse event within 24 hours; and Obs_{Nor} includes the observation that were not followed by major adverse events.

	Mean (SD)	
Variable	Obs_{Nor} (N = 31419)	Obs_{Abn} (N = 1226)
Heart Rate	85 (15)	97 (22)
Respiratory Rate	17 (3)	20 (6)
SpO_2[a]	97 (95-97-98)	96 (94-96-98)
Temperature	36.5 (0.6)	36.6 (0.7)
Systolic BP	127 (22)	126 (29)
Oxygen support[b]	16458 (52.4)	1120 (91.4)
Level of consciousness (AVPU scale):[b]		
- A	30843 (98.2)	1182 (96.4)
- V	535 (1.7)	33 (2.7)
- P	8 (0.0)	6 (0.5)
- U	33 (0.1)	5 (0.4)

[a]Values presented as mean (25^{th}-50^{th}-75^{th} quantiles); [b]Values presented as number of observations (%).

number of observation sets followed by a major adverse event within 24 hours was 1226 (3.8%). Table 3.9 summarises the characteristics of the observation sets (vital signs means and standard deviations) included in each class that were used to evaluate the performance of the systems: (1) observation sets followed by a major adverse event within 24 hours, Obs_{Abn}; and (2) observation sets not followed by a major adverse event, Obs_{Nor}.

Using the 26 EWS systems evaluated in the preceding section, we determined their performance in the CALMS-2 dataset. Table 3.10[1] shows the performance measures computed for each scoring system on the post-operative population of patients. Figure 3.3 shows the results displayed as pairs of sensitivity and specificity values; recall that the optimal model parameter (optimal threshold)

[1]Again, for ease of reading, the standard error associated with the mean values displayed in the table are omitted. The complete version of the table can be found in Appendix A.

Table 3.10: Performance metrics for the twenty-six track-and-trigger systems evaluated in the 407 patients in the CALMS-2 dataset (for the combined outcome of cardiac arrest, unanticipated admission to ICU and death occurring within 24 hours of a given observation set). The results are presented in descending order of AUROC, and the best values for each performing metric are underlined.

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[23]	0.841	0.826	0.801	0.765	0.084	0.209
[25]	0.835	0.747	0.829	0.715	0.073	0.190
[21]	0.833	0.753	0.639	0.886	0.131	0.251
[26]	0.829	0.816	0.773	0.768	0.083	0.201
[22]	0.791	0.807	0.711	0.791	0.084	0.193
[10]	0.786	0.677	0.700	0.784	0.081	0.185
[12]	0.784	0.825	0.622	0.872	0.116	0.227
[9]	0.782	0.680	0.697	0.777	0.078	0.179
[4]	0.782	0.720	0.778	0.730	0.072	0.180
[20]	0.781	0.678	0.697	0.779	0.078	0.180
[5]	0.779	0.703	0.752	0.751	0.075	0.183
[6]	0.777	0.713	0.762	0.738	0.073	0.179
[17]	0.777	0.712	0.757	0.742	0.073	0.180
[1]	0.776	0.714	0.762	0.736	0.072	0.178
[7]	0.776	0.711	0.757	0.742	0.073	0.180
[2]	0.776	0.712	0.758	0.739	0.073	0.178
[8]	0.776	0.618	0.603	0.875	0.115	0.221
[13]	0.770	0.710	0.752	0.736	0.072	0.175
[18]	0.766	0.623	0.593	0.861	0.103	0.203
[19]	0.764	0.782	0.570	0.845	0.090	0.179
[15]	0.764	0.802	0.565	0.885	0.117	0.216
[11]	0.764	0.803	0.801	0.612	0.053	0.135
[14]	0.763	0.802	0.801	0.613	0.053	0.135
[24]	0.759	0.693	0.702	0.748	0.070	0.163
[16]	0.744	0.609	0.522	0.868	0.096	0.178
[3]	0.717	0.760	0.501	0.898	0.117	0.202

was determined on the Portsmouth dataset using ROC analysis (previous section). The results are presented in ascending order (bottom to top) of best pair of

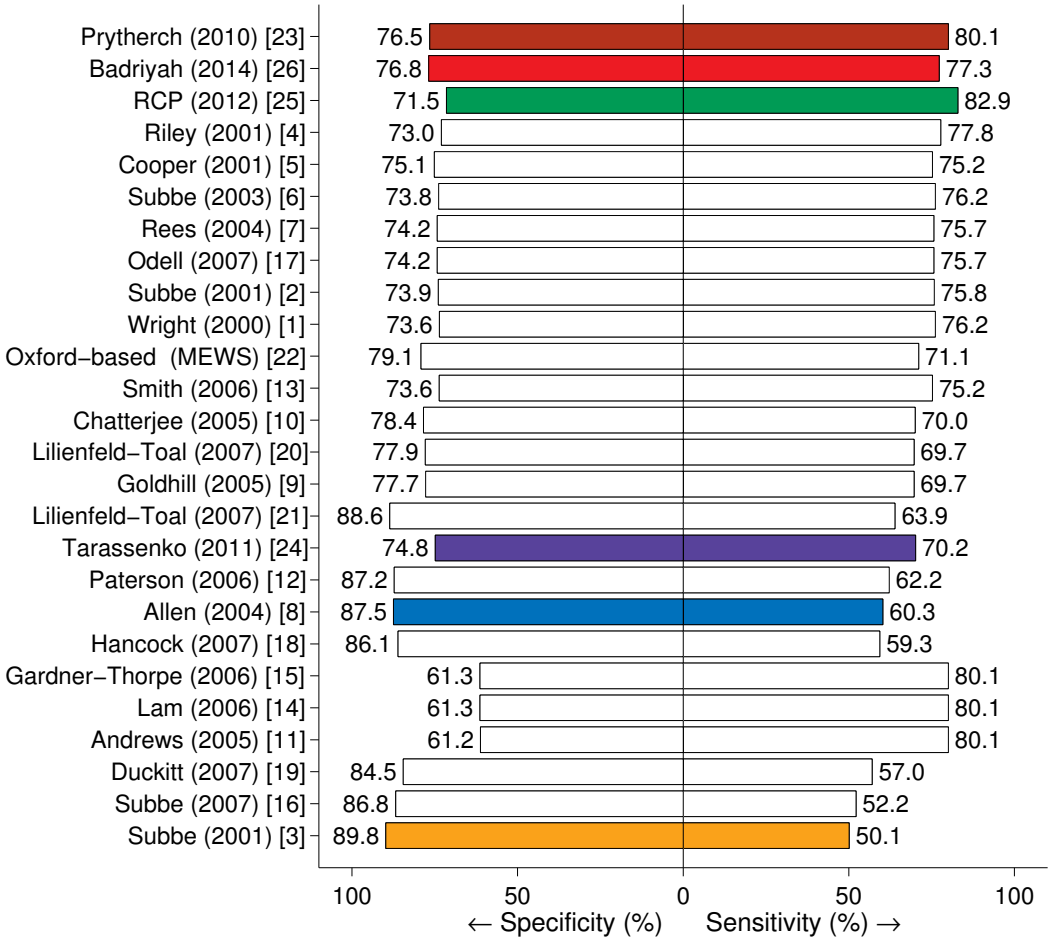


Figure 3.3: Performance of the EWS systems for the combined outcome of cardiac arrest, unanticipated ICU admission or death occurring within 24 hours of a given observation set evaluated in the CALMS-2 dataset.

sensitivity and specificity values, which was determined by minimising the cost function discussed in section 3.1.1 (same procedure as that used in ROC analysis to identify the optimal operating point).

In general, we observe that the values of the AUROC, pAUROC, sensitivity and specificity are lower than those calculated for the Portsmouth dataset; for example, the AUROC ranged from 0.779 (0.007) to 0.881 (0.006) in the Portsmouth dataset, and from 0.717 (0.010) to 0.841 (0.008) in the CALMS-2 dataset. This conveys the idea that these EWS systems have a lower performance when applied

to post-operative patients. Conversely, the ranges of PPV and MCC values were higher in the CALMS-2 dataset. This is an expected result as the imbalance ratio of the positive and negative classes is substantially different from that of the Portsmouth dataset. As discussed in Fawcett [2006], these last two metrics are dramatically affected by the class imbalance of the datasets and, therefore, may not reflect an improvement of the performance of the EWS systems.

We observe that the three best performing systems in the CALMS-2 dataset are the same as those in the Portsmouth dataset in terms of AUROC, sensitivity and specificity values. The EWS system [23] obtained a sensitivity of 80.1% and a specificity of 76.5%. On the other hand, the worst performing system in the test set was system [3], with a higher specificity of 89.8%, but a substantially reduced sensitivity of 50.1%. In terms of PPV and MCC values, the EWS system [21], proposed by Lilienfeld-Toal et al. [2007], achieved the best performance in the post-operative patient population.

A different way of characterising the performance of different systems is by considering the amount of "early warning" of adverse events provided by each of the scoring systems (i.e., how long before the event the scoring systems can identify that the patient is deteriorating). The time to a major adverse event for those patients in the "abnormal" group was determined for the best- and worst-performing scoring systems (at their respective "optimal" threshold). Figure 3.4 represents the cumulative proportion of events identified as a function of the first time (up to 24 hours before the event) that the EWS is at, or above, the optimal threshold. As an example, the best-performing EWS system would generate an alert for 50% of all adverse events 18 hours before the event. That is, the majority of the events would be detected earlier, and clinical review would be prompted in advance of the event. System [3], on the other hand, would predict 50% of the major adverse events 4 hours later than system [23].

3.3.5 Discussion

In order to characterise the performance of currently-used early-warning scoring systems, we used two independent vital-sign datasets acquired from patients admitted to different hospitals and hospital settings. It is, therefore, important

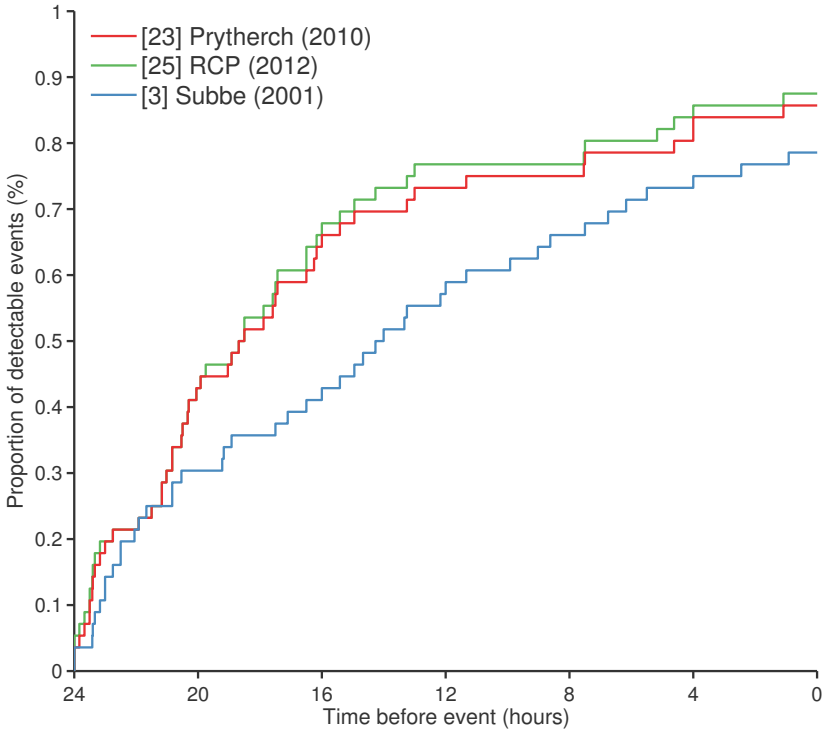


Figure 3.4: Cumulative percentage of adverse events detected as a function of the time before the respective event. The results for the best and worst performing EWS systems are shown in red and blue, respectively (the system recommended by the RCP [2012] is shown in green).

to highlight some substantial differences between the two databases, which may explain some of the results obtained in this analysis.

As mentioned above, the two databases were developed from clinical data obtained from patients in different hospital settings. The Portsmouth dataset includes a more heterogeneous patient population than that included in the CALMS-2 dataset. Another important difference between the two databases is the derived outcome variable used to evaluate these systems. While in the CALMS-2 we used a composite outcome of death, cardiac arrest, and unanticipated ICU admission within 24 hours of an observation set, the only outcome measure available for the Portsmouth dataset was death within 24 hours. Smith [2013] conducted an evaluation of the NEWS system in the same Portsmouth database using the composite outcome as used in the CALMS-2 database, and

no significant differences in the performance of the EWS systems evaluated were observed, in comparison with the single outcome of death within 24 hours.

An important observation is that, from all scoring systems tested in this analysis, the three best performing EWS systems were developed using data-driven approaches. Nevertheless, we observe that the system proposed by Tarassenko et al. [2011], which was also built using a data-driven approach, did not perform as well as one would expect (in comparison with the other data-driven based scoring systems). Chapter 4 will show that the fact that the system proposed by Tarassenko et al. [2011] was constructed using a large dataset of *continuously-recorded* vital-sign data, rather than vital signs *collected manually*, is the main reason for this difference.

The results from the analysis conducted provide a few other insights that are important to consider. An important point to note is the similarity of the top-four best-performing systems on both datasets (Table 3.7 and Table 3.10). A common characteristic shared by these four EWS systems is that they include information about whether the patient is on oxygen support or not (where an additional score is added if the patient is on oxygen support). As shown in Table 3.6, the number of observation sets which were followed by death (within 24 hours) for which the patient was breathing with oxygen support was 1,208 (76.2%), which is substantially higher, in proportion, than the same figure for those observation sets included in the Obs_{Nor} group (23.1%). A similar observation can be made for the CALMS-2 dataset, 91.4% vs. 52.4% (Table 3.9). In both cases, patients who are at higher risk of deterioration will often be on oxygen support, while patients who are about to be discharged from the hospital will be breathing room air. Nevertheless, the inclusion of this extra information as a marker for identifying deterioration may be problematic. Firstly, in the CALMS-2 patient population, patients are generally on oxygen support for the first two/three days after surgery (as part of the clinical protocol for post-operative care), which corresponds to the time during which a post-operative complication is more likely to occur. In medical assessment units (from which the Portsmouth clinical data were acquired), oxygen support is given to patients who are in a state of hypoxia (lack of oxygen). Secondly, the use of an oxygen mask may be seen as the result of a clinical intervention (i.e., a treatment) which will have a direct influence on

the physiological variables (certainly on SpO_2) that are measured during that period. Therefore, although the use of oxygen support is an important factor that increases the performance of scoring systems for predicting a major adverse event, its inclusion in a scoring system may not be straightforward, and may depend on the clinical protocol employed in different hospital settings.

Additionally, the inclusion of age in scoring systems may also not be a simple task. In fact, if we consider systems [2] and [3], in which the same physiological components, ranges and cutoffs are used, except for age, we observe that the system that includes a score for age has a better performance (Table 3.7). This is an expected result according to studies in the literature (Prytherch et al. [2010]; Subbe et al. [2001, 2007]), which have reported that the inclusion of age in the calculation of the EWS conveys some benefit, as older patients are more likely to deteriorate in the hospital than younger patients. On the other hand, age does not appear to have a significant impact for identifying deterioration in the post-operative population included in the CALMS-2 dataset. This result may be justified, however, by the cutoff values used in the scoring system and the significant differences between the populations studied. Some of the scoring systems studied add an additional score if the patient is older than 70 years old. Most studies report the use of EWS systems for patients in medical assessment units, to which patients of all ages are admitted. The post-operative patient population in the CALMS-2 study is more selective and homogeneous: patients undergoing major operations are typically not young; and older patients, due to their increased risk factors, may not be scheduled for major operations. Therefore, the cutoffs and ranges for age would have to be modified according to the demographic information of this patient population. As also observed by Prytherch et al. [2010], most of the systems do not include age as it adds further complexity, conveys little benefit (or none in the case of our study population), and has the potential to raise significant ethical issues.

3.4 Conclusion

The majority of the early warning systems presented here are based on a combination of clinical experience and knowledge of the signs of derangement for

individual physiological variables in hospitalised patients. While these scoring systems are often used on general wards, they may not be suitable for all in-hospital patients. That is, it can be argued that scoring systems need to be developed and validated in specific patient groups and not for large heterogeneous groups of hospital patients (Cuthbertson et al. [2007]). Despite recognition that scores need to be developed taking into account the population for which they will be used, commonly recommended scores have not been developed using specific post-surgical populations (Badriyah et al. [2014]; Prytherch et al. [2010]; Tarassenko et al. [2011]).

Furthermore, we also observe that the "source" of the data used to build EWS systems may have some consequences on their performance when evaluated in both vital-sign databases. Such implications are explored and further discussed in the next chapter.

Chapter 4

The modified centile-based early-warning score

In the preceding chapter, we evaluated the performance of scoring systems for recognising patient deterioration using both the Portsmouth and CALMS-2 vital-sign databases. While the systems proposed by Prytherch et al. [2010] and Badriyah et al. [2014] (ViEWS and NEWS) were developed using the Portsmouth database in which the vital signs were *collected manually*, Tarassenko et al. [2011] used a dataset of *continuously-recorded* vital-sign data to build the centile-based EWS (CEWS). Several studies have reported some fundamental differences between continuously-aquired and manually collected vital-sign data in hospital settings (Sneed and Hollerbach [1992]; Taenzer et al. [2014]; Villegas et al. [1995]), and so we might expect CEWS and ViEWS to differ.

In chapter 2, we described the continuous monitoring system deployed in the CALMS-2 clinical trial. During their first few days on the post-operative ward, patients were connected to conventional bedside monitors, from which a limited subset of continuous physiological variables was acquired (which comprises HR, RR and SpO_2). During the remainder of their stay on the post-operative ward, patients were connected to portable monitoring devices, from which an even smaller subset of vital signs (HR and SpO_2) was recorded, together with the PPG waveforms. In this chapter, we explore the characterisation of the automated continuous data acquired from post-operative patients, and consider

the clinical implications of deploying a system that relies on continuous data.

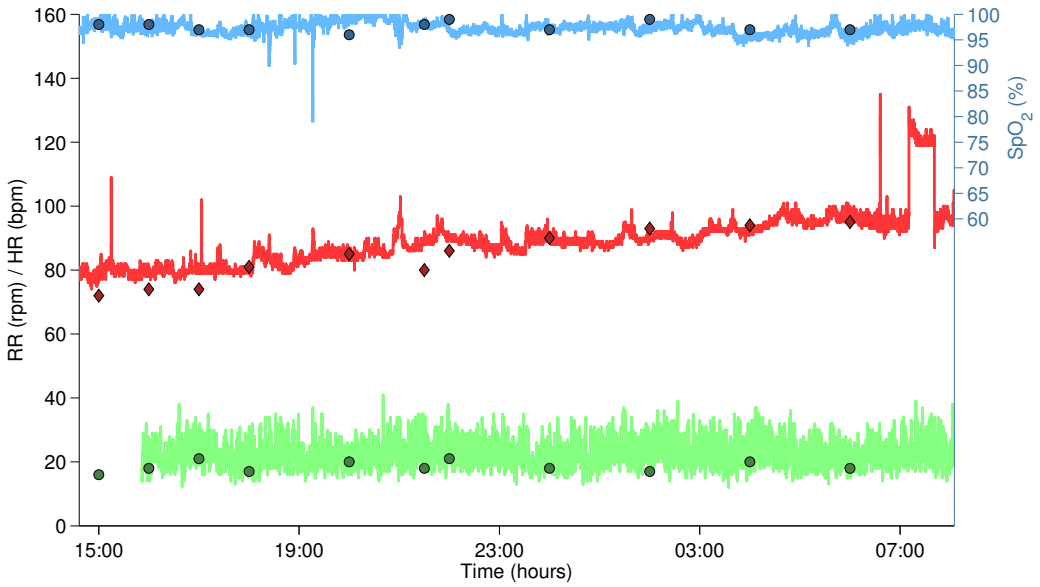
4.1 Comparison of automated continuous monitoring and manual charting

The accuracy of obtaining BP, HR, RR, and SpO_2 values through intermittent manual data collection and charting has been questioned (Bianchi et al. [2013]; Sneed and Hollerbach [1992]; Taenzer et al. [2014]; Villegas et al. [1995]). This uncertainty may have an impact on the reliability of EWS systems. In this chapter, we investigate the reliability of intermittent, manually collected vital-sign data (observational data) with respect to measurements recorded automatically by continuous monitoring systems.

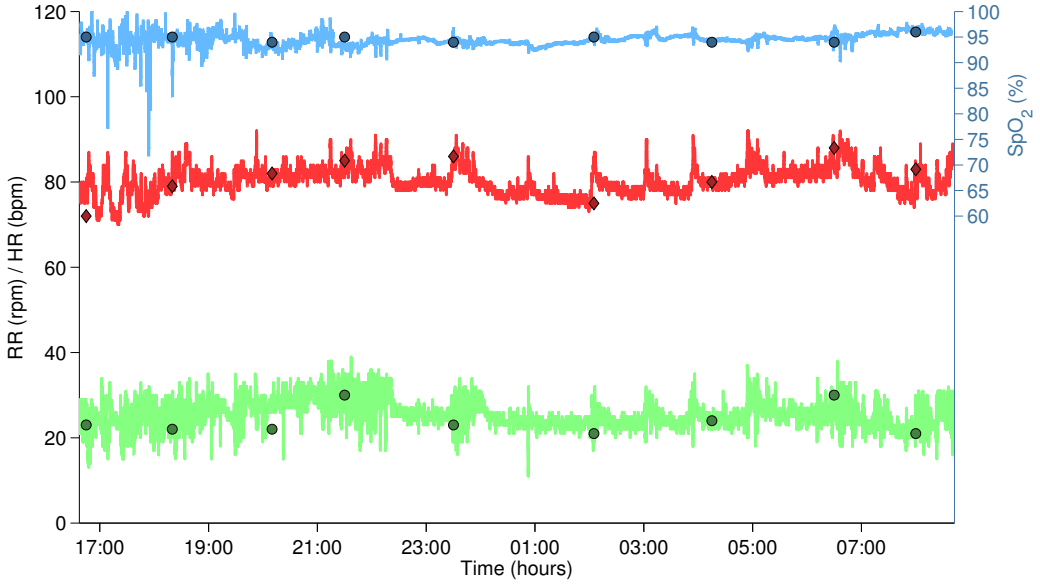
4.1.1 Experimental setting

We here consider data acquired from bedside monitors during the CALMS-2 study (HR, RR, SpO_2, sampled at 1 Hz). Data from the portable monitoring devices were not considered in the analysis described by this chapter, due to the smaller subset of vital signs measured by the mobile monitors. The bedside monitors used include a system of 3 channels of ECG, from which HR and RR are derived, and an in-built pulse oximeter that provides measurements of SpO_2. While patients were connected to bedside monitors, manual charting of vital signs during routine observations by nursing staff was performed, as is normal care on the ward. Figure 4.1 shows examples of vital-sign data for two patients during the period of bedside monitoring for each. Note that the manual observations made by nursing staff and recorded on track-and-trigger charts take place every two to four hours.

From the cohort of 407 patients in our study, we selected patients for whom two or more hours of continuous data had been acquired from the bedside monitors. As a result, data from 269 patients were considered for analysis. We then took a similar approach to that proposed by Taenzer et al. [2014], which we describe next. The first step in the continuous data analysis involved extraction of data for each patient for the 10-minute interval (ΔT_1) that ended 5 minutes before



(a) Patient 12



(b) Patient 158

Figure 4.1: Vital-sign data measurements for two example patients during the period for which each patient was connected to bedside monitors: solid lines correspond to data acquired from the bedside monitors, and manual observations are shown with darker-colour markers. RR is shown in green, HR in red, and SpO_2 in blue (refer to the right-hand axis for the latter). We only display the three vital signs analysed in this study. (Measurements of blood pressure and temperature were also performed, but not continuously.)

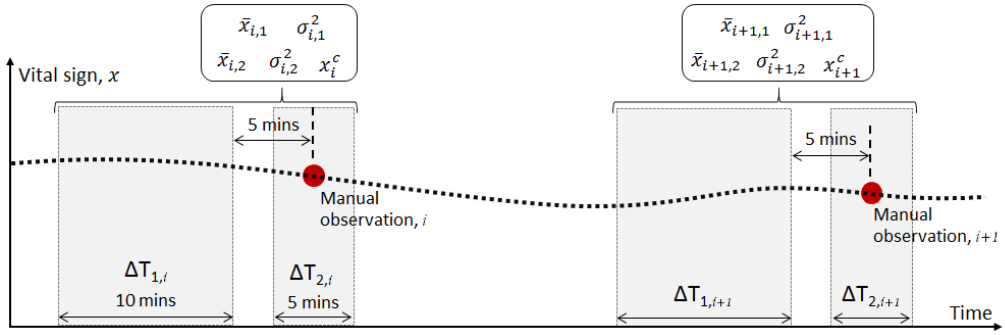


Figure 4.2: Timeline of patient continuous data collection (dotted line) and manual observations (red circles), showing the 10-minute interval ending 5 minutes before each manual data charting occurrence (ΔT_1), and the 5-minute interval centered on each manual data charting time (ΔT_2). For the three vital signs, for each manual observation, i, we extracted the mean and variance of values, $\bar{x}_{i,j}$, and $\sigma^2_{i,j}$, with $j = \{1, 2\}$, corresponding to intervals ΔT_1 and ΔT_2 for respective value of j. The charted vital sign is denoted x^c_i.

the recorded time for each manual observation, i (see Figure 4.2). Continuous data were also retrieved from a 5-minute interval around the manual observation (± 2.5 minutes), ΔT_2, for comparison. For each sampling interval, $j = \{1, 2\}$, the mean and variance were calculated:

$$\bar{x}_{i,j} = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{ij,k} \tag{4.1}$$

$$\sigma^2_{i,j} = \frac{1}{n_j} \sum_{k=1}^{n_j} (x_{ij,k} - \bar{x}_{ij})^2, \tag{4.2}$$

where $\bar{x}_{i,j}$ and $\sigma^2_{i,j}$ correspond to the mean and variance of the n_j values for the sampling interval j that is associated with observation i; and n_j corresponds to the number of samples in each interval, $n_j = 600$ if $j = 1$ (for ΔT_1), and $n_j = 300$ if $j = 2$ (for ΔT_2).

Taenzer et al. [2014] considered the analysis of all patients who had a mean SpO_2 of 90% or less over the period ΔT_1 (which in their case corresponded to a 15-minute period). The rationale for this was to focus on the high-risk population of

patients with long periods of low SpO_2. In our study, we extended this analysis to the other two vital signs under consideration, HR and RR. To be able to determine whether a discrepancy between automatically- and manually-recorded vital signs exists, the difference between the mean vital-sign values acquired during the period ΔT_2 (i.e., $\bar{x}_{ij}$, with $j = 2$) and the charted vital sign x_i^c was computed.

Data were then partitioned according to their preceding values as follows:

- **SpO_2.** Data were split into two groups: observations that were preceded by a 10-minute period in which the mean SpO_2 was 90% or less (i.e., periods in which $\bar{x}_{i,j} \leq 90\%$, with $j = 1$), and observations that were preceded by a 10-minute period in which the mean SpO_2 was over 90% (i.e., periods in which $\bar{x}_{i,j} > 90\%$, again with $j = 1$).

- **RR.** Data were split, similarly, into three groups according to the mean values over the period ΔT_1: 12 rpm or less, 20 rpm or more, and the normal range of 12 to 20 rpm.

- **HR.** Data were also split into three groups according to the mean value over ΔT_1: 50 bpm or less, 110 bpm or more, and the normal range of 50 to 110 bpm.

These ranges of values were selected according to standard definitions of what is considered to be the "normal range" of each vital sign (as discussed in Taenzer et al. [2014]).

4.1.2 Results

Data for the partitions described above are presented in Table 4.1, displaying the average (μ) and standard deviation (σ) values over all observations considered for each vital sign, HR, RR and SpO_2, such that $\mu_j = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_{i,j}$, and $\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x}_{i,j} - \mu_j)^2}$, where $j = \{1, 2\}$, and N corresponds to the number of observations considered. In addition, to provide a measure of the *variability* within each sampling interval (δ), we determined the average of the standard deviation values σ_{ij} such that $\delta_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_{i,j}$. We observe that the variation (δ) of HR, RR and SpO_2 is approximately 3 bpm, 2 rpm and 1%, respectively, over

Table 4.1: Data for sampling intervals ΔT_1 and ΔT_2, displaying the number of observations in each vital-sign range (N), the mean (μ), the standard deviation (σ), and averaged variability (δ).

	Heart rate in ΔT_1			Respiratory rate in ΔT_1			SpO_2 in ΔT_1	
	≤50	50−110	≥110	≤12	12−22	≥22	≤90	90−100
N	104	4996	401	1179	3581	589	109	5633
$\mathbf{\Delta T_1}$								
μ (SD)	47 (3)	80 (13)	121 (12)	10 (2)	16 (2)	26 (4)	87 (4)	97 (2)
δ	3	3	5	1	2	3	2	1
$\mathbf{\Delta T_2}$								
μ (SD)	49 (9)	81 (14)	120 (14)	12 (3)	17 (4)	25 (5)	92 (5)	98 (3)
δ	2	3	4	2	2	3	1	1
Chart								
μ (SD)	53 (10)	82 (13)	117 (16)	14 (3)	17 (3)	24 (5)	95 (5)	98 (2)

both sampling periods. That is, no substantial differences between the variability of each channel in the two intervals were found.

Figure 4.3 shows the difference between the automatically-acquired vital-sign measurements at the time of the nurse visit and the corresponding vital sign values recorded by the nurse on the paper chart ($\bar{x}_{i,2} - x_i^c$). It may be seen that the manually-charted HR, RR and SpO_2 values x_i^c are significantly different from the mean of the automated values $\bar{x}_{i,2}$ ($p < 0.01$, for all groups of values, using a Mann-Whitney U test with the null hypothesis that the two groups have similar values). For example, the SpO_2 values recorded manually were higher than the mean SpO_2 values recorded by the continuous monitors during ΔT_2 by an average of approximately 8% in patients with prolonged desaturations. Large differences may likewise be observed for low and high values of HR and RR.

SpO_2 values from ΔT_1 and ΔT_2 were similar; with a median (and interquartile range, IQR) difference between the mean value for ΔT_2 and ΔT_1 of 1% (IQR -0.3−2.1%). For HR and RR, the median differences between the two sampling periods were 2 bpm (IQR -1−4 bpm) and 1 rpm (IQR -1−3 rpm), respectively. Although these differences are statistically significant ($p < 0.01$), there appears to be no physiologically-relevant systematic difference between the values of SpO_2,

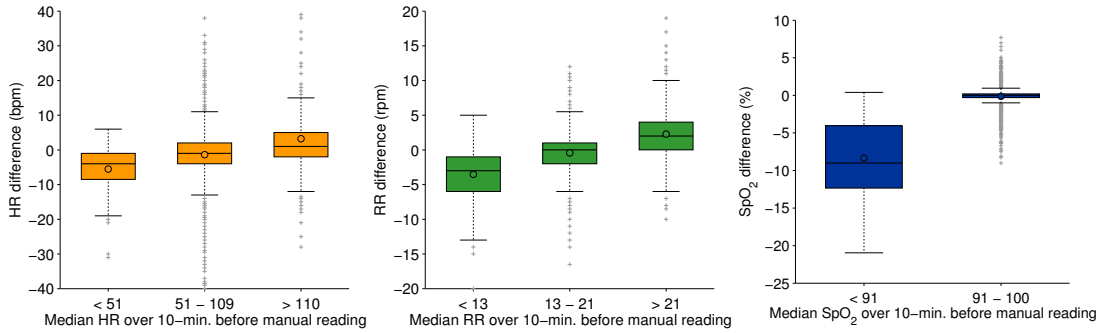


Figure 4.3: Boxplots representing the difference between the automatically-acquired vital-sign measurements at the time of the nurse visit and the corresponding charted vital sign values. From left to right, the differences are shown for HR, RR, and SpO_2. For each vital sign, the differences are computed for the different ranges defined according to data from the 10-minute intervals that precedes the manual observation (ΔT_1). The 25^{th}, 50^{th} and 75^{th} quantile values are displayed as lower, middle and upper horizontal lines of the boxes, respectively. Whiskers are used to represent the most extreme values within 1.5 times the interquartile range from the central box. Outliers (data with values beyond the ends of the whiskers) are displayed as crosses. Circles represent the mean values.

HR and RR during the two time frames.

4.1.3 Discussion

Fundamental differences between the vital-sign values charted by the nurses and those recorded by the bedside monitors at approximately the same time (nurse observation) are evident. Firstly, our findings support the results obtained in the recently published study by Taenzer et al. [2014]. SpO_2 values obtained by intermittent manual charting were biassed upwards, and they did not reflect the values recorded by bedside monitors, for observations that were preceded by prolonged desaturations. A mean difference of 8% above the actual physiological value (as given by the bedside monitor) was obtained in our study, which is similar to that reported by Taenzer et al. [2014] (6.5%). We also considered HR and RR. Values of these vital signs obtained by intermittent manual observations were also biassed upwards for observations that were preceded by periods of low HR and low RR. Mean differences of 7 bpm and 4 rpm above the actual physiological values (as given by the bedside monitors) were obtained, respectively, for these

vital signs. Conversely, for observations that were preceded by 10-minute periods of high HR or high RR, the charted values were 3 bpm and 3 rpm (in average) *below* the values acquired with the bedside monitors. As also observed in Taenzer et al. [2014], there were no clear arousal (or "wake-up") effects created by the nurse visit itself that were reflected in the results obtained in this analysis. That is, the presence of the nurse, during ΔT_2, did not "produce" substantial changes in the vital signs of the patients.

It may be argued that the differences we have observed may be due to artefacts or inaccurate measurements made by patient monitors. For this analysis, we only considered those data acquired with bedside monitors that were connected to patients during their first two days after surgery. That is, during this post-surgical period, patients were generally confined to a hospital bed after a major operation; this suggests that the presence of significant motion artefacts is unlikely.

Several other studies have questioned the accuracy of obtaining vital-sign data manually. In a previous study comparing electronically-acquired vital-sign measurements with those charted by nurses, a discrepancy greater than 20 mmHg for systolic BP was found for 7.5% of patients, and a discrepancy of over 6% for SpO_2 was found for 1% of patients (Sapo et al. [2009]); however, most of the patients in the study had values in the normal range. Sneed and Hollerbach [1992] reported that in patients with atrial fibrillation, 86% of nurses underestimated the HR. RR measurements have also been found to vary significantly between *counting periods* (Bianchi et al. [2013]; Lovett et al. [2005]). The current recommended method for measuring RR is the manual observation of chest-wall movements for a counting period of 60 seconds. Yet, in clinical practice, nursing staff count breaths for 15 or 30 seconds, and then scale up the number (i.e., multiply the result by 4 or 2, respectively) to obtain an estimate of RR in breaths per minute (rpm). Several studies have demonstrated the inaccuracy of this practice, such as Lovett et al. [2005], who compared respiratory rates recorded by triage nurses in Emergency Departments (counting chest wall movements for 15 seconds) to the gold standard (counting chest wall movements for 60 seconds) and found low correlation between the two methods. Perhaps more relevant to our study, Bianchi et al. [2013] found that 77% of tachypnoeic (rapid breathing) cases were missed in the Emergency Department when chest wall movements were only counted for 15

seconds. Respiratory rates derived from this method also did not reflect the wide spread of respiratory rates when compared to counting chest-wall movements for 60 seconds.

As with the work of Taenzer et al. [2014], our study has some limitations. The number of observations which were preceded by periods of "abnormal" physiology is small (Table 4.1). Also, it is important to note that the time at which manual charting occurred may be inaccurate with respect to the timestamps of data from the bedside monitor. However, the accuracy of the charting times was verified by checking and comparing with the time at which the corresponding blood pressure measurement (from the data associated with the bedside monitors) was performed. Bedside monitors are programmed to take periodic blood pressure measurements every 30 or 60 minutes, and these measurements are synchronised with the automatic readings of the other vital signs collected by the bedside monitors. During nurse observations, a blood pressure reading is always taken, and this reading can thus be used to confirm the time of the charted observation. Observations for which the difference between the time recorded on the chart and the time of the blood pressure measurement from the bedside monitor was higher than 5 minutes were removed from this analysis.

In conclusion, there are some important differences between the values of continuously-acquired vital signs and those that are collected manually. For RR, in particular, the difference appears to be physiologically-significant given the chest-wall motion method used to estimate this physiological variable and the smaller range of physiologically-plausible values for this variable in comparison with other variables (such as HR). This is an important consideration that needs to be taken into account when designing an early warning scoring system, as demonstrated in the next section.

4.2 Implications on the design of an early-warning scoring system

Recently proposed EWS-based systems rely on data-driven approaches to derive the cutoff values for the scores assigned to each physiological variable (Badriyah

et al. [2014]; Prytherch et al. [2010]; Tarassenko et al. [2011]), as discussed in chapter 3. While the systems proposed by Prytherch et al. [2010] and Badriyah et al. [2014] are based on a large database of vital signs *collected manually*, Tarassenko et al. [2011] used a large dataset of *continuously-recorded* vital-sign data. These data were used to investigate the statistical distributions of each vital sign and define an aggregate centile-based alerting system with seven bands of high and low activity for each vital sign (i.e., scores of 3, 2, 1, 0, 1, 2 and 3, as used in other systems). Observations were then treated as being abnormal if they occur at the extremes of the distributions of the vital signs; e.g., a score of 3 corresponds to vital signs above the 99^{th} quantile, a score of 2 corresponds to the vital sign being between the 95^{th} and 99^{th} quantiles, and a score of 1 corresponds to vital signs being between the 90^{th} and 95^{th} quantiles. Although the average values for each vital sign were remarkably similar to previously-published values for hospitalised patients, the threshold values in CEWS are different to those in many other EWS systems, and differ most for the upper limits of RR (Tarassenko et al. [2011]).

The cutoff values for the upper limits of RR in CEWS are 25, 28, and 33 (for a score of 1, 2, and 3, respectively), as shown in Table 4.2. These are substantially different from the values considered in other EWS systems (see Appendix A). We have just seen that the reason for this may be the fact that in the analysis underpinning the CEWS system, RR data were collected automatically, whereas for every other system, *and for vital-sign observations on the ward*, RR is estimated manually. While HR, SpO_2, temperature and BP are measured using standard methods (normally using a pulse oximeter, temperature probe and blood pressure cuff, respectively), RR is measured by clinical staff using the chest-wall movement counting strategy, as described earlier, whereas the RR data used to build the CEWS model were derived from continuous impedance pneumography measurements. However, the primary use of EWS systems is based on their application to manual measurements, and so, in the next section we investigate the cutoff values for RR derived from observational data (counting of chest-wall movements).

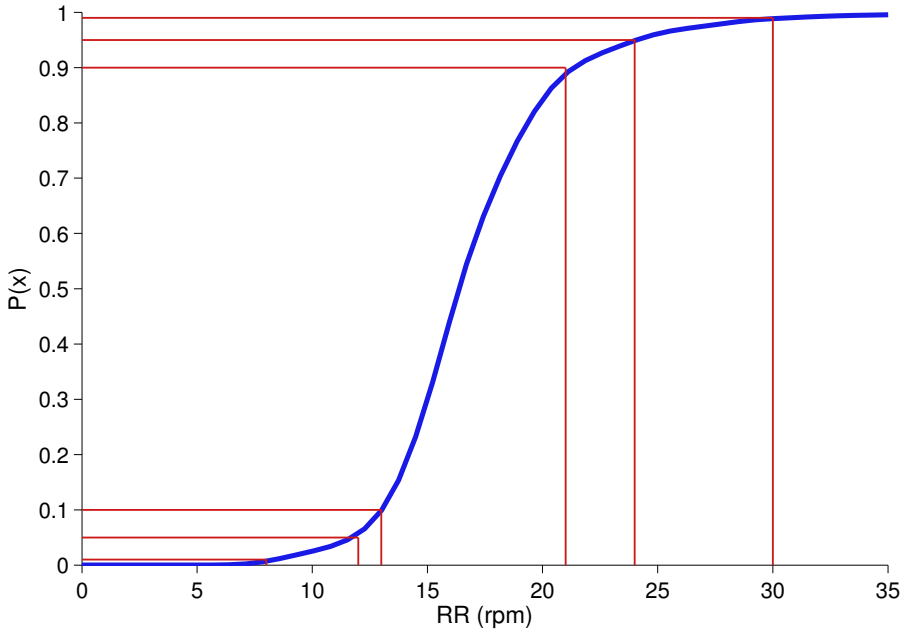


Figure 4.4: Cumulative distribution function $P(x)$ for RR (shown in number of breaths per minute, rpm), computed from the observational data from a subset of the patient-episodes included in the Portsmouth dataset. The 1^{st}, 5^{th}, 10^{th}, 90^{th}, 95^{th}, and 99^{th} quantiles are shown on the vertical axis and the corresponding cutoff values on the horizontal axis (red lines).

4.2.1 Methodology

Using the database of observation sets (which includes manual measurements of RR), and using the same approach as Tarassenko et al. [2011], we can identify new limits for the lower and upper thresholds for RR (see Figure 4.4). To maximise the use of the dataset for robust estimation of the quantiles in the tails of the distribution, a smoothed distribution of RR was obtained using a kernel-based density estimator. The bandwidth of the Gaussian kernels used, h, was computed using the *normal distribution approximation*, given by $h = 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of the n samples (Chakravarty et al. [1967]). Using this distribution of RR computed from training data, the lower cutoff values can then be set to the integer values that correspond to the 10^{th}, 5^{th} and 1^{st} quantiles, and the upper cutoff values can be set to be the integer values that correspond to the 90^{th}, 95^{th} and 99^{th} quantiles (as shown in Figure 4.4), for a score of 1, 2, and 3,

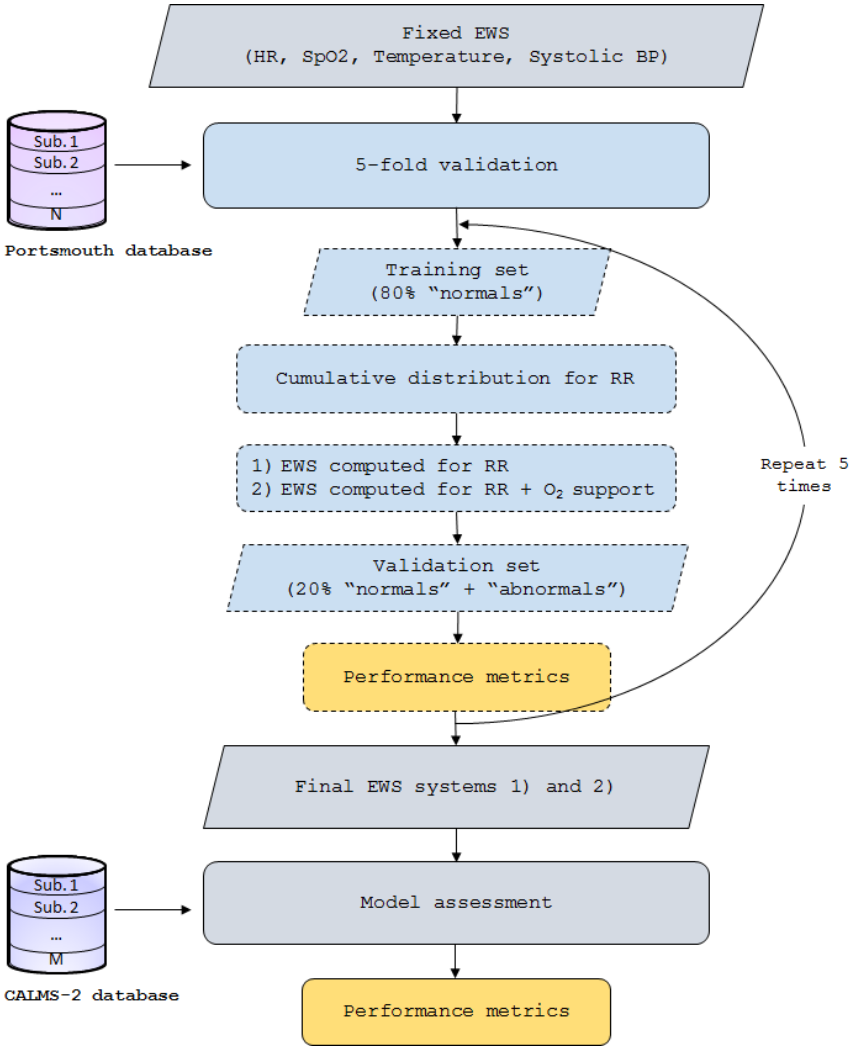


Figure 4.5: Flowchart of model development process for the two modified CEWS systems. Dashed lines indicate steps that are repeated 5 times (i.e., they are included in the 5-fold validation process using the Portsmouth dataset).

respectively.

The performance of this approach was evaluated using the procedure illustrated in Figure 4.5. The cutoff values for HR, SpO_2, systolic BP, and temperature, were assumed to be the as same as those in the original CEWS system. We used a five-fold cross-validation procedure using the Portsmouth database for deriving the new cutoff values for RR. Data partitioning was performed on a patient-by-patient basis (see Figure 4.6); i.e., by including all the data from

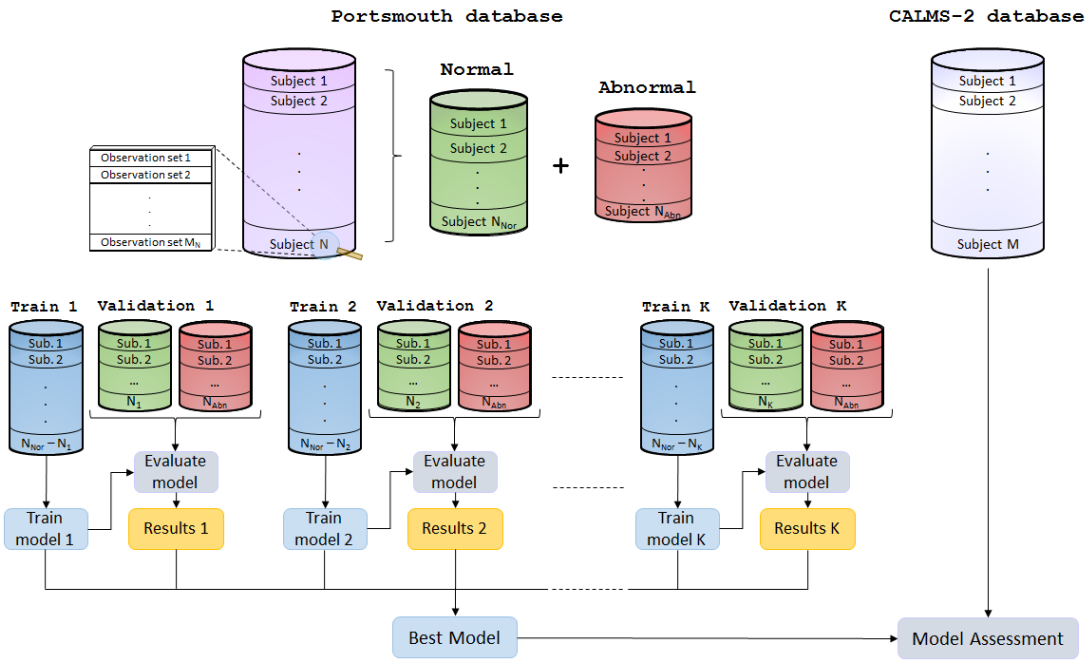


Figure 4.6: Schematic representation of the data partitioning method used for designing and assessing the performance of the models studied. A K-fold cross-validation using the Portsmouth dataset was employed to design the proposed scoring system, and the entire CALMS-2 dataset was used to assess the models' performance. Note that the training set contained only data from "normal" patients; only the validation sets (and test set) contained "abnormal" samples.

one patient *either* in the training or the validation set, in order to avoid the presence of observation sets from the same patient appearing in both training and validation sets. We note that training sets contain only data from "normal" patients. Data from $K-1$ subsets of "normal" patients (80%) were used for designing (training) the proposed scoring system, and the remaining subset of "normal" patients (20%) and the set of "abnormal" patients were used for validation. During "training", new cutoff values for the lower and upper bands of RR were determined from the observational data included in the training set. The resulting system (with the new cutoff values for RR and "fixed" cutoff values for the remaining four vital signs) is termed the *modified CEWS* system. Also, taking into account the importance of the supplemental use of oxygen (as noted in the previous chapter), we modelled the effect of adding an additional score of 2 for this variable, as in the system proposed by the RCP [2012] (this resulted

in a second modified CEWS system, in which the only difference from the previous one is this additional feature). During validation, data for each fold of the cross-validation were used to evaluate the performance of the EWS systems for the derived outcome of death within 24 hours of an observation set (as in the previous chapter), and then used to select the EWS threshold that gave the best performance for the model from each fold (the "best" threshold was selected using ROC curve analysis, as previously defined in the chapter 3).

The selected modified CEWS systems (based on the best performance obtained on the validation set) were then assessed using the CALMS-2 dataset for the identification of a major adverse event (composite outcome of death, cardiac arrest and unanticipated ICU admission) 24-hours post-observation set.

Data pre-processing was carried out prior to the evaluation of these systems as described in section 3.3.1 (by removing artefactual data and data outside the range of physiologically-plausible values). The performance of the proposed systems were compared with that of the twenty-six track-and-trigger systems previously presented in Table 3.3 (page 64).

4.2.2 Results

The cutoff values obtained for each of the five folds during the training/validation phase were very similar; i.e., there were no major differences between the integer cutoff values for RR obtained on the five folds. The mean (standard deviation, SD) values obtained for each cutoff value were 8 (0), 12 (0), 13 (0), 21 (0), 24 (0), and 30 (0), for a score of 3, 2, 1, 1, 2, and 3, respectively. Table 4.2 highlights the differences between the original CEWS system and the modified version of the CEWS system proposed in this study (two modified CEWS systems were tested: one without the supplemental oxygen feature, and one with this additional feature).

Table 4.3[1] summarises the performance measures computed for the proposed systems and the other 26 scoring systems on the population of patient-episodes in the Portsmouth dataset. We observe that, compared with the best perform-

[1] As in the previous chapter, for ease of reading, the standard error associated with the mean values displayed in the table are omitted. The complete version of the table can be found in Appendix A.

Table 4.2: CEWS (EWS system [24]) and modified CEWS. Differences between the systems are highlighted in red.

Centile-based early warning score							
	Score						
Variable	3	2	1	0	1	2	3
Heart Rate	≤ 42	$43 - 49$	$50 - 53$	$54 - 104$	$105 - 112$	$113 - 127$	≥ 128
Resp. Rate	≤ 7	$8 - 10$	$11 - 13$	$14 - 25$	$26 - 28$	$29 - 33$	≥ 34
Temperature	≤ 35.4		$35.5 - 35.9$	$36.0 - 37.3$	$37.4 - 38.3$		≥ 38.4
Systolic BP	≤ 85	$86 - 96$	$97 - 101$	$102 - 154$	$155 - 164$	$165 - 184$	≥ 185
SpO_2	≤ 84	$85 - 90$	$91 - 93$	≥ 94			
Inspired O_2							
AVPU scale				A	V		P, U

Modified centile-based early warning score							
	Score						
Variable	3	2	1	0	1	2	3
Heart Rate	≤ 42	$43 - 49$	$50 - 53$	$54 - 104$	$105 - 112$	$113 - 127$	≥ 128
Resp. Rate	≤ 8	$9 - 11$	$12 - 13$	$14 - 20$	$21 - 23$	$24 - 29$	≥ 30
Temperature	≤ 35.4		$35.5 - 35.9$	$36.0 - 37.3$	$37.4 - 38.3$		≥ 38.4
Systolic BP	≤ 85	$86 - 96$	$97 - 101$	$102 - 154$	$155 - 164$	$165 - 184$	≥ 185
SpO_2	≤ 84	$85 - 90$	$91 - 93$	≥ 94			
Inspired O_2				Air			Any O_2
AVPU scale				A	V		P, U

ing system [26], proposed by Badriyah et al. [2014] (which was developed using the Portsmouth vital-sign database), the original CEWS system [24][1] has a significantly lower AUROC of 0.796 (0.006). With the limits for RR derived from the quantiles of the distribution of the *manual* measurements of this variable rather than the (continuously-acquired) electronic measurements as in the original CEWS model, the modified CEWS system [24][2] now has an AUROC value of 0.817 (0.004). We note that among all EWS-based systems that do not take into account the use of oxygen support, the modified CEWS achieves a reasonable performance in terms of AUROC values in this database. When an extra score of +2 for the use of any oxygen support was also included, then the "modified-CEWS" system [24][3] has an AUROC of 0.856 (0.005), with a 95% CI close to those of the three best-performing EWS systems. A better performance was also achieved with system [24][3] in terms of all performance metrics other than AUROC.

The performance measures computed on the post-operative population are

Table 4.3: Performance metrics for the EWS systems studied, including the two modified versions of CEWS (system [24]), evaluated using the Portsmouth dataset. The results are presented in descending order of AUROC, and the best values for each performing metric underlined.

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[26]	0.881	0.827	0.780	0.834	0.037	0.146
[25]	0.877	0.854	0.841	0.760	0.028	0.125
[23]	0.877	0.862	0.793	0.815	0.034	0.139
[24]^3	0.856	0.855	0.731	0.835	0.030	0.133
[21]	0.854	0.777	0.721	0.826	0.033	0.128
[19]	0.836	0.728	0.772	0.758	0.025	0.110
[3]	0.835	0.712	0.771	0.771	0.027	0.115
[9]	0.827	0.752	0.780	0.719	0.022	0.099
[20]	0.827	0.744	0.780	0.728	0.023	0.102
[12]	0.826	0.680	0.703	0.816	0.030	0.119
[24]^2	0.817	0.790	0.684	0.812	0.023	0.117
[11]	0.815	0.799	0.835	0.656	0.019	0.092
[15]	0.814	0.798	0.835	0.657	0.019	0.092
[14]	0.814	0.798	0.835	0.657	0.019	0.092
[5]	0.812	0.764	0.806	0.695	0.021	0.097
[10]	0.811	0.748	0.765	0.715	0.021	0.095
[22]	0.811	0.739	0.766	0.723	0.022	0.097
[18]	0.810	0.829	0.647	0.838	0.032	0.117
[6]	0.809	0.779	0.818	0.677	0.020	0.094
[13]	0.809	0.778	0.813	0.679	0.020	0.094
[1]	0.809	0.779	0.818	0.677	0.020	0.094
[2]	0.809	0.778	0.818	0.677	0.020	0.094
[4]	0.809	0.785	0.821	0.668	0.020	0.093
[17]	0.808	0.776	0.814	0.680	0.020	0.094
[7]	0.808	0.775	0.814	0.680	0.020	0.094
[24]^1	0.796	0.746	0.750	0.711	0.021	0.090
[16]	0.780	0.749	0.751	0.685	0.019	0.084
[8]	0.779	0.668	0.640	0.815	0.027	0.104

[24]^1 Original CEWS system; [24]^2 CEWS system with modified limits for RR; [24]^3 CEWS system with modified limits for RR and with an additional score of 2 for any oxygen support.

Table 4.4: Performance metrics for the EWS systems studied, including the two modified versions of CEWS (system [24]), evaluated using the CALMS-2 dataset. Results are presented in descending order of AUROC. The best values for each performing metric are underlined.

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[23]	0.841	0.826	0.801	0.765	0.084	0.209
[24]^3	0.839	0.830	0.789	0.797	0.132	0.251
[25]	0.835	0.747	0.829	0.715	0.073	0.190
[21]	0.833	0.753	0.639	0.886	0.131	0.251
[26]	0.829	0.816	0.773	0.768	0.083	0.201
[24]^2	0.796	0.737	0.653	0.860	0.112	0.228
[22]	0.791	0.807	0.711	0.791	0.084	0.193
[10]	0.786	0.677	0.700	0.784	0.081	0.185
[12]	0.784	0.825	0.622	0.872	0.116	0.227
[9]	0.782	0.680	0.697	0.777	0.078	0.179
[4]	0.782	0.720	0.778	0.730	0.072	0.180
[20]	0.781	0.678	0.697	0.779	0.078	0.180
[5]	0.779	0.703	0.752	0.751	0.075	0.183
[6]	0.777	0.713	0.762	0.738	0.073	0.179
[17]	0.777	0.712	0.757	0.742	0.073	0.180
[1]	0.776	0.714	0.762	0.736	0.072	0.178
[7]	0.776	0.711	0.757	0.742	0.073	0.180
[2]	0.776	0.712	0.758	0.739	0.073	0.178
[8]	0.776	0.618	0.603	0.875	0.115	0.221
[13]	0.770	0.710	0.752	0.736	0.072	0.175
[18]	0.766	0.623	0.593	0.861	0.103	0.203
[19]	0.764	0.782	0.570	0.845	0.090	0.179
[15]	0.764	0.802	0.565	0.885	0.117	0.216
[11]	0.764	0.803	0.801	0.612	0.053	0.135
[14]	0.763	0.802	0.801	0.613	0.053	0.135
[24]^1	0.759	0.693	0.702	0.748	0.070	0.163
[16]	0.744	0.609	0.522	0.868	0.096	0.178
[3]	0.717	0.760	0.501	0.898	0.117	0.202

[24]^1 Original CEWS system; [24]^2 CEWS system with modified limits for RR; [24]^3 CEWS system with modified limits for RR and with an additional score of 2 for any oxygen support.

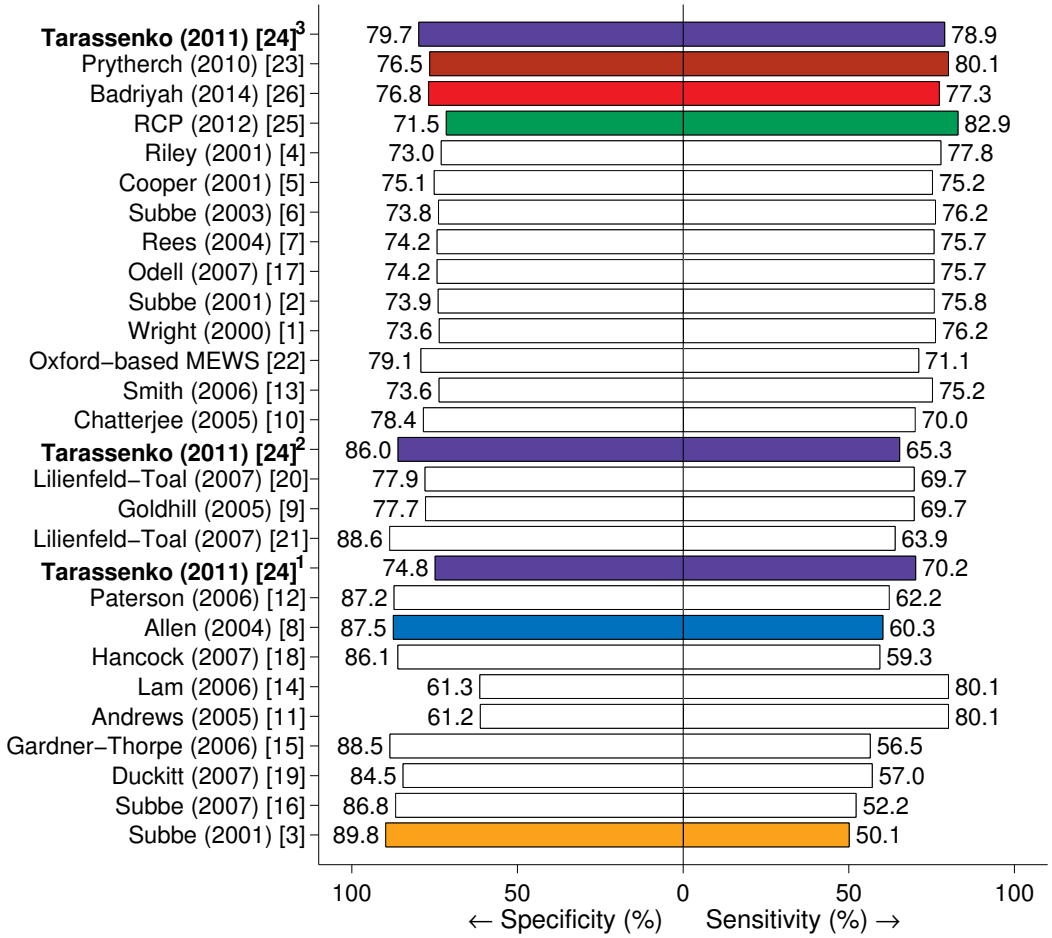


Figure 4.7: Performance of the EWS systems for the combined outcome of cardiac arrest, unanticipated ICU admission or death occurring within 24 hours of a given observation set tested with the CALMS-2 dataset. [24]1 Original CEWS system; [24]2 CEWS system with modified limits for RR; [24]3 CEWS system with modified limits for RR and with an additional score of 2 for any oxygen support.

shown in Table 4.4[1]. In Figure 4.7, results are represented as pairs of sensitivity and specificity values, which are displayed in ascending order of best pair of values in the figure. A substantial improvement in diagnostic performance in the CALMS-2 dataset is also obtained by the CEWS system with the modified RR limits and the additional score of +2 for use of oxygen support: the original CEWS system gave a sensitivity of 70.2% and a specificity of 74.8%, while system

[1]Refer to Appendix A for the complete version of the table.

[24][3] achieves a sensitivity of 78.9% and a specificity of 79.7%. We observe that the overall performance of the proposed modified CEWS system is comparable to that of the best-performing EWS systems.

4.2.3 Discussion

These results support the findings of the previous analysis, and demonstrate their relevance for both clinical practice and the design of data-driven scoring systems. We conclude that the method of recording physiological variables on the ward has a significant effect on scoring systems. With the increasing prevalence of automated methods to measure, acquire and record vital-sign data from hospitalised patients, the results presented in this chapter suggest that EWS-based systems will have to be adjusted accordingly when respiratory rate is no longer estimated manually.

In this study, we focused on the analysis of RR, as there are fundamental differences between the methods used to measure this variable manually and those used to measure it automatically. Apart from the upper limits of RR, no significant differences were observed between the cutoff values obtained for other vital signs using modified CEWS and those of the original CEWS system; i.e., the difference between CEWS and the other EWS-systems is predominantly in the upper limits considered for RR. For this reason, we focused on the effect on the performance metrics of this *small* modification to the CEWS system. It is important to take into account the fact that RR is still estimated manually on most post-operative wards, hence EWS systems should be designed for manual measurements of respiratory rate. There is also a clear effect associated with the inclusion of supplemental oxygen in the scores for post-operative patients (as already discussed in chapter 3).

These findings show that the design of future data-driven EWS systems should take into account not only the different sources of data used to build those systems, but also the different methods of recording the physiological variables *for which the systems will be used.*

4.3 Conclusion

In this study, we demonstrated that the method of recording physiological variables on the ward may play a fundamental role on both the design and performance of current early warning scoring systems. Conversely, in the era of electronic medical records and automated methods for acquiring vital-sign data, scoring systems will increasingly be based on automatically-collected data. When RR, for example, is eventually recorded electronically rather than by counting chest-wall movements, then the various EWS systems may have to be adjusted accordingly.

To conclude, we note that the methods employed in the most recent data-driven systems (including those proposed here) treat each variable independently and correlations between the different components of the system are not taken into account. Alternative strategies based on machine learning techniques can be adopted for multivariate data analysis, with for example, the inclusion of other variables that may significantly improve the performance of current early warning scoring systems. Such approaches are explored in the following chapters.

Chapter 5

Machine learning approach to patient monitoring

In the preceding chapters, we evaluated the performance of scoring systems that were designed using univariate analysis; i.e., they treat each variable independently and correlations between the different components of the system are not taken into account. Alternative strategies based on machine learning are explored in this chapter. Firstly, the key concepts of novelty detection, a "task" of machine learning methods, are introduced, and the theoretical framework behind novelty detection techniques is described. The performance of such techniques for identifying deterioration in hospitalised patients is then presented and discussed.

5.1 Novelty detection

Conventional pattern recognition and machine learning classification tasks typically focus on the classification of two or more classes. General multi-class classification problems are often decomposed into multiple two-class classification problems, where the two-class (or binary) case is considered the basic classification task (Barber [2012]; Sammut and Webb [2011]). For example, the prediction of mortality in a population of ICU patients for which data and outcomes have previously been collected is often formulated as being a two-class (binary) classification problem (Knaus et al. [1985]; Mayaud [2014]). This defines a set of N

training examples $\mathbf{X} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^D, i = 1, ..., N\}$, for which each example (which may correspond to a patient observation set) consists of a D-dimensional vector $\mathbf{x}_i$ (in which D corresponds to the number of variables considered) and its label $y_i \in \{0, 1\}$. From the labelled dataset, a function $h(\mathbf{x})$ is constructed such that, for a given input test vector $\mathbf{x}^*$, an estimate of one of the output labels is obtained, $y^* = h(\mathbf{x}^*|\mathbf{X})$:

$$h(\mathbf{x}^*|\mathbf{X}) : \mathbb{R}^D \rightarrow [0, 1] \tag{5.1}$$

In many practical situations, however, a representative set of labelled examples for either or both classes may be too costly or difficult to obtain. The scarcity of examples from a class corresponding to unusual events can also be due to the low frequency at which these abnormal events occur. For example, in a machine monitoring system, we require an alarm to be triggered whenever the machine exhibits "abnormal" behaviour. Measurements of the machine during its normal operational state are inexpensive and easy to obtain. Conversely, measurements recorded during failure of the machine would require the destruction of similar machines in all possible ways. The same rationale can be applied to human patients. Therefore, it is difficult, if not impossible, to obtain a very well-sampled "abnormal" class, or classes (He and Garcia [2009]; Lee and Cho [2006]). Furthermore, the complexity of modern high-integrity systems is such that only a limited understanding of the relationships between the (typically very large number of) system components can be obtained. An inevitable consequence of this is that there may exist a large number of possible modes of behaviour, some of which may not be known *a priori*, which makes conventional multi-class classification schemes unsuitable for such applications. This problem is often compounded when monitoring critical systems, such as human patients, for which there is significant variability between individuals, thereby limiting the ability of population-based models to generalise to the monitoring of previously-unseen individuals.

This often occurs in medical diagnosis, for which data from non-healthy patients are extremely hard to obtain. For example, detection of mass-like struc-

tures in mammograms for breast cancer identification (Tarassenko et al. [1995]), prediction of protein-protein interactions (Reyes and Gilbert [2007]), recognition of cognitive brain functions (Boehm et al. [2011]), lung tissue categorisation of patients affected by lung diseases (Depeursinge et al. [2010]), identification of patients with infections (Cohen et al. [2008]), and detection of physiological deterioration in hospitalised patients (Hann [2008]; Tarassenko et al. [2006]; Wong [2011]). A solution to this is offered by *novelty detection*.

5.1.1 Novelty detection as one-class classification

Novelty detection can be defined as the task of recognising that test data differ in some respect from those data that are available during training. Its practical importance and challenging nature have led to many approaches being proposed to address this problem. These methods are typically applied to datasets in which a very large number of examples of one class, often termed the "normal" class is available and where there are insufficient data to describe the other classes (which are often associated with "abnormalities"). Novelty detection has gained much research attention in application domains involving large datasets acquired from critical systems.

The problem of novelty detection can be tackled within the framework of one-class classification (Moya et al. [1993]), in which *one* class (the specified normal class) has to be distinguished from all other possibilities. It is usually assumed that the normal class is very well sampled, while the other classes are under-sampled. In condition monitoring applications, "normal" patterns $\mathbf{X}$ are available for training, while "abnormal" patterns are typically not used during training. A model of normality $H(\mathbf{X}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the free parameters of the model, is inferred and used to assign novelty scores $z(\mathbf{x})$ to previously unseen test data $\mathbf{x}$. Larger novelty scores $z(\mathbf{x})$ correspond to increased "abnormality" with respect to the model of normality. A novelty threshold $z(\mathbf{x}) = k$ is defined such that $\mathbf{x}$ is classified "normal" if $z(\mathbf{x}) \leq k$, or "abnormal" otherwise. Thus, $z(\mathbf{x}) = k$ defines a decision boundary. Different types of models H, methods for setting their parameters $\boldsymbol{\theta}$, and methods for determining novelty thresholds k have been proposed in the literature.

Two interchangeable synonyms of novelty detection (Bishop [1994]; Tarassenko et al. [1995]) often used in the literature are *anomaly detection* and *outlier detection* (Ritter and Gallegos [1997]). The different terms originate from the different domains of application for one-class classification, and there is no universally-accepted definition. Anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably (Chandola et al. [2009]). An "outlier" is a term that is typically used to describe a small fraction of "normal" data that lie far way from the majority of "normal" data in the data space (Markou and Singh [2006]). Therefore, outlier detection aims to handle those "rogue" observations in a set of data, that would otherwise have a substantial effect on the analysis of the data. Outliers are thus assumed to contaminate the dataset under consideration and the goal is to cope with their presence during the model-construction stage. A different goal is to learn a model of normality $H(\mathbf{X}|\boldsymbol{\theta})$ from a set of data that is considered "normal", in which the assumption is that the data used to train the learning system constitute the basis to build a model of normality and the decision process on test data is based on the use of this model. The notion of normal data as expressed in anomaly detection is often not the same as that used in novelty detection. Anomalies are often taken to refer to irregularities or transient events in otherwise "normal" data. These transient events are typically noisy events, which give rise to artefacts that act as obstacles to data analysis, to be removed before analysis can be performed. From this definition, novel data are not necessarily anomalies; this distinction has also been drawn by recent reviews in anomaly detection (Chandola et al. [2009]). Nevertheless, the term "anomaly detection" is typically used synonymously with "novelty detection", and the solutions and methods used in novelty detection, anomaly detection, and outlier detection are often the same.

5.1.2 Methods of novelty detection

Approaches to novelty detection include both frequentist and Bayesian approaches, involving information theory, extreme value statistics, support vector methods, other kernel methods, and neural networks. In general, all of these methods build some model $H(\mathbf{X}|\boldsymbol{\theta})$ of a training set $\mathbf{X}$ that is selected to contain no examples

(or very few) of the important (i.e., novel) class. Novelty scores $z(\mathbf{x}^*)$ are "assigned" to new data $\mathbf{x}^*$, and deviations from normality are detected according to a decision boundary, referred to as the novelty threshold $z(\mathbf{x}) = k$.

A number of surveys of novelty detection and outlier or anomaly detection methods have been published in the last decade (Agyemang et al. [2006]; Bakar et al. [2006]; Chandola et al. [2009]; Hodge and Austin [2004]; Khan and Madden [2010]; Markou and Singh [2003a,b]; Marsland [2003]). We have recently reviewed the different methods of novelty detection proposed in the literature (Pimentel et al. [2014]), in which we aimed to provide an updated and structured overview of recent studies, including their main domains of application. We classified the approach taken by these techniques according to the following five general categories: probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic techniques.

Probabilistic approach. This approach uses probabilistic methods that often involve a density estimation of the "normal" class, $p(\mathbf{X}|\boldsymbol{\theta})$. The assumption is that low-density areas in the training set indicate that these areas have a low probability of containing "normal" examples. Broadly, these methods fall into *parametric* and *nonparametric* approaches, in which the former impose a restrictive model on the data (i.e., the number of free parameters of the model $\boldsymbol{\theta}$ is finite), which can result in a large bias when the model does not fit the data well; conversely, the latter lead to more flexible models by making fewer assumptions: the parameter space is not limited to some finite value *a priori*, and may grow in size as data are observed. Examples of probabilistic approaches to novelty detection include component mixture models and kernel density estimates.

Distance-based approach. This may include the concepts of nearest-neighbour and clustering analysis that are typically used in conventional classification problems. These methods rely on well-defined distance metrics to compute the distance (as a similarity measure) between test data and the "normal" training data. The assumption here is that "normal" data are tightly clustered, while novel data occur far from the training set, in the data space.

Reconstruction-based approach. This involves training a regression model

using the training set. When "abnormal" data are mapped using the trained model, the "reconstruction error", defined to be the distance between the test point and the output of the model, gives rise to a high novelty score. Neural networks, for example, can be used in this way and can offer many of the same advantages for novelty detection as they do for regular classification problems. This category of methods includes different configurations of neural networks and principal component analysis.

Domain-based approach. This approach uses methods that are typically insensitive to the distribution of the "normal" class. A *domain* containing "normal" data is characterised by defining a boundary around the "normal", but does not provide an explicit estimate of the distribution. Classification of test data is then determined by their location with respect to the boundary. The support vector machine (SVM) is a popular technique for forming decision boundaries that separate data into classes and has been applied for novelty detection in two related approaches in the one-class setting. The one-class SVM approach (Schölkopf et al. [2000]) defines the novelty boundary by mapping the "normal" data into a feature space corresponding to a kernel (a Gaussian kernel is typically used) and separating the resulting projections from the origin with maximum margin. The *support vector data description* (Tax and Duin [1999]) defines the novelty boundary as being a hypershere with minimum volume that encloses all (or most) of the data in the "normal" class. Novelty is assessed by determining if a test point lies within the hypersphere.

Information-theoretic approach. This computes the information content in the training data using information-theoretic measures, such as entropy or Kolmogorov complexity. This is based on the hypothesis that novel data significantly alter the information content in a dataset, whereas examples of normal data (being similar to the training data) do not.

Novelty detection is well-suited to the problem addressed in this thesis, in which only a few patients (fewer than 10%) suffer a major adverse event, while the majority of patients undergo a "normal" recovery during the course of their stay on the ward. Novelty detection is an important learning paradigm and

has drawn significant attention within the research community, as shown by the increasing number of publications in this field (Pimentel et al. [2014]).

We observe that a precise definition of novelty detection is difficult to achieve, nor is it possible to suggest an "optimal" method or learning scheme for novelty detection. The variety of methods employed is a consequence of the wide variety of practical and theoretical considerations that arise from novelty detection in real-world datasets; many of these methods have data-specific parameters due to factors such as availability of training data, the type of data (including its dimension, continuity, and format), and application domain investigated. It is perhaps because of this great variety of considerations that there is no single universally applicable novelty detection method, and therefore, data knowledge is necessary.

5.2 Modelling for patient monitoring

There exists a growing body of literature that has focused on investigating machine learning methods for patient monitoring data (Lehman et al. [2013]; Quinn et al. [2009]; Saria et al. [2010, 2011])[1]. Such approaches have typically been applied to continuous time-series data collected from bedside monitors in the ICU. Outside the ICU, however, data are recorded manually at different time intervals, and it is unclear how such approaches can cope with irregularly-sampled data and missing data. As opposed to equally-spaced time-series, in which these methods have been applied, irregularly-sampled time-series data are characterised by variable intervals between successive measurements; i.e., the spacing of observation times is not constant. Recordings from different patients typically contain different numbers of observations and the times at which different observations within the observation set were recorded may not be aligned.

While several scoring systems exist for patient monitoring outside critical care and ICUs, very few are based on machine learning algorithms. Tarassenko et al. [2006] proposed an i.i.d. approach to patient monitoring based on novelty detection, in which a multivariate, multimodal model of the distribution of vital-sign

[1]These methods are further described in chapter 7.

data from a population of "normal" high-risk patients was constructed. Multivariate test data can then be compared with this model to give a novelty score, and an alert is generated when the score exceeds the novelty threshold. This approach has been further described by Hann [2008], and explored and extended in other studies (Clifton et al. [2011a,c]; Wong [2011]). Given its status as one of the few machine learning systems that is used in clinical practice (and the fact that this system has undergone regulatory approval and gained FDA approval), the method is described in this section as the "baseline" machine learning model in this thesis. Its extensions and derived approaches are also briefly outlined.

5.2.1 The dataset

The original model was trained using 3,500 hours of continuous vital-sign data collected from 150 high-risk patients at the John Radcliffe Hospital, Oxford, between 2001 and 2003 as part of an observational study (described in Hann [2008]). The patient population included patients who had severe heart failure, acute respiratory problems (such as acute asthma, pneumonia, or pulmonary embolism), trauma injuries, and those who were being continuously monitored following myocardial infarction. The physiological variables included in the model were HR, RR, SpO_2, systolic and diastolic BP, and temperature, which were acquired from bedside patient monitors. HR, RR, SpO_2 and temperature were sampled at a frequency of approximately 1 Hz, while BP values were measured at 30-minute intervals during the day, and at hourly intervals during the night (when the patient was asleep).

The different data channels (corresponding to the measurements of the physiological variables) were recorded asynchronously. In total, 2.6×10^5 sets of vital signs, with each set having six elements (one per physiological variable), were produced by aligning the different channels and re-sampling at 5-second intervals. Data were then pre-processed (using the procedure described in section 3.3.1), and the number of sets was thereby reduced to 2.4×10^5.

The model proposed by Tarassenko et al. [2006] assumes that each vital sign has equal importance, and should therefore have an equal weighting in the model. Hence, the systolic and diastolic BP values are combined into one parameter, the

systolic-diastolic average, by calculating their arithmetic mean. Also, each variable was scaled to have approximately the same dynamic range to ensure that variables with large changes (e.g., heart rate in beats per minute) do not dominate variables with smaller changes (e.g., temperature in °C). Each vital-sign measurement, x, is then normalised using the zero-mean unit-variance transformation, so that $x' = \frac{x-\mu}{\sigma}$, where μ and σ denote the mean and the standard deviation, respectively, computed from the training data points for the given vital sign.

5.2.2 Kernel density estimates

Kernel density estimate, KDE (a special case of which is referred to as "Parzen windowing"), allow the underlying D-dimensional data distribution to be estimated. It is a generalisation of the histogram technique, in which smoother functions are used instead of the rectangular volumes typically used in histogram binning (Parzen [1962]).

Define a finite dataset consisting of N data samples $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}$, where the feature vector variable $\mathbf{x}_i \in \mathbb{R}^D$ follows an unknown probability density function (pdf) $p(\mathbf{x})$. A general kernel-based estimate of $p(\mathbf{X})$ is given by

$$\hat{p}(\mathbf{x}|\mathbf{g}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} g_i K(\mathbf{x}, \mathbf{x}_i|\sigma_i) \tag{5.2}$$

where g_i are the kernel weights, subject to $g_i \geq 0$ for $i = 1, ..., N$ and $\mathbf{g} = [g_1, ..., g_N]$ with $\mathbf{g}^T\mathbf{1} = 1$, in which $\mathbf{1}$ is a vector of elements 1 with an appropriate dimension; $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_N]$ is the kernel "bandwidth" vector; $K(\mathbf{x}, \mathbf{x}_i|\sigma_i)$ is a kernel function with parameter σ_i. In Parzen windows, a Gaussian kernel denoted by

$$K(\mathbf{x}, \mathbf{x}_i|\sigma_i) = \frac{1}{(2\pi\sigma_i^2)^{D/2}} \exp\left(-\frac{||\mathbf{x} - \mathbf{x_i}||^2}{2\sigma_i^2}\right) \tag{5.3}$$

is typically used. Let the well-known Parzen window estimator be denoted by $\hat{p}(\mathbf{x}|\mathbf{g}, \boldsymbol{\sigma})$. In conventional Parzen window methods, $\mathbf{g} = [\mathrm{g}_1, ..., \mathrm{g}_N]$ and

$g_i = 1/N, \forall i$ (that is, all kernels have the same weight). Also, most approaches proposed in the literature consider an isotropic kernel (with a single bandwidth for all D dimensions of the data); i.e., $\sigma_i = \sigma, \forall i$. Combining equations (5.2) and (5.3), the estimated pdf $\hat{p}(\mathbf{x}|\sigma)$ is given by

$$\hat{p}(\mathbf{x}|\sigma) = \frac{1}{N(2\pi\sigma^2)^{D/2}} \sum_{i=1}^{N} \exp\left(-\frac{||\mathbf{x} - \mathbf{x}_i||^2}{2\sigma^2}\right) \tag{5.4}$$

The kernel parameter σ has the effect of controlling the level of detail in the pdf. A large value of σ leads to a distribution that is very smooth and which may not capture the details of the density, while a small value causes the pdf to be over-fitted to the data. Different methods for estimating σ have been proposed in the literature (see Appendix B).

5.2.3 Sparse kernel density estimates

A disadvantage of the Parzen windows method is its high computational cost for determining the likelihood of a point $\mathbf{x}$ when the training set is very large. By using a smaller number of mixture components, the finite-mixture model can be regarded as being a condensed representation of the data (Hong et al. [2008]). A commonly-employed preprocessing step is to partition the training dataset into n disjoint subsets (where $n \ll N$) using the K-means algorithm (Bishop [2006]). The resulting centroids of each cluster $\mathbf{s}_j$, with $j = 1, ..., n$, can then be used to estimate the density of the data, such that

$$\hat{p}(\mathbf{x}|\sigma) \approx \hat{p}(\mathbf{s}|\sigma) = \frac{1}{n(2\pi\sigma^2)^{D/2}} \sum_{i=1}^{n} \exp\left(-\frac{||\mathbf{s} - \mathbf{s}_i||^2}{2\sigma^2}\right) \tag{5.5}$$

From the 2.4×10^5 feature vectors in the training set, a subset of 500 cluster centres were then selected using the K-means algorithm. The 100 cluster centres furthest from the centroid of the 500 centres (using Euclidean distance) were discarded, thereby retaining the most "normal" cluster centres.

The kernel parameter σ was set using a heuristic suggested by Bishop [1994],

who calculates the mean of the local estimates of the variance at each data point location, as follows:

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j\in Q_i}||\mathbf{s}_i - \mathbf{s}_j||\right) \tag{5.6}$$

where Q_i is the set of m nearest-neighbours to point $\mathbf{s}_i$. The number of nearest-neighbours m was set to 10. A different approach for choosing the value of σ is to maximise the leave-one-out likelihood $J(\sigma)$ of the training data,

$$J(\sigma) = \frac{1}{n}\sum_{j=1}^{n}\log\left(\frac{1}{n-1}\sum_{i=1,i\neq j}^{n}K(\mathbf{s}_j, \mathbf{s}_i|\sigma)\right) \tag{5.7}$$

where $K(\cdot)$ is the Gaussian kernel introduced in Eq. (5.3). Among the different techniques proposed in the literature to compute σ, the maximum leave-one-out likelihood approach has provided a simple and reliable way of determining the kernel parameter (see Appendix B).

5.2.4 Patient Status Index

In order to estimate the "abnormality" of a test data point $\mathbf{x}$, the departure from normality is quantified using a *novelty score* $z(\mathbf{x})$ defined as follows:

$$z(\mathbf{x}) = \log\left(\frac{1}{p(\mathbf{x}|\boldsymbol{\theta})}\right) - \log\left(\frac{1}{p_{max}}\right) = \log\left(\frac{p_{max}}{p(\mathbf{x}|\boldsymbol{\theta})}\right) \tag{5.8}$$

where $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood of point $\mathbf{x}$, $\boldsymbol{\theta} = \{\mathbf{s}_i, \sigma\}$ with $i = 1, ..., n$, and p_{max} is the maximum possible value of p, where the subtraction is to adjust the scale so that the novelty score is close to zero when all vital signs are normal (this value can be determined using a gradient descent method). "Normal" data, which have higher likelihoods $p(\mathbf{x}|\boldsymbol{\theta})$, therefore generate low novelty scores $z(\mathbf{x})$; conversely, "abnormal" data generate high novelty scores.

It is important to consider that the original model proposed in Tarassenko et al. [2006] was designed to be applied to continuously-acquired data from bedside monitors. The model assumes that all physiological variables are available so that a novelty score can be computed. However, with most vital-sign monitors, each channel of data is treated independently and hence data are received asynchronously. To deal with this, values of all physiological variables are held for some duration, and a new novelty score is calculated each time new values from a single channel become available.

In practice, vital-sign data may be unavailable for extended periods of time due to disconnection of the sensors; this often occurs, for example, because electrodes become detached, or because patients remove the pulse oximeter finger probe. Also, medical data are specific in the sense that the presence (and thereby absence) of a measurement can reflect the clinician's decision to collect (or not) a certain vital sign. This is particularly true on general wards where data are said to be *not missing at random*. There are different missing data mechanisms: missing at random, in which a missing value of a variable does not depend on the value of that variable, and missing not at random, in which the probability of a missing value depends on the variable that is missing. Imputing a missing value with the mean of the variable only gives half of the story: this measurement may be non-informative. A different strategy to account for the non-randomness of missing values consists of adding a novel binary covariate indicating whenever the value is missing. However, this strategy potentially doubles the amount of covariates, which can be handled by a feature selection technique. In the system described above, data are typically assumed to be missing at random. A simple heuristic is used to deal with this situation: if a physiological variable value is missing over a one-minute period, the median value of that physiological variable over the last five minutes is substituted (this heuristic is used for all physiological variables, except for blood pressure, which is sampled less frequently). If a variable is missing for over 30 minutes, the mean value in the training dataset is used instead. The short-term median filter and population mean heuristics are only deemed valid in the cases where up to two physiological variables are missing. In the case of any further data drop-out, no novelty score is calculated.

5.2.5 Extensions and other approaches

Some extensions to this approach have been proposed to overcome some of its shortcomings. Also, other approaches for patient monitoring have been explored.

Sparse-weighted kernel density estimates. Wong [2011] noted that the subset of 400 prototype centres extracted from the full training dataset might not be the optimal representation of the normal data. This is because the K-means clustering algorithm may produce clusters with unequal population size. When Parzen windows is then used with the cluster centres, all clusters are given equal weight in the pdf estimate independent of the number of points in each cluster (i.e., the same weight g). This argument holds true as long as there are differences in the cluster populations, and the effect depends on how different the cluster sizes are. It was demonstrated that different number of data points were assigned to each one of the 400 clusters centres used to build the model (Wong [2011]). Sparse-weighted kernel density estimates were explored as a solution to this problem. This method yields an estimate

$$\hat{p}(\mathbf{x}|\sigma) \approx \hat{p}(\mathbf{s}|\sigma) = \frac{1}{(2\pi\sigma^2)^{D/2}} \sum_{i=1}^{n} w_i \exp\left(-\frac{||\mathbf{s}-\mathbf{s}_i||^2}{2\sigma^2}\right) \tag{5.9}$$

where $w_i = k_i/N$ corresponds to the weight of each centroid, and where k_i corresponds to the relative size of each cluster (i.e., k_i is the number of data points assigned to cluster i, with $i = 1, ..., n$ in the K-means algorithm).

Gaussian mixture models. Mixture of Gaussians has also been considered for patient monitoring (Clifton et al. [2014]). These methods assume that data are distributed according to a mixture of M multivariate Gaussian distributions,

$$\hat{p}(\mathbf{x}|\vartheta) = \sum_{i=1}^{M} \pi_i \; g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5.10}$$

where π_i, with $i = 1, ..., M$, are the mixture weights, and $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, ..., M$ are the component Gaussian densities. Each component density is a D-variate

Gaussian function of the form,

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{5.11}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ correspond to the centre and covariance matrix for the multivariate Gaussian i, respectively. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} \pi_i = 1$. Hence, $\vartheta = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, with $i = 1, ..., M$. There are several techniques available for estimating the parameters of a GMM (Bishop [2006]). By far the most popular and well-established method is expectation-maximisation (Dempster et al. [1977]) which yields maximum likelihood estimates.

One-class support vector machines. SVM-based approaches have also been considered to perform novelty detection for patient monitoring (Clifton et al. [2014]; Wong [2011]). We consider a binary classification problem, in which each point $\mathbf{x}_i$ is associated to a class label $y_i \in \{-1, 1\}$, indicating the class membership. As described earlier, the SVM can create a non-linear decision boundary by projecting the data through a non-linear function ϕ to a feature space of higher dimension. This means that points which cannot be separated by a straight line in their original space $\mathbb{R}^D$, are transformed to feature space $\mathbb{F}$ where a linear hyperplane can separate one class from the other. The corresponding decision boundary in the input space takes the form of a curve. The hyperplane in $\mathbb{F}$ is defined by $\omega^\top \mathbf{x} + b = 0$, with $\omega \in \mathbb{F}$ and $b \in \mathbb{R}$. The hyperplane determines the *margin* between classes; all points for class $y = -1$ are on one side, and all points for class $y = +1$ appear on the other. The distance from the closest point of each class to the hyperplane is equal; thus the hyperplane is defined by the maximal margin between the classes. To prevent the SVM classifier from over-fitting with noisy data (or to create a *soft margin*), slack variables ξ_i are introduced to allow some points to lie on the "wrong side" of the decision boundary, and a constant $C > 0$ penalises these "misclassifications" by a greater or lesses amount; higher values of C penalise misclassifications more than smaller values of C, and thus result in more detailed (less smooth) decision boundaries in an attempt to classify points correctly. The objective function of the SVM classifier is the following

minimisation problem:

$$\begin{array}{lll} \underset{\omega\in\mathbb{F},\ \xi\in\mathbb{R}^N,\ b\in\mathbb{R}}{\text{minimise}} & \dfrac{||\omega||^2}{2} + C\sum_{i=1}^{N}\xi_i & \\ \text{subject to} & y_i(\omega^\top\phi(\mathbf{x}_i)+b) \geq 1-\xi_i & \text{for all } i=1,\ldots,N \\ \text{and} & \xi_i \geq 0 & \text{for all } i=1,\ldots,N \end{array} \tag{5.12}$$

which is typically solved using Lagrange multipliers. The decision function (classification) rule for a data point $\mathbf{x}$ then becomes:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N}\alpha_i y_i k(\mathbf{x},\mathbf{x}_i) + b\right) \tag{5.13}$$

where $\alpha_i > 0$ are the Lagrange multipliers (which are weighted in the decision function and thus "supports" the "machine"); the function $k(\mathbf{x},\mathbf{x}_i) = \phi(\mathbf{x})^\top\phi(\mathbf{x}_i)$ is known as the kernel function. Since the outcome of the decision function only relies on the dot-product of the vectors in the feature space $\mathbb{F}$ (i.e., all the pairwise distances between points), it is not necessary to perform an explicit projection into $\mathbb{F}$; hence a kernel is used, exploiting Mercer's theorem. The latter shows that if k is a kernel function (i.e., positive semidefinite), then there exists some Reproducing Kernel Hilbert space $\mathbb{F}$ s.t. $k(\mathbf{x},\mathbf{x}_i) = \phi(\mathbf{x})^\top\phi(\mathbf{x}_i)$, as required. Popular choices for the kernel function are polynomial, sigmoidal, and the linear functions and the Gaussian radial basis function (RBF), given by:

$$k(\mathbf{x},\mathbf{x}_i) = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}_i||^2}{2\sigma^2}\right) \tag{5.14}$$

where $\sigma \in \mathbb{R}$ is the kernel parameter.

In the one-class SVM approach, Schölkopf et al. [2000] separate all the points from the origin (in the feature space $\mathbb{F}$) and maximise the distance from the hyperplane to the origin. This results in a binary classification that captures regions in the input space where the pdf of the data has most support. The

objective function is slightly different from the original in equation (5.12),

$$\begin{array}{lll} \underset{\omega\in\mathbb{F},\ \xi\in\mathbb{R}^N,\ \rho\in\mathbb{R}}{\text{minimise}} & \dfrac{||\omega||^2}{2} + \dfrac{1}{\nu N}\displaystyle\sum_{i=1}^{N}\xi_i - \rho & \\ \text{subject to} & (\omega\cdot\phi(\mathbf{x}_i)) \geq \rho - \xi_i & \text{for all } i = 1,\ldots,N \\ \text{and} & \xi_i \geq 0 & \text{for all } i = 1,\ldots,N \end{array} \tag{5.15}$$

where ν is the parameter that defines the smoothness of the boundary (it sets an upper bound on the proportion of training observations that lie on the "wrong" side of the hyperplane, and is also a lower bound on the number of training examples used as support vectors), and ρ is an offset parameter of the hyperplane. Again, via Lagrange multipliers and the use of a kernel function, the decision function becomes

$$f(\mathbf{x}) = \text{sign}(\omega\cdot\phi(\mathbf{x}_i) - \rho) = \text{sign}\left(\sum_{i=1}^{N}\alpha_i k(\mathbf{x},\mathbf{x}_i) - \rho\right) \tag{5.16}$$

In short, this method creates a hyperplane characterised by ω and ρ, which has maximal distance from the origin in the feature space $\mathbb{F}$ and separates all the data points from the origin. As described earlier, a different approach by Tax and Duin [1999] is to define a circumscribing hypersphere around the data in the feature space. For RBF kernels, this procedure gives the same result as the one-class SVM (Hoffmann [2007]).

Other novelty detection schemes. Other novelty detection schemes have been employed in the same way for building a scoring system (a review of which can be found in Pimentel et al. [2014]). Popular choices include different distance-based approaches (such as K-means and nearest neighbour-based methods) and reconstruction-based approaches (which include principal component analysis and different neural network configurations). The performances of these methods, however, have been reported to be similar to those of the previously described methods when applied to a number of one-class classification tasks (Chandola et al. [2009]; Clifton et al. [2014]; Kemmler et al. [2013]; Miljkovic [2010]; Pimentel et al. [2014]; Tax [2001]).

5.3 Coping with mixed numerical and categorical data

We previously observed that including information about whether the patient is on oxygen support or not increases the performance of scoring systems for identifying patient deterioration in the period before a major adverse event. We define the variable $S_O \in \{0, 1\}$, where $S_O = 1$ if the observation set was recorded while the patient was on oxygen support, and $S_O = 0$ if the patient was breathing room air (i.e., was not on oxygen support). The use of oxygen support already reflects a clinical intervention, and its inclusion in a kernel density estimate or SVM, for example, is not straightforward, being a categorical variable, as opposed to the vital-sign measurements which are *numerical* variables.

The ability to deal with datasets that contain both numerical and categorical attributes is an important characteristic because datasets with mixed types of attributes are common in other real-life data-mining applications. Most standard methods are not directly applicable to categorical data, for various reasons. The sample space for categorical data is discrete, and does not have a natural origin. Therefore, a Euclidean distance function on such a space is not really meaningful. Nevertheless, there are some strategies that can be used to cope with this problem. In the specific case of nonparametric kernel density estimation methods, conventional approaches do not handle mixed categorical and numerical data in a satisfactory manner. Although it is widely appreciated that one can use a frequency estimator to obtain consistent nonparametric estimates of a joint pdf in the presence of categorical variables, this frequency-based approach splits the dataset into parts ("cells"), and the number of observations lying in each cell may be insufficient to ensure the accurate nonparametric estimation of the pdf of the remaining numerical variables. Furthermore, it is not uncommon to encounter situations in which the number of cells exceeds the number of observation sets, hence the conventional frequency estimator cannot even be applied.

5.3.1 Joint density with categorical data

Li and Racine [2003] proposed a method for estimating a joint pdf defined over

mixed categorical and numerical data. This approach is briefly described in the next sub-sections. We first consider the estimation of a joint pdf over discrete (categorical) data.

Consider a finite dataset consisting of N samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ is an E-dimensional binary variable, $\mathbf{x}_i \in \{0,1\}^E$. Here, $\mathbf{x}_i$ denotes the i^{th} row of $\mathbf{X}$, with $i = 1, \ldots, N$, and $\mathbf{x}_{\cdot j}$ the j^{th} column of $\mathbf{X}$, with $j = 1, \ldots, E$.

A univariate binomial kernel function k_c can be defined as $k_c(x_{\cdot j}, x_{ij}|\lambda) = 1-\lambda$ if $x_{\cdot j} = x_{ij}$, and $k_c(x_{\cdot j}, x_{ij}|\lambda) = \lambda$ if $x_{\cdot j} \neq x_{ij}$, where λ is a smoothing parameter. For the multivariate case, a product kernel can be defined:

$$K_c(\mathbf{x}, \mathbf{x}_i|\lambda) = \prod_{j=1}^{E} k_c(x_{\cdot j}, x_{ij}|\lambda) = (1-\lambda)^{E-d_i}\lambda^{d_i}, \tag{5.17}$$

where $d_i = E - \mathbf{1}(\mathbf{x} - \mathbf{x}_i)$ corresponds to the number of "disagreement components" between $\mathbf{x}$ and $\mathbf{x}_i$, and $\mathbf{1}(A)$ is the usual indicator function, which equals one if $A \neq 0$, and zero otherwise. We note that d_i takes values in $\{0, 1, 2, \ldots, E\}$. We also note that λ can be an E dimensional vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_E]$; however, we consider the isotropic kernel, and the same value of λ is assumed for all dimensions of $\mathbf{x}$.

The final estimate of the pdf $p(\mathbf{x}|\lambda)$ is given by

$$\hat{p}(\mathbf{x}|\lambda) = \frac{1}{N}\sum_{i=1}^{N} K_c(\mathbf{x}, \mathbf{x}_i|\lambda) \tag{5.18}$$

We note that we have only described the case for a multivariate set of binary variables (i.e., variables that can only take one of two possible values). Other variables, such as level of consciousness, may have more than two possible categories. These variables, however, can be transformed into $E = m_c - 1$ binary variables, where m_c is the number of categories. The level of consciousness, for example, takes values in $\{A, V, P, U\}$ and which may be converted into $V = (1, 0, 0)$, $P = (0, 1, 0)$, and $U = (0, 0, 1)$.

5.3.2 Joint density with mixed data

We now consider the case involving mixed categorical and numerical data. As in the previous sub-section, $\mathbf{X}_C \in \{0,1\}^E$ represents the collection of categorical variables, and $\mathbf{X} \in \mathbb{R}^D$ to denote the collection of D numerical (real) variables; each collection with the same number N data samples. Let $\mathbf{x}_i$ denote the i^{th} row of $\mathbf{x}$, with $i = 1, \ldots, N$, and $\mathbf{x}_{.j}$ the j^{th} column of $\mathbf{x}$, with $j = 1, \ldots, D$. Let also $k_r(\cdot)$ be a univariate kernel function, and let $K_r(\cdot)$ be a product kernel function for the real variables:

$$K_r(\mathbf{x}, \mathbf{x}_i|\sigma) = \prod_{j=1}^{D} k_r(x_{.j}, x_{ij}|\sigma), \tag{5.19}$$

where σ is the smoothing parameter (or kernel width), if k_r is the Gaussian kernel[1]. Here, σ is assumed to be the same for all D dimensions. We further define $\mathbf{B} = (\mathbf{X}, \mathbf{X}_C)$, and we use $p(\mathbf{b}) = p(\mathbf{x}, \mathbf{x}_C)$ to denote the joint pdf based on $\mathbf{B}$. The pdf $p(\mathbf{b}|\lambda, \sigma)$ is estimated by

$$\hat{p}(\mathbf{b}|\lambda, \sigma) = \frac{1}{N} \sum_{i=1}^{N} W(\mathbf{b}, \mathbf{b}_i|\lambda, \sigma), \tag{5.20}$$

where $W(\mathbf{b}, \mathbf{b}_i|\lambda, \sigma) = K_c(\mathbf{x}_C, \mathbf{x}_{Ci}|\lambda) K_r(\mathbf{x}, \mathbf{x}_i|\sigma)$. The values of the smoothing parameters, λ and σ, may be found by maximising the leave-one-out likelihood of the data,

$$J(\sigma, \lambda) = \frac{1}{N} \sum_{j=1}^{N} \log \left(\frac{1}{N-1} \sum_{i=1, i \neq j}^{N} W(\mathbf{z}_j, \mathbf{z}_i|\lambda, \sigma) \right) \tag{5.21}$$

Using this approach, we can construct a model of normality that can cope with both numerical and categorical variables. We note that this approach has

[1] We note that the product kernel $K_r(\cdot)$ used in Li and Racine [2003] is different from that presented in Equation 5.3, which uses a multivariate kernel function.

not previously been applied to the problem of patient monitoring.

5.4 Modelling physiological deterioration using novelty detection approaches

The results presented in the previous chapter suggest that reasonable performance may be achieved using the observation sets acquired by nursing staff during their routine observations of the patients, which is agrees with results available in the literature. We have therefore implemented the different modelling methods introduced in this chapter using the same vital-sign data attributes included in most EWS systems. The occurrence of a major adverse event within 24 hours of a given observation set was again used as the outcome for both Portsmouth and CALMS-2 datasets considered previously. Observational data were pre-processed in order to remove artefactual vital-sign values from the observation sets (as was done in the previous chapters and described in section 3.3.1). Observation sets with more than two missing variables were not included in the analysis; for those observation sets with one or two missing variables, the missing value was imputed using the mean of that variable in the training set.

All real variables included in each model were normalised using the same zero-mean, unit-variance transformation (as was described above). All categorical variables (oxygen support and level of consciousness) were converted to binary variables using the procedure described above.

5.4.1 Models considered

The models explored here comprise modelling strategies that have been applied to patient monitoring in previous studies. We used different sets of variables to construct these models, and applied well-established techniques to tune the parameters of each model, as described below.

Baseline method (BAS). We employed the model proposed by Tarassenko et al. [2006] as our baseline approach. Before normalising all variables, we combined the systolic and diasbolic BP into a single parameter by computing their

arithmetic mean. Each observation set is a 5-dimensional vector that comprises HR, RR, SpO_2, temperature, and the systolic-diastolic average. We then re-used the 400 cluster centres and the value of σ from Hann [2008] to build the model (as detailed in section 5.2.1). The threshold on the novelty score $z(\mathbf{x})$ was determined using 5-fold cross-validation.

Kernel density estimate (KDE). We retrained a KDE using our training data. In this model (as in the models subsequently described), we used the systolic BP rather than the systolic-diastolic average (used for the BAS model), to allow a direct comparison with EWS systems used in practice. The value of σ was found using the leave-one-out likelihood $J(\sigma)$ of the training data (as denoted in equation 5.7). A simple gradient descent method was employed to find the best parameter $\sigma \in \mathbb{R}_+$. Again, the threshold on novelty scores $z(\mathbf{x})$ was found using cross-validation. We then implemented a KDE using mixed categorical and real data as described in the previous section. Here, we used the 5-dimensional real feature-vector as before, augmented with categorical variables corresponding to the use of oxygen and level of consciousness. Again, the values of the parameters were determined via leave-one-out likelihood $J(\sigma, \lambda)$ of the training data (equation 5.21), for which a gradient descent method was employed to find the best parameter set $\sigma, \lambda \in \mathbb{R}^2_+$. The threshold of the model was also found using cross-validation.

One-class support vector machine (SVM). The one-class SVM was implemented using the LIBSVM library (Chang and Lin [2011]). As in Schölkopf et al. [2000], we used a Gaussian kernel. To adjust the decision threshold, the one-class SVM has basically two parameters ν and σ to be set. The value of the kernel width σ was varied over the values $\sigma = [0.01, 0.1, 0.5, 1, 1.5, 2, 3, 4]$. The value of ν, which is related to the fraction of false positives in a novelty classification task (Schölkopf et al. [2000]), was varied over values in the interval from 0.1 to 0.9, with increments of 0.05, in order to avoid edge effects for small or large values of ν (as suggested by Schölkopf et al. [2000] and Rabaoui et al. [2007]). A search over (σ, ν) was performed. We also evaluated the ability of this approach to cope with mixed categorical and real variables, and, as with the KDE approach, two SVM models were built using the two sets of variables: (1) the 5-dimensional

feature vector comprising the five vital signs; and (2) the 5-dimensional feature vector augmented with the categorical variables for oxygen support and level of consciousness. In addition, in both cases, we implemented the heuristic introduced by Wong [2011]: any SpO_2 values that were greater than the mean of the training data were replaced by that mean in the training set. For example, an SpO_2 value of 99% would be replaced by the mean value of the training data (which is commonly around 96%). This heuristic was required to avoid the SVM model having to cope with SpO_2 values above 100%.

Active outlier method (AOM). This method, proposed by Abe et al. [2006], is based on an "ensemble-based minimum-margin active learning", which augments the one-class dataset with synthetic "abnormal" samples obtained via rejection sampling. Rejection sampling, also commonly called *accept-reject algorithm*, is a type of Monte Carlo sampling method that works for any distribution in $\mathbb{R}^m$. We used an implementation from Erdoğan [2011]. A decision tree was used as described in Abe et al. [2006], and we assumed the sampling distribution as uniform. We varied the number of ensemble learners $N_c = [4, 8, 12, 16, 20, 30, 40]$, and the outlier threshold between 0.1 and 0.9, increasing by 0.05. Again, we constructed models using the two sets of variables used in the two previous approaches (i.e., with and without categorical data).

Gaussian mixture model (GMM). A GMM was also implemented and used to determine the density of the real training data. Different numbers of mixture components M were used to train the model, $M = N \cdot \{1/2, 1, 2, 5, 10, 15, 20, 25, 30\}/100$, where N is the number of data points in the training set. For each value of M, the maximum likelihood estimates of the model parameters ϑ were determined using EM (Bishop [2006]), and the K-means algorithm (with ten random initialisations) was used to initialise ϑ. The Bayesian information criterion, BIC, is

$$\mathrm{BIC} = -2 \log \hat{p}(\mathbf{x}|\hat{\vartheta}) + k \log N \tag{5.22}$$

where $p(\mathbf{x}|\hat{\vartheta})$ is the likelihood of the data given the model defined by $\hat{\vartheta}$, and k

is the number of parameters to be estimated. The BIC was evaluated for each value of M and we used that value of M that minimised BIC for the final model. The value of the threshold on $z(\mathbf{x})$ was selected via cross-validation, as before.

5.4.2 Evaluation of models

The performance of each method was evaluated during the cross-validation procedure described in section 4.2; i.e., a five-fold validation with the Portsmouth dataset was used to train/validate the proposed methods, and the entire CALMS-2 dataset was used to assess the classification performance (identification of major adverse event within 24 hours of the observation set). Figure 5.1 illustrates the procedure used for evaluating the performance of the models considered in this study. We emphasise that the training set contained data from "normal" patients; only the (cross-) validation and test sets contained data from "abnormal" patients. The optimal parameter setting for each fold was selected based on the best pair of sensitivity/specificity values on the validation set (as in the previous two chapters).

For simplicity of exposition, we denote each method by its 3-letter acronym, and with the use of real and categorical attributes used in subscript; for example, $\text{SVM}_{r,c}$ corresponds to the support vector machine built using the 5-dimensional real feature-vector (i.e., the five vital signs comprising HR, RR, SpO_2, temperature and systolic BP), augmented with the categorical variables corresponding to the use of oxygen support and level of consciousness.

5.4.3 Results

The results of this experiment are shown in Tables 5.1 and 5.2, which show the different models and their performance measures evaluated on the validation sets during the five-fold cross-validation, and on the test set, respectively. We observe that, with the validation sets, the one-class SVM approach that included both real and categorical variables resulted in the highest AUROC and partial AUROC, followed by the KDE that used the same set of variables. In fact, the difference in AUROC between the best performing two models, the one-class SVM and KDE for mixed data, was not found to be significant. Interestingly, on the test set,

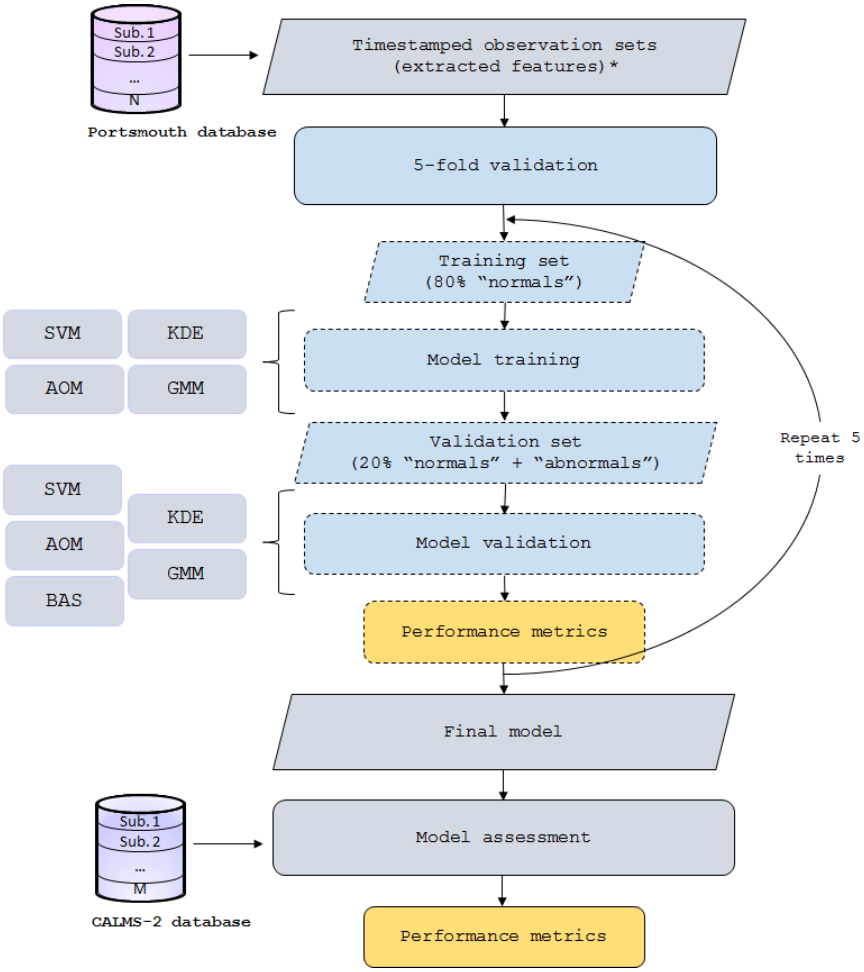


Figure 5.1: Flowchart of model development process for the machine learning models considered. Dashed lines indicate steps that are repeated 5 times (i.e., they are included in the 5-fold validation process using the Portsmouth dataset).

the better performances were achieved by models that were trained without the categorical variables (i.e., only numerical attributes were considered). Of all the models evaluated, the baseline approach (which uses the original set of five vital sign values as features) provided the lowest performance.

5.4.4 Discussion

In this chapter, we have described machine learning techniques for novelty detection based on multivariate density estimation, SVMs, and decision trees. Some

Table 5.1: Performance metrics for the different approaches described in this chapter evaluated using the Portsmouth dataset. Results are presented with mean and standard error of the Wilcoxon statistic. Best values for each performing metric are underlined.

Model	AUROC	pAUROC	Sens.	Spec.	PPV
BAS_r	0.772 (0.006)	0.809 (0.008)	0.714 (0.005)	0.745 (0.003)	0.019 (0.001)
KDE_r	0.818 (0.009)	0.810 (0.007)	0.729 (0.011)	0.770 (0.011)	0.023 (0.004)
$KDE_{r,c}$	0.848 (0.010)	0.863 (0.010)	0.781 (0.011)	0.808 (0.009)	0.031 (0.005)
SVM_r	0.827 (0.008)	0.807 (0.005)	0.734 (0.010)	0.767 (0.010)	0.024 (0.003)
$SVM_{r,c}$	0.857 (0.025)	0.876 (0.006)	0.792 (0.011)	0.795 (0.010)	0.034 (0.004)
AOM_r	0.794 (0.010)	0.816 (0.003)	0.731 (0.012)	0.745 (0.009)	0.021 (0.002)
$AOM_{r,c}$	0.813 (0.024)	0.822 (0.018)	0.744 (0.013)	0.760 (0.015)	0.25 (0.007)
GMM_r	0.810 (0.012)	0.804 (0.005)	0.733 (0.011)	0.747 (0.011)	0.018 (0.004)

attempts have previously been made to compare different one-class classification methods (Irigoien et al. [2014]; Khan and Madden [2010]) showing the superiority of some types of techniques including those presented here. Beyond these trends, the advantage of one technique over another is very much dataset-dependent: the number of observations and features, the nature of the relation between them, and the presence of artefacts. Because not all existing machine learning algorithms could be investigated in this work, we restricted our analysis to those described in this chapter.

Models with only real-valued data. Interestingly, among all methods in which only the set of real variables corresponding to the five main vital signs was considered, our results suggest that the difference in performance between the novelty detection schemes explored in this work is small (on the test set). SVM_r provided the highest AUROC among all methods considered. We note, however, that this performance is not significantly higher than, for example, that obtained with a KDE. The similarity of the performance achieved by the GMM and the KDE, which are both density estimation-based methods, and the performance of the SVM, which is a boundary method, is anticipated. Vert and Vert [2006] determined the asymptotic behaviour (in the situation where the number of examples tends to infinity) of SVMs using an RBF kernel. The output of the SVM provides an estimate of the level sets of the density (i.e., it approaches the

Table 5.2: Performance metrics for the different approaches described in this chapter evaluated using the CALMS-2 dataset. Results are presented with mean and standard error of the Wilcoxon statistic. Best values for each performing metric are underlined.

Model	AUROC	pAUROC	Sens.	Spec.	PPV
BAS_r	0.779 (0.010)	0.802 (0.009)	0.716 (0.016)	0.738 (0.003)	0.085 (0.003)
KDE_r	0.797 (0.013)	0.825 (0.010)	0.765 (0.016)	0.781 (0.003)	0.111 (0.004)
$KDE_{r,c}$	0.788 (0.009)	0.821 (0.008)	0.757 (0.014)	0.770 (0.003)	0.101 (0.003)
SVM_r	0.801 (0.010)	0.822 (0.009)	0.767 (0.016)	0.777 (0.003)	0.104 (0.003)
$SVM_{r,c}$	0.792 (0.008)	0.819 (0.008)	0.758 (0.015)	0.768 (0.002)	0.098 (0.003)
AOM_r	0.773 (0.010)	0.804 (0.009)	0.696 (0.016)	0.755 (0.002)	0.093 (0.003)
$AOM_{r,c}$	0.771 (0.010)	0.797 (0.010)	0.694 (0.016)	0.748 (0.003)	0.087 (0.004)
GMM_r	0.781 (0.009)	0.810 (0.008)	0.733 (0.015)	0.744 (0.002)	0.093 (0.003)

level sets on the pdf of the data in its tails), which is the overall goal of density estimation methods. We note, however, that the GMM provided a slightly worse performance than the other methods tested. This may be due to poor optimisation of the number of mixture components. In our experiments, we used a model selection metric for selecting the number of mixture components, which balances a term that measures how well the model fits the data with a complexity penalty term that favours simpler models. Other approaches, based on Bayesian nonparametric models (Orbanz and Teh [2010]), provide a different approach to this problem. Rather than comparing models that vary in complexity, the aim of a Bayesian nonparametric approach is to fit a single model that can adapt its complexity to the data. The hierarchical Dirichlet process mixture model implements a variant of the traditional GMM with a potentially infinite number of components using the Dirichlet process. It does not require the number of components to be selected, and at the expense of extra computational time, only a loose upper bound on this number needs to be specified beforehand. Nevertheless, the kernel density estimates provide an alternative, nonparametric approach which can be compared with GMMs.

We also observe that the baseline model provided the lowest performance among the models tested. There are two major differences that may have contributed to this result. Unlike previous work, we used observational (i.e., manually-

recorded) data acquired periodically by nurses at the bedside, whereas the baseline model was trained using data acquired continuously (via bedside monitors) from a large population of acutely-ill patients. As observed in the last chapter, there may be substantial differences between data acquired continuously from patient monitors and the periodic data from nurses' observations. In addition, the baseline model uses the systolic-diastolic average as one of its features, which has no precedent in physiological monitoring and may not accurately model the change in a patient's condition (as also noted by Wong [2011]). Hence, the performance of the baseline model was slightly lower than that of the other approaches, which all used vital-sign data from periodic observations made by nurses.

Models with mixed categorical and real-valued data. As observed in the previous chapters, the inclusion of information such as whether the patient is on oxygen support or not dramatically improves the performance of scoring systems. This is the case for both the Portsmouth and CALMS-2 datasets. The results obtained in this chapter support a similar conclusion when we consider those resutls from the validation set. We implemented different machine learning algorithms to deal with the presence of mixed categorical and real-valued attributes. As shown in Table 5.1, for the validation set, the performance of the models trained with mixed categorical and real-valued data was superior to that of models that did not consider information about the level of consciousness and the use of oxygen support. We note that for the AOM method, the improvement (from AOM_r to $\text{AOM}_{r,c}$) was not as pronounced as that seen for the other two methods. The main reason for this is that the "active outlier" approach uses a frequency estimator based on each individual categorical variable to estimate the joint pdf of the data. This frequency-based approach splits the dataset, and the number of samples in each partition may be insufficient for an accurate estimation of the pdf of the real-valued variables. For the other two methods ($\text{KDE}_{r,c}$ and $\text{SVM}_{r,c}$), a substantial improvement in AUROC and partial AUROC was obtained with the cross-validation data.

Results for the same method applied to the test data (Table 5.2) show that models without the use of categorical data (KDE_r and SVM_r) outperformed those models that included the categorical data. One of the main reasons that may

explain this result is the different cohort of patients included in the CALMS-2 dataset. In fact, the test set comprised data acquired from post-operative patients, while the Portsmouth dataset used for training and validating was acquired from medical patients admitted to a Medical Assessment Unit (MAU). Apart from the differences between the two patient populations and derived outcome variables, that have already been highlighted in chapter 3, there are also different clinical protocols used in these different hospital settings, which may affect not only the distribution of the data but also the relationship between the variables included in the model. For example, in a MAU, oxygen support is provided to patients who are in a state of hypoxia, or are at risk of becoming hypoxic, while in the post-operative ward, patients are generally on oxygen support for the first few of days after surgery (as part of the clinical protocol for post-operative care), and not necessarily because patients are (or are at risk of) becoming hypoxic. The use of oxygen support has a direct influence on the values of the vital signs that are measured during those periods of time in which observations take place; they certainly have an effect in SpO_2. Conversely, the values of the vital signs may affect the decision of whether or not to provide oxygen support to the patient. Figure 5.2 shows the difference between the vital signs that were measured while the patient was on oxygen support and while the patient was breathing room air, for both datasets. Only "normal" observations sets (i.e., observation sets that are not followed by an adverse event within the following 24 hours) were taken into account in this analysis. We observe that, for example, in the CALMS-2 dataset, the presence or absence of oxygen support appears to affect the SpO_2 the most: a patient's SpO_2 is higher when on oxygen support. However, in the Portsmouth dataset, there is very little difference between the two cases. Because patients are hypoxic, oxygen support is used to raise the levels of oxygen to "normal" values (approximately 96%). We also observe that, when patients from the Portsmouth dataset are on oxygen support, their heart rate and respiratory rate are elevated (which is a normal physiological compensation mechanism during hypoxia), whereas oxygen support makes no difference to the heart rate and respiratory rate of CALMS-2 patients.

Multivariate models have the ability to capture these correlations between the different variables. The varying effect of oxygen support between the two datasets

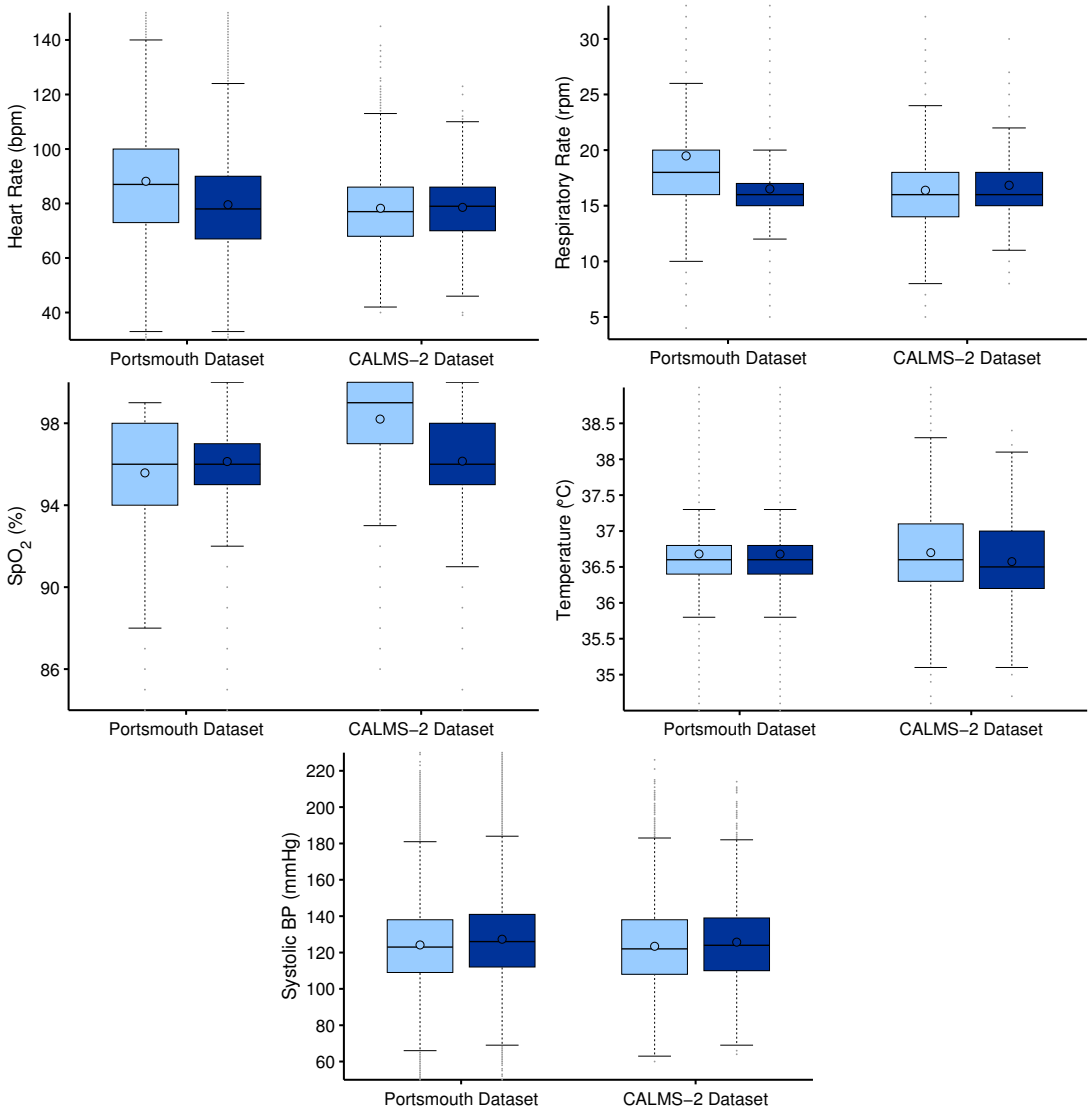


Figure 5.2: Summary statistics of the five vital signs (from each dataset) included in the analysis for (1) observations performed while the patient was on oxygen support (lighter blue), and (2) observations performed while the patient was not on oxygen support (darker blue). Summary statistics are represented for each group in a boxplot: the length of the coloured box represents the interquartile range (from the 25^{th} to the 75^{th} quantile); the dot in the box interior represents the mean; the horizontal line in the box interior represents the median; the vertical dotted lines issuing from the box extend to the most extreme data points not considered outliers; outliers are plotted individually with gray dots. Data points are drawn as outliers (for visualisation purposes only) if they are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25^{th} and 75^{th} quantiles, respectively.

may thus account for the fact that inclusion or exclusion of categorical variables results in different classification performance between these two datasets.

Machine learning models vs. early-warning scores. In principle, data-fusion methods avoid three of the major drawbacks of the track-and-trigger scoring systems:

1. these methods are derived from data collected from patients that are representative of the study population, thus providing an objective, data-driven score;

2. unlike early warning scores, these methods are not constrained to integer outputs, and allow for small changes in vital signs which provide coarse estimates of vital-sign abnormality;

3. machine learning methods based on multivariate models have the ability to capture correlations between variables, which may help to improve their performance.

Notwithstanding the above, we observe that the performance of machine learning methods was similar to that of the best-performing EWS systems on the validation sets. The best-performing model has a mean (SD) AUROC of 0.857 (0.025), compared with an AUROC of 0.881 (0.006) for the best-performing EWS system (section 3.3.3). The best-performing model from this chapter had a partial AUROC of 0.876 (0.006), compared to 0.827 (0.006) for the best EWS system. This difference in the partial AUROC values may be explained by the quantisation effect created by the integer outputs of EWS systems. A similar performance in terms of specificity, sensitivity and positive predictive values was also obtained. It is important to note that the best-performing EWS systems were derived from the data used for training and cross-validation of the machine learning methods.

However, when applied to test data, the performance of the machine learning models was slightly worse than that obtained by the EWS systems. The $\text{KDE}_{r,c}$ achieved an AUROC of 0.797 (0.013), which is lower than that achieved with the best-performing EWS system, with a mean of 0.841 (0.008). Also, the best-performing model achieved a sensitivity and specificity of 0.765 (0.016) and 0.781

(0.003), while the same figures for the EWS system proposed by Prytherch et al. [2010] were 0.801 (0.014) and 0.765 (0.002), respectively. As explained above, the different relationships that were captured by the multivariate models during training may not generalise to a different (test) patient population. Furthermore, these correlations between the physiological variables may evolve over time as patients recover from a major intervention. This may explain the relatively poor performance of the machine learning models on the CALMS-2 dataset. Models, or scoring systems, based on univariate analysis may generalise better to different populations as these relationships are not captured and are not implicitly modelled. These findings strongly support the recommendation of Cuthbertson et al. [2007], which was that scoring systems for the identification of deterioration need to be developed and validated for specific patient groups. That is, scoring systems that are developed for one patient population may not be suitable for a different patient population, as demonstrated in this analysis.

5.5 Conclusion

In this chapter, we have briefly introduced the key concepts of machine learning and described the theoretical framework behind techniques that are currently used in novelty detection and one-class classification tasks. In particular, different multivariate (kernel-based) density estimation and boundary-based methods were described with the aim of capturing the "normal" physiological behaviour of patients, so that "abnormal" observation sets may be identified and the occurrence of an adverse event detected earlier.

Other data-fusion models have been proposed in the literature for identifying deterioration in patients outside the ICU (such as the studies by Churpek et al. [2014], Alvarez et al. [2013], and Escobar et al. [2012]). These models are typically based on fitting a logistic regression model to a large dataset (over 200,000 hospital admissions) that includes not only vital signs, but also results from laboratory tests and demographic data. In a data-rich environment, in which all these variables are available, different approaches have to be employed in order to provide *parsimonious* models that can be easily interpreted, for which acceptance amongst clinical staff members is likely to be higher. It is important to

notice, however, that laboratory tests are not performed as frequently as the measurement of vital signs on general wards. Given this, we have concentrated on building models with data that are monitored more frequently on general wards.

The most important result in this section is the fact that these data-driven machine learning models can achieve levels of performance that are comparable to those provided by the scoring systems currently in clinical use. However, these multivariate models need to be population specific, in order to capture the distributions and correlations between the different variables that are representative of the patient population under study. This is relevant because the same principled approach needs to be applied to continuously-acquired data, as robust machine learning methods are needed to deal with the noisy and artefactual data typical of these monitors. Furthermore, these methods can be used to explore the incorporation of additional information that may increase predictive power as demonstrated in the next chapter.

Chapter 6

Physiological trajectory and variability for post-operative patients

The analysis presented in earlier chapters has been carried out without exploiting prior knowledge concerning the physiology of the post-operative patient population. Previous results have hinted that a time-varying trend exists in the physiological variables recorded in the period immediately after surgery. This may simply result from the fact that a major operation represents a substantial physiological insult, from which the patient is expected to recover to their "normal" physiological status. In this chapter, we focus our analysis on the CALMS-2 patient population, and explore the trajectories associated with recovery following surgery. We also introduce a new concept for EWS systems: the variability of physiological variables over a 24-hour period. A strategy is then proposed for incorporating this information into the data-fusion models.

6.1 Data visualisation

In order to analyse the physiological trajectories of post-operative patients, it is important to consider the different periods of time for which these patients had their vital signs recorded. Table 6.1 lists the main characteristics of the two

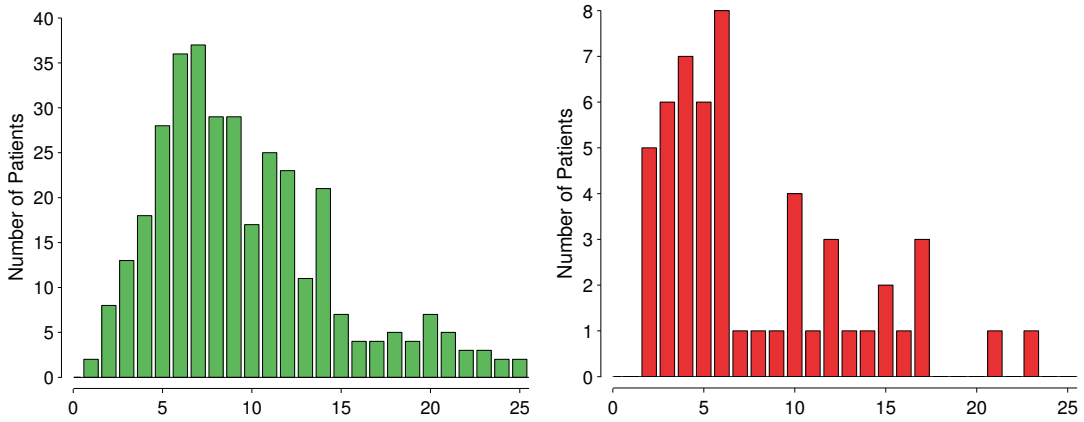
Table 6.1: Patient characteristics for the normal and abnormal groups of patients. Values are presented as median values and interquartile range within brackets (unless otherwise stated).

	Normal (N = 357)	Abnormal (N = 50)
Age, yr	63 (53-69)	63 (51-72)
Gender, N males (%)	206 (57.7)	26 (52.0)
Length of stay, days	9 (6-13)	23 (16-45)
Time to first event, days	-	6 (5-10)

groups of patients (N = 407)[1]. Histograms of the length of stay for patients in the normal group and time to the adverse event for patients in the abnormal group are shown in Figure 6.1. In this analysis, the length of stay on the ward is defined to be the time between the patient's operation and their discharge from the post-operative ward. For example, some patients were not admitted to the ward on the day of surgery. Some patients were (electively) admitted to the ICU from the operating theatre before going to the post-operative ward. The median length of stay on the ward after the operation for the normal group of patients was 9 days (25^{th} quantile: 6 days; 75^{th} quantile: 13 days). The equivalent figures for the abnormal group are 23 days (25^{th} quantile: 16 days; 75^{th} quantile: 45 days); i.e., the abnormal group has much higher 50^{th} and 75^{th} quantiles than for the normal group, because the length-of-stay figures are skewed by the time spent in the ICU after emergency admission from the ward. Of more relevance for the abnormal group of patients is the median time to the first major adverse event: 6 days with an IQR of 5 days.

We first concentrate our analysis on the trajectory of each individual vital sign for patients who recovered from surgery and who did not suffer any major adverse event in the course of their stay on the ward (i.e., patients in the normal group).

[1]Some characteristics of the cohorts given in this table have been further detailed in Table 2.3 from section 2.3.



(a) Day of discharge (b) Day of event

Figure 6.1: **(a)** Histogram of the post-operative length of stay of the 357 patients in the normal group on the ward (data are shown up to the 95$^{\text{th}}$ quantile of the length of stay). **(b)** Histogram of the time to the adverse event (from the day of surgery) for the 50 patients in the abnormal group.

6.1.1 Univariate vital-sign data distribution

The changes in vital-sign data distributions between admission to the post-operative ward and subsequent discharge, when the patient was deemed sufficiently stable to go home, were evaluated for the patients in the normal group. For this, we selected all the observations performed during the first 24 hours on the ward, and during the last 24 hours. Normalised histograms (unit area under the curve) were subsequently computed for each physiological variable (HR, RR, SpO_2, temperature, systolic and diastolic BP).

The empirical pdfs (histograms) for each physiological variable for both admission and discharge periods are shown in Figure 6.2[1]. Table 6.2 gives the corresponding means and standard deviations. We observe that apart from HR, the distributions represented for each of the other five vital signs vary from admission to discharge, as the patient recovers from major surgery. The HR distributions are similar and approximately symmetrical. For the distribution of SpO_2 at admission, a mode occurs at 100%. Patients are likely to achieve 100% oxygen saturation only if they are receiving additional oxygen support through an oxygen

[1] On the bottom of each histogram, the presence (or absence) of the values are denoted with small "sticks".

Table 6.2: Vital-sign data means (standard deviation) for the first (admission) and last (discharge) 24 hours on the ward for patients in the normal group.

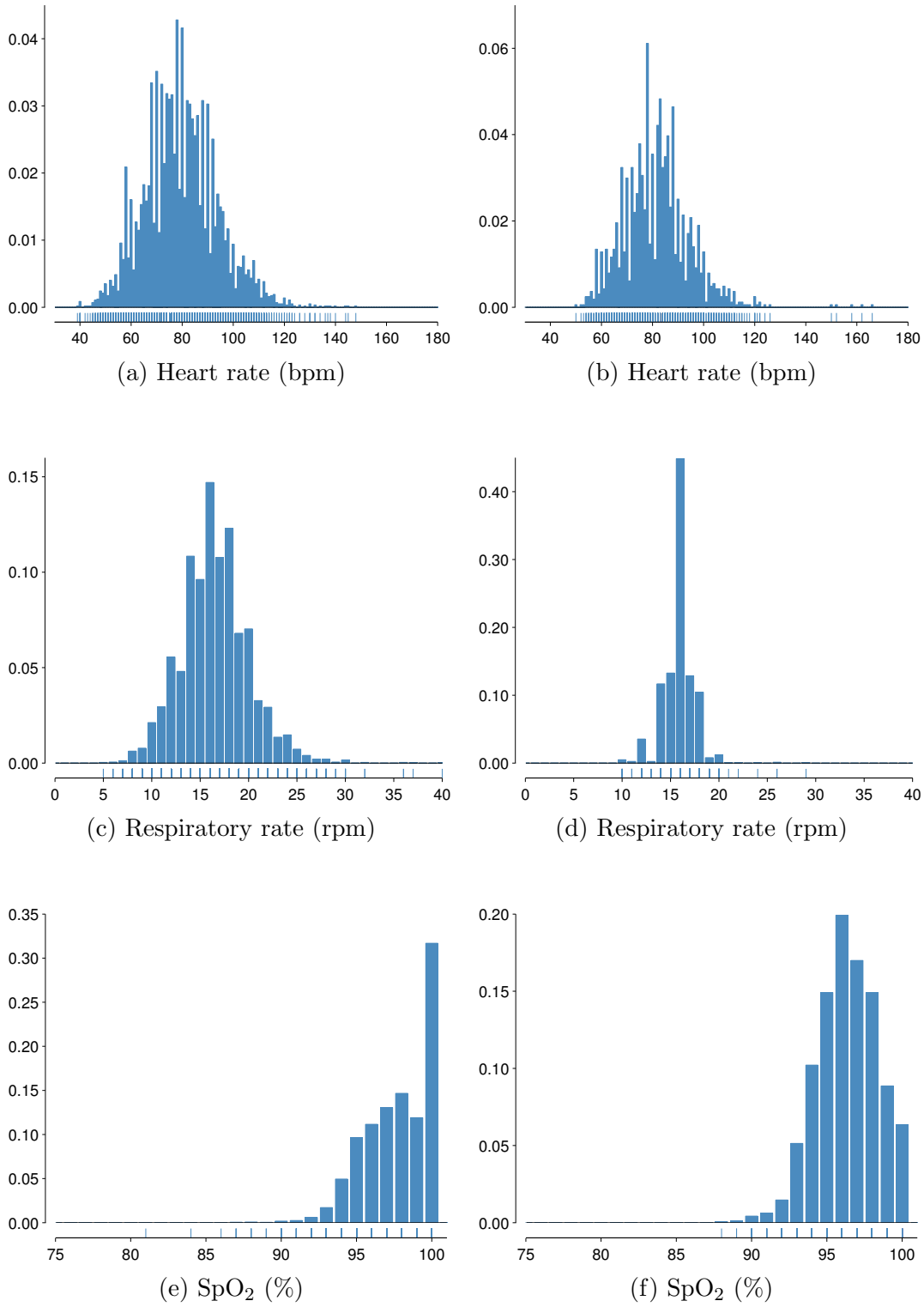
	HR (bpm)	RR (rpm)	SpO_2 (%)	Temp. (°C)	Sys BP (mmHg)	Dia BP (mmHg)
Admission	80 (14)	17 (3)	98 (2)	36.6 (0.5)	115 (22)	60 (13)
Discharge	82 (13)	16 (2)	96 (2)	36.4 (0.5)	130 (18)	74 (10)

mask or cannula. This is more likely to occur in the first days after surgery, which may explain the shift towards high SpO_2 values at admission (with respect to the period just before discharge). RR distributions differ essentially with respect to the spread of values around the mode (which is 16 rpm in both distributions). We also note that the values of RR are discretised; i.e., there is a pronounced peak at 16 rpm, which may be related to the counting method used to record this vital sign on general wards[1]. The distributions of temperature and blood pressure (both systolic and diastolic) show that patients are, in general, mildly pyrexic (high temperature) and hypotensive (low blood pressure) when admitted to the ward following surgery. They subsequently show decreasing temperature (returning to more "normal" values) and increasing blood pressure (returning to more "normal" values) by the last day of their stay on the ward.

From these results, we assume that there are obvious changes in the vital signs of post-operative patients between admission to and discharge from the ward. We conducted a subsequent analysis that considers different periods of time throughout the patient's stay on the ward. We examined the following five subgroups of observations:

- $\mathbf{G}_1$: the set of averages of all observations performed on the first day of the patient's stay on the ward (admission day);
- $\mathbf{G}_2$: the set of averages of all observations performed on the day that corresponds to a quarter (25%) of the length of the patient's stay on the ward;
- $\mathbf{G}_3$: the set of averages of all observations performed on the day that corresponds to half (50%) of the length of the patient's stay on the ward;

[1]This is a topic that was discussed in greater detail in chapter 4, in which continuous data were compared to observational data.



(a) Heart rate (bpm)

(b) Heart rate (bpm)

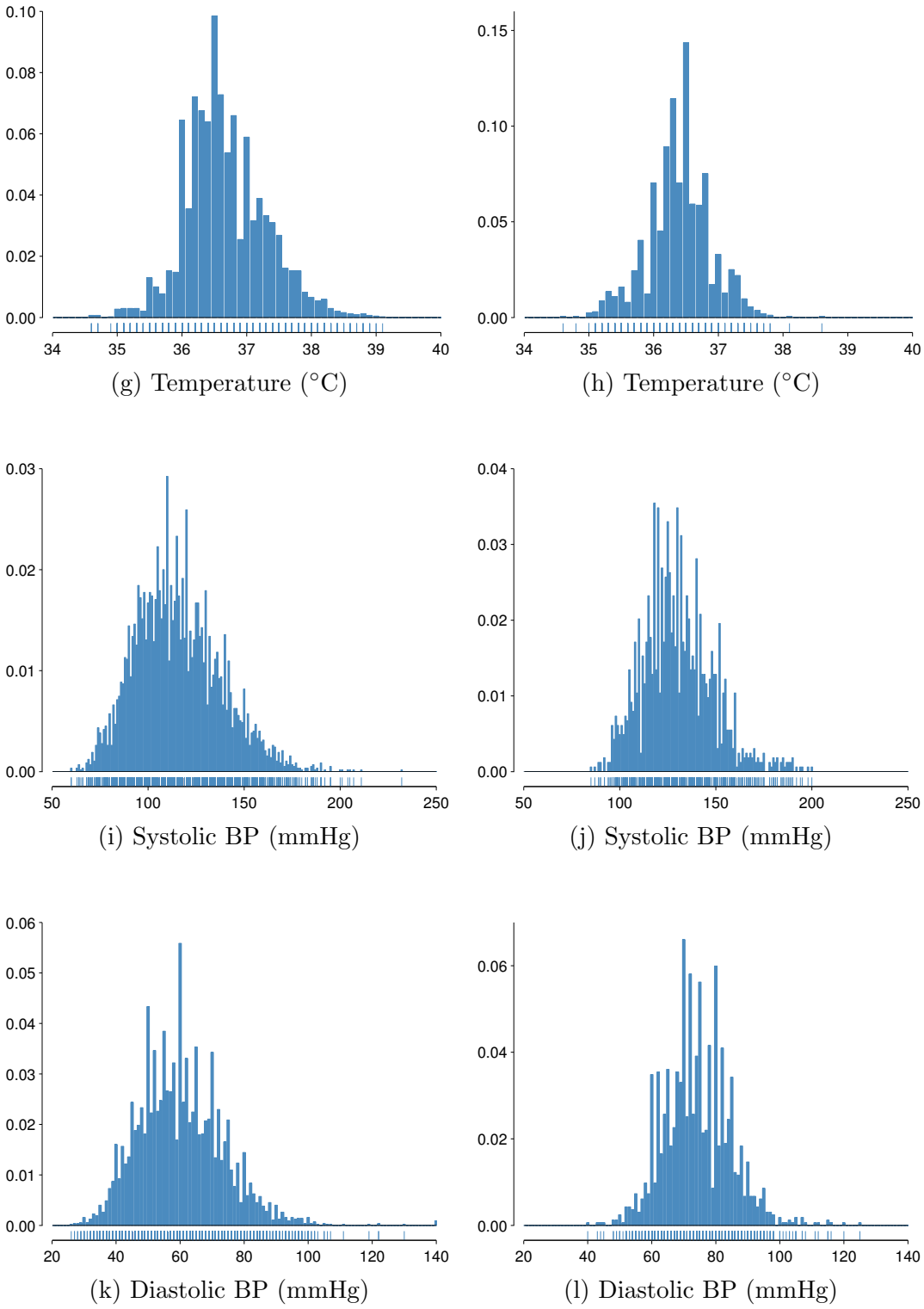
(c) Respiratory rate (rpm)

(d) Respiratory rate (rpm)

(e) SpO_2 (%)

(f) SpO_2 (%)

Figure 6.2: Normalised histograms for the vital signs acquired from patients in the normal group at admission to the post-operative ward (left, (a)-(c)-(e)) and during the 24 hours before discharge (right, (b)-(d)-(f)). (This figure continues on the following page.)



(g) Temperature (°C) (h) Temperature (°C)

(i) Systolic BP (mmHg) (j) Systolic BP (mmHg)

(k) Diastolic BP (mmHg) (l) Diastolic BP (mmHg)

Figure 6.2: (Continuation) Normalised histograms for the six vital signs acquired from patients in the normal group at admission to the post-operative ward (left, (g)-(i)-(k)) and during the 24 hours before discharge (right, (h)-(j)-(l)).

- $\mathbf{G}_4$: the set of averages of all observations performed on the day that corresponds to 75% of the length of the patient's stay on the ward;
- $\mathbf{G}_5$: the set of averages of all observations performed on the last day of the patient's stay on the ward (discharge day).

These subgroups were defined in this way because of patient's different lengths of stay on the ward. To avoid the presence of the same observation set of one subject appearing in more than one group, we focused on patients who have a length of stay on the ward of at least 5 days for this analysis. As a result, 297 patients were considered.

In order to compare the resulting vital-sign data distributions between the different groups, three different metrics were used: the Kolmogorov-Smirnov metric (Chakravarty et al. [1967]), the symmetrical Kullback-Leibler distance (Kullback and Leibler [1951]; Veldhuis [2002]), and the Bhattacharyya distance (Bhattacharyya [1946]).

Kolmogorov-Smirnov distance. This is a metric that quantifies the distance between the empirical distributions of two sample sets (Chakravarty et al. [1967]). Considering two probability densities of the same random variable x, $p(x)$ and $q(x)$, if $P(x)$ and $Q(x)$ are the respective cumulative distribution functions (cdfs), the KS distance (ΔKS) between them is defined by

$$\Delta KS(p,q) = \sup\big(|P(x) - Q(x)|\big) \tag{6.1}$$

where $\sup(d)$ is the supremum of the set of distances d.

Symmetrical Kullback-Leibler divergence. This metric compares the entropy of two distributions over the same random variable (Kullback and Leibler [1951]). It measures the number of additional bits required when encoding a random variable with a distribution $p(x)$ using the alternative distribution $q(x)$. This metric is asymmetrical, but it can be easily modified to be symmetrical, ΔKL (Veldhuis [2002]), which is defined as

$$\Delta KL(p,q) = \sum_{x \in X} (p(x) - q(x)) \log \frac{p(x)}{q(x)} \tag{6.2}$$

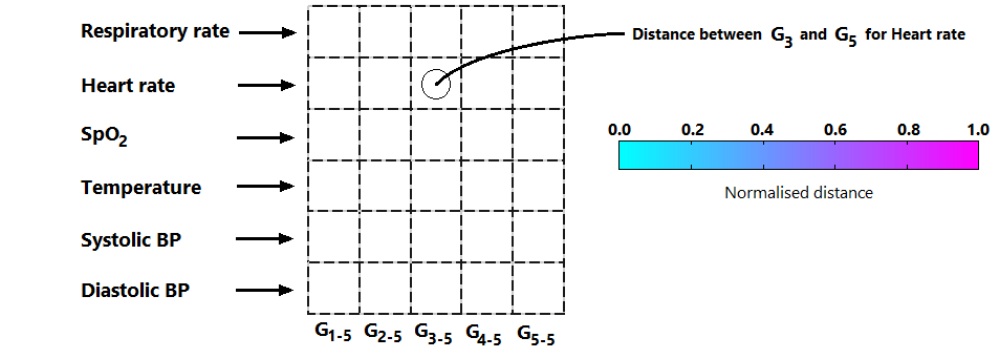
Bhattacharyya distance. The Bhattacharyya distance (Bhattacharyya [1946]) measures the amount of overlap between two distributions ($\Delta Bhat$), and is defined by

$$\Delta Bhat(p,q) = -\log \left(\sum_{x \in X} \sqrt{p(x)q(x)} \right) \tag{6.3}$$

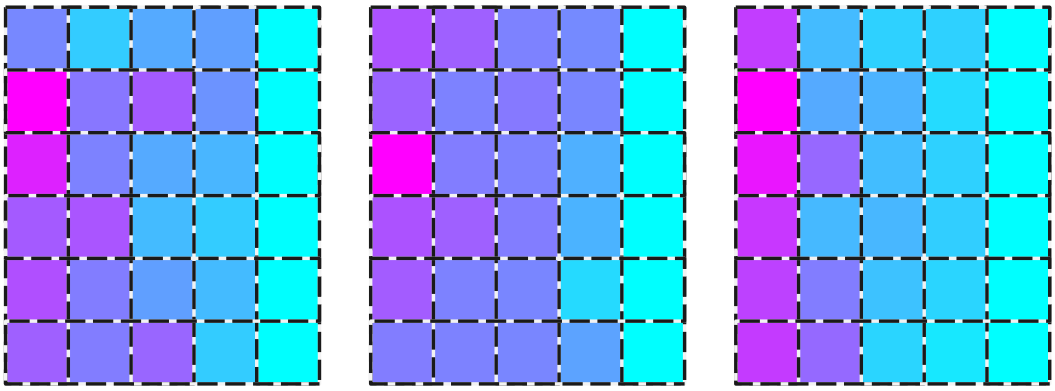
where X is the domain of x.

To study the physiological trajectory of the "normal" patients, the distributions of each vital sign, for each of the first 4 subgroups described (G_1, G_2, G_3, and G_4) were compared with G_5 using the three metrics defined by Eqs. (6.1), (6.2) and (6.3). The rationale for this was to consider the most physiologically stable period of the patient's stay on the ward, which should correspond to the period immediately before the patient was discharged from the ward. This most-stable state would then be used as a reference for all the other periods of time during the patient's stay.

Figure 6.3 shows the three metrics determined between each of the vital-sign data distributions, for the 4 subgroups (G_1, G_2, G_3, and G_4) and the distribution for the G_5 subgroup. In each map, the subgroups involved (G_{i-5}, with $i = \{1, 2, 3, 4, 5\}$) are represented on the x-axis, and the physiological variables are represented on the y-axis. The colour code is associated with the values of the calculated distances. The results obtained for the three metrics are very similar, in the sense that the patterns in the distances for each physiological variable are identical; e.g., the distances between G_5 and G_1 distributions (for most vital signs) are greater than the distances between G_5 and G_3 distributions. From these results, we can easily see the pattern of recovery with time: as the period of time gets closer to the period immediately before discharge, the distribution of the vital signs gradually becomes more similar to that of G_5.



(a) Legend Scheme



(b) ΔKS (c) ΔKL (d) $\Delta Bhat$

Figure 6.3: Representation of the distances between each of the 4 subgroups (G_1, G_2, G_3, and G_4) and the subgroup G_5 for each vital sign: (a) schematises the legend of the normalised metrics calculated using (b) the Kolmogorov-Smirnov, (c) symmetrical Kullback-Leibler, and (d) Bhattacharyya metrics. G_{i-5} represents the distance between subgroups G_i and G_5. The computed distances were normalised by dividing all values by the maximum value calculated for the correspondent metric (i.e., values close to 0 correspond to very similar distributions). The colour is associated with the values of the calculated distances, as shown in the colourbar in (a). See electronic version for a correct display of figure.

It is clear, from the results presented above, that there is a gradual change in the distributions of individual vital signs during the period of recovery from surgery. Given the significant correlation between the different physiological variables, it may be instructive to visualise the data in their D-dimensional space (in this case, $D = 6$).

6.1.2 Multivariate vital-sign data distribution

To investigate changes in the vital signs via projection of the underlying multivariate data distribution into a two-dimensional representation, the data were visualised using the NeuroScale algorithm (Tipping and Lowe [1998]). The mapping of a dataset onto a two-dimensional space for visualisation purposes is known in the literature as *multi-dimensional scaling*. A well-known example of such a dimensionality-reduction mapping is Sammon's mapping (Sammon [1969]), which seeks to find a configuration of image points in the two-dimensional visualisation space, such that the (Euclidean) distances d_{ij} between image points are as close as possible to the corresponding distances δ_{ij} in the original high-dimensional data space. Since it is not possible to find a configuration for which $d_{ij} = \delta_{ij}$, Sammon's mapping uses the following sum-of-squared-error criterion, E_S, for assessing the suitability of a configuration with respect to the others:

$$E_S = \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{\delta ij}. \tag{6.4}$$

Sammon's mapping aims to minimise E_S, which can be achieved by initialising the image points to have random locations in the two-dimensional space and iteratively adjusting these locations in the direction which gives the maximum change in E_S using gradient descent. The NeuroScale algorithm is an extension of Sammon's mapping, in which the mapping from the high-dimensional input space to the two-dimensional visualisation space is parameterised using an RBF neural network (Tipping and Lowe [1998]). This allows interpolation between points in the training set such that new points can be displayed using the map constructed from the training set.

The data from all five subgroups described above, $\mathbf{X} \in \mathbb{R}^6$, were visualised using the NeuroScale algorithm. Each variable was first scaled to have approximately the same dynamic range using the zero-mean unit-variance transformation described in earlier chapters (i.e., the data were standardised with respect to G_5). The RBF neural network was trained using data from G_5. The number of basis functions was selected (from 5 to 20 basis functions) in order to minimise

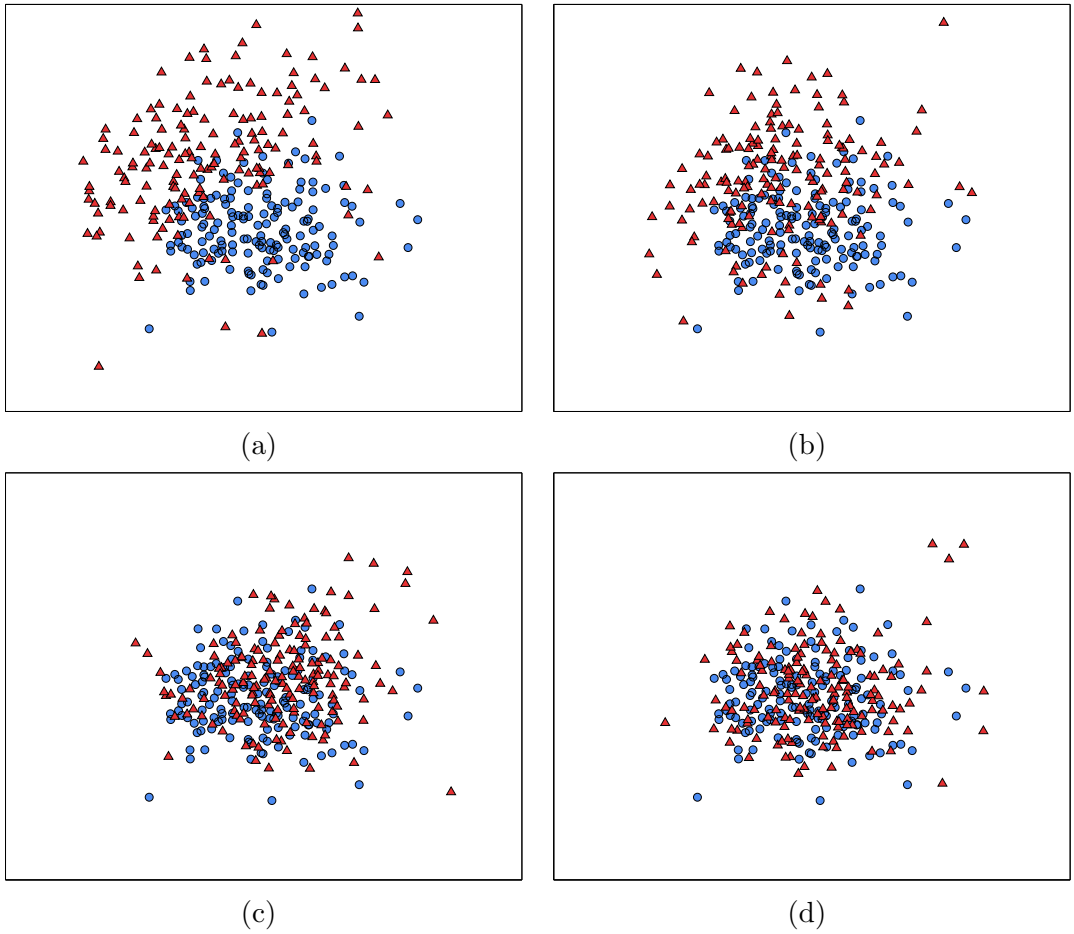


Figure 6.4: NeuroScale visualisation maps for the projected six-dimensional vectors of vital signs for the subgroups: **(a)** G_1 and G_5, **(b)** G_2 and G_5, **(c)** G_3 and G_5, and **(d)** G_4 and G_5. Projected data from subgroup G_5 are shown by blue '$\circ$'; projected data from all other groups are shown by red '$\triangle$'. Axes' tick marks and corresponding labels are omitted as the values in the projected space do not have a direct interpretation (but the same limits are used for the axes in the four plots).

the sum-of-squared-error (after initialisation using principal component analysis, PCA). The neural network was then used to project into the 2-dimensional space the 1485 normalised vectors contained in the five subgroups (G_1, G_2, G_3, G_4, and G_5) from the 297 patients that belong to the normal group.

The resulting maps obtained are shown in Figure 6.4. Represented in each map are the projected data points from G_1, G_2, G_3, and G_4 superimposed on the projected data points from G_5 subgroup. The maps show that these projected

data form clusters with some overlap between them, but that there are subgroups with visually separable distributions. The cluster G_1 is the most diffuse (shown with red $\triangle$ in Figure 6.4a), while the projected data from G_3, G_4 and G_5 are more concentrated, and similar to each other in their locus in the projection plane. This suggests that are no large changes in data distributions from halfway through a patient's stay to the time of their discharge from the ward. That is, normal patients appear to have stabilised at around halfway through their stay on the ward. These results suggest that patients included in the normal group could have been considered for earlier discharge, or provided with a lower level of care from halfway through their stay, as no major changes in their physiology occur during their last days on the ward.

6.2 Physiological trajectory

The overall trajectory of each physiological variable for the two groups of patients may also be evaluated considering the average value of each variable for each post-operative day. Figure 6.5 (left-hand column) shows the averaged values, for each day, of the six vital signs for the 357 patients in the normal group, and for the 50 patients in the abnormal group. These values are displayed for the length of stay (or time to first major adverse in the case of the "abnormal" group) up to the 75^{th} percentile (13 and 10 days, respectively) for each group of patients. For all patients, day 1 is the day on which surgery took place. As patients have varying lengths of stay, the patient's sample size used for each daily summary mean and standard error is provided in Figure 6.6. Each data point corresponds to data from an entire day. Thus, for example, 229 patients from the normal group had data available to calculate the average value on the first day post-operatively, and the values of each vital sign correspond to data collected between 00:00 and 23:59 on that day.

We note that post-surgical recovery lasts, on average, approximately four days (Figure 6.5, left-hand column). By day 5, for example, the systolic blood pressure of the normal group of patients has reached its steady-state range of values (between 120 and 130 mmHg). The systolic BP for the abnormal group of patients has a very similar trajectory to that of the normal group. In both cases,

the values are well within the bounds of what would be considered to be normality; i.e., if we consider a typical EWS system, a score of 0 (i.e., normality) for systolic BP extends from 100 to 170 mmHg (Gao et al. [2007]). The same pattern is observed for diastolic BP. We can also observe that there is no physiologically-significant difference in the values of RR between the two groups. The mean values for each day are within 1 rpm of each other throughout. The averaged HR values are mostly between 75 and 90 bpm, with a peak at around 93 bpm on day 6 for the abnormal group, which is caused by a few extremely high HR values for two abnormal patients that coincidentally occur on the sixth post-operative day for these patients. There is no clear distinction between the trajectories of the two groups regarding SpO_2. The overall trajectory may be explained by the fact that many patients are on oxygen support during the first two days after surgery. Finally, the mean temperature readings from day 1 to day 10 are broadly similar for both normal and abnormal groups.

Given the small number of patients in the abnormal group and the fact that the major adverse events do not occur all on the same day (see Figure 6.6), it may be argued that trajectory alignment according to the date of surgery is not suitable for representing the trajectories of abnormal patients. In fact, if we consider one patient who has an adverse event on day 5, and one patient who has an event on day 9, the latter may have normal physiology on day 5, masking the effect on the vital signs of the former. Hence, trajectories were aligned according to the day of event for all patients in the abnormal group. Figure 6.5 (right-hand column) shows the mean values, per day, of the vital signs for the six days that preceded the adverse event for the "abnormal" patients, with the last six days on the ward for the normal group (as a reference). We observe that on the last two or three days before the major event, abnormal patients exhibit gradually higher values of HR, RR, and temperature compared to those for the normal patients, while the differences observed for the other vital signs appear not to be as pronounced.

Thus far, we have considered the trajectory of each vital sign individually. We next propose a method to represent the overall physiological trajectory of these patients by considering the contribution of all vital signs. This provides a single representation of the physiological behaviour of post-operative patients.

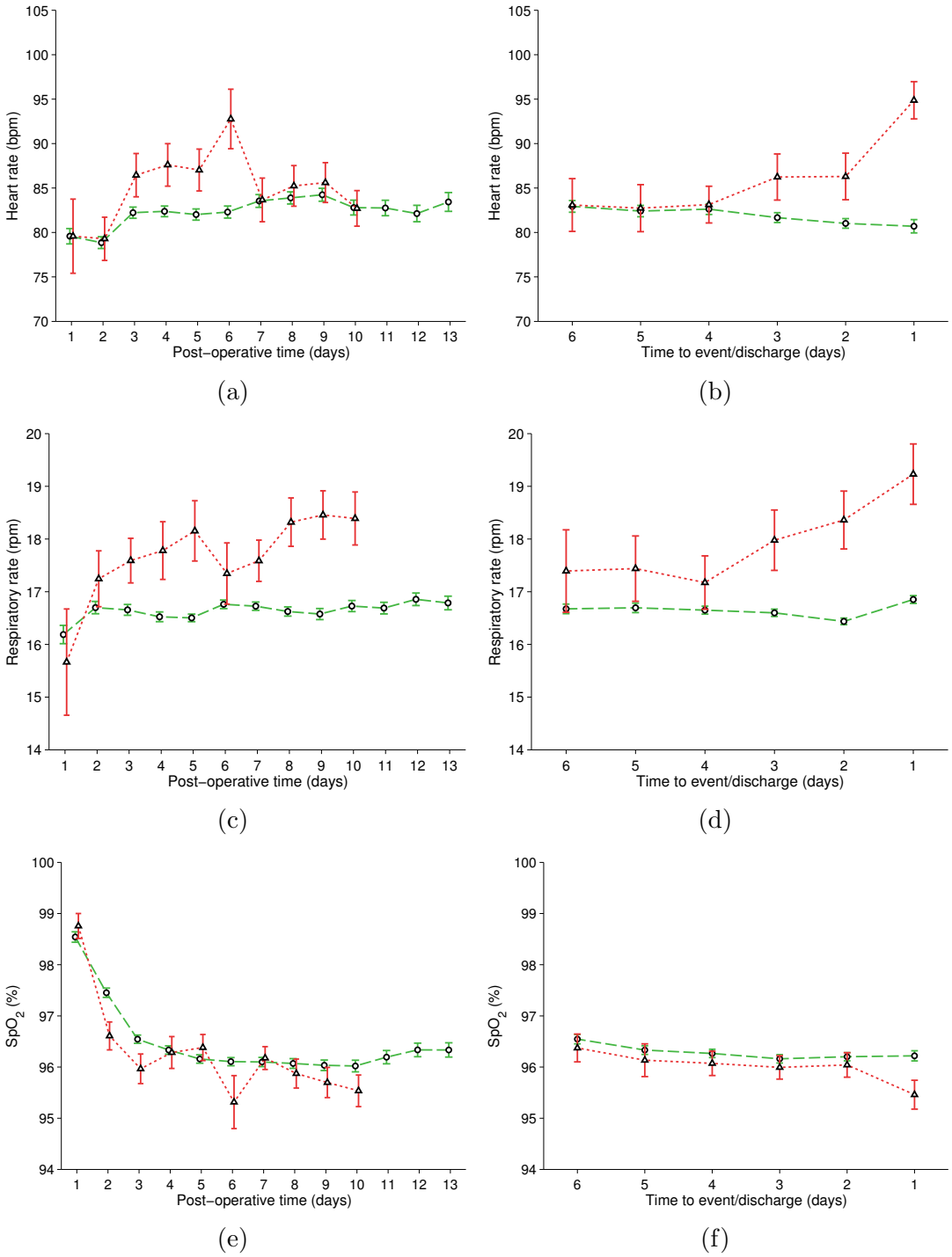


Figure 6.5: On the left, (a)-(c)-(e): the averaged values for the vital signs are shown for the first 13 days post-operatively, for patients from the normal group (shown in green, '○'), with the first 10 days of the patients from the abnormal group (shown in red, '△'); trajectories are aligned according to the date of surgery. On the right, (b)-(d)-(f): the averaged values for the vital signs are shown for the last 6 days before a major adverse event for the abnormal group, with the last 6 days on the ward for the normal group as a reference (see main text for further details). Error bars denote one standard error of the group mean. (This figure continues on the following page.)

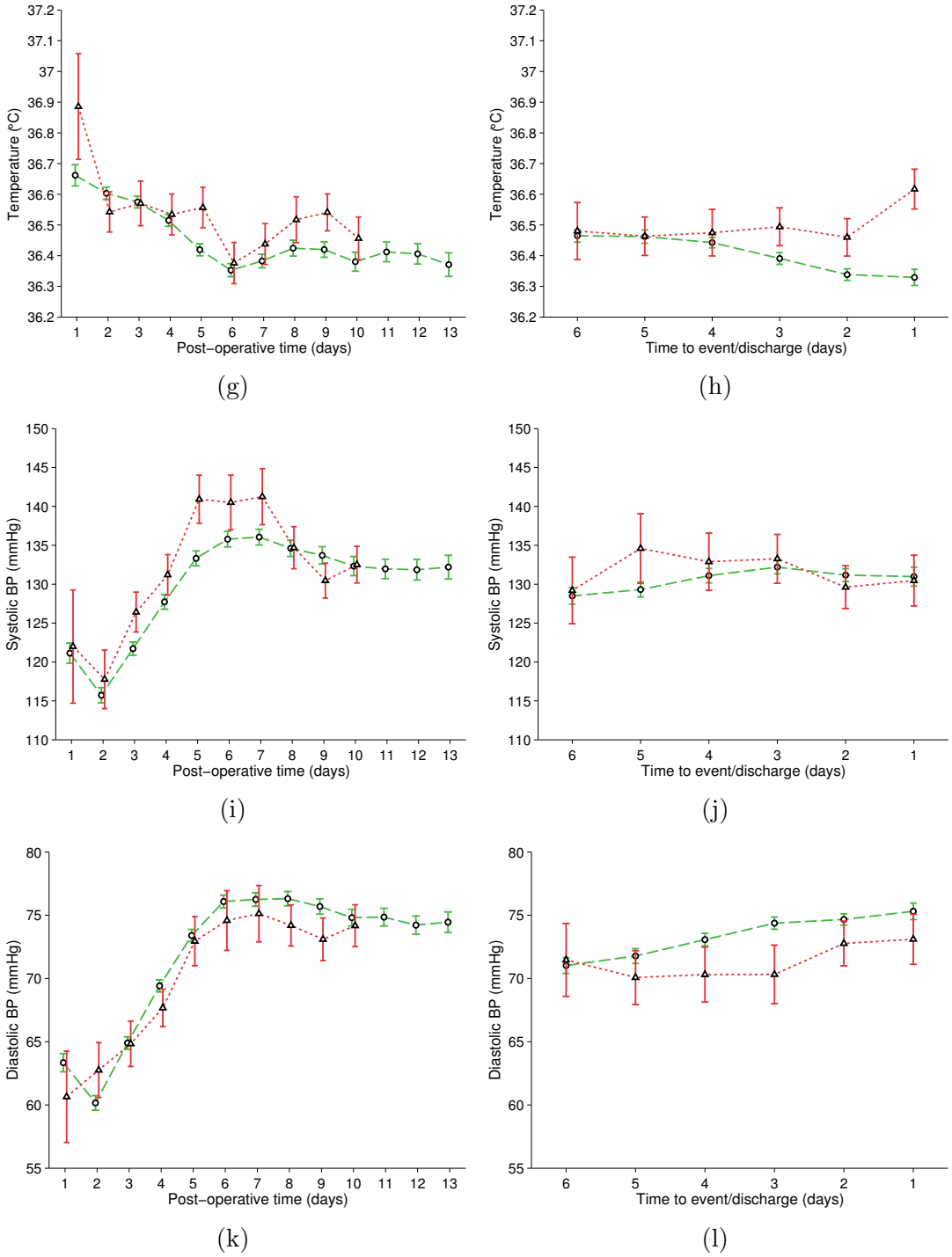


Figure 6.5: (Continuation) On the left, (g)-(i)-(k): the averaged values for the vital signs are shown for the first 13 days post-operatively, for patients from the normal group (shown in green, '○'), with the first 10 days of the patients from the abnormal group (shown in red, '△'); trajectories are aligned according to the date of surgery. On the right, (h)-(j)-(l): the averaged values for the vital signs are shown for the last 6 days before a major adverse event for the abnormal group, with the last 6 days on the ward for the normal group as a reference. Error bars denote one standard error of the group mean.

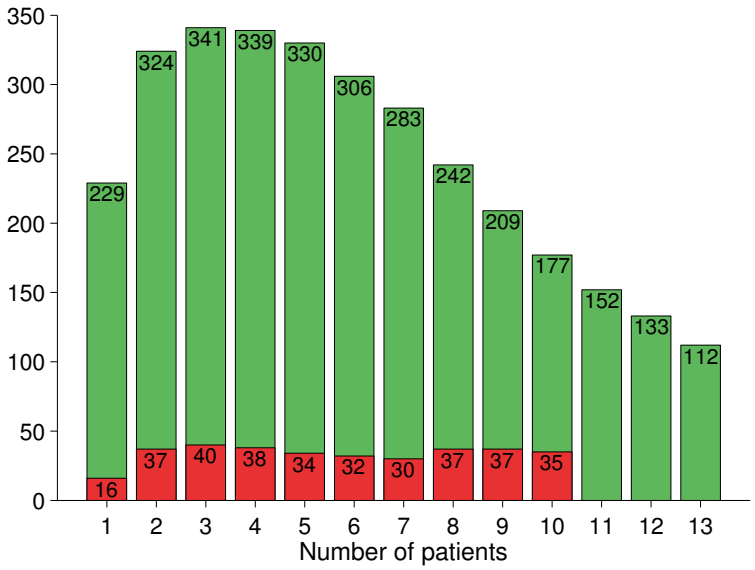


Figure 6.6: Number of patients with observations for each post-operative day for the "normal" (in green) and "abnormal" (in red) patients.

6.2.1 Multivariate model of normality

In order to represent the physiological trajectory of post-operative patients, we now consider the construction of a model of normality based on the vital-sign observations recorded on the last day on the ward (discharge day) for each patient in the normal group. This subset of data contains the vital signs from the most physiologically stable period of the patient population, because these data were acquired immediately prior to discharge from the ward, when patients are in their most "normal" state, following recovery from surgery. The rationale for selecting this subset of data is to create a reference dataset to which data from other periods of time (from other patients) may be compared. As in the previous analysis, for all patients, day 1 is the day on which surgery took place.

A KDE, which was introduced in the previous chapter (section 5.2.2), was used to estimate the pdf of the underlying D-dimensional pre-discharge vital-sign data, $\mathbf{X} \in \mathbb{R}^D$. For the analysis conducted in this experiment, we considered the five vital signs that were used in the models constructed in the last chapter ($D = 5$), which comprise HR, RR, SpO_2, temperature and systolic BP. All

variables included in each model of normality were standardised as before, using the zero-mean unit-transformation, using the mean and standard deviation values computed from the training data. The kernel width, σ, of the isotropic KDE was optimised by maximising the leave-one-out likelihood of the training data (as described in section 5.4.1).

6.2.2 Multivariate physiological trajectory

To represent the physiological trajectory of post-operative patients, the likelihood for all data recorded for each patient was calculated with respect to the model of normality. In order to estimate "abnormality" of the test data $\mathbf{x}'$, the departure from normality was quantified using the novelty score $z(\mathbf{x})$ defined as before, $z(\mathbf{x}) = \log\left(\frac{1}{p(\mathbf{x}|\boldsymbol{\theta})}\right)$, where $p(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood of the test point $\mathbf{x}$, and $\boldsymbol{\theta} = \{\mathbf{X}, \sigma\}$. "Normal" data, which have higher likelihoods $p(\mathbf{x}|\theta)$, therefore generate low novelty scores $z(\mathbf{x})$; conversely, "abnormal" data, which have lower likelihoods, generate high novelty scores $z(\mathbf{x})$.

As before, five-fold cross-validation was used with the CALMS-2 database, as illustrated in Figure 6.7, with $K = 5$. We note again that only a subset of vital-sign data from normal patients (corresponding to the last 24 hours on the ward) is included in the training set; the distinction between normal patients used for training and normal patients used for validation is kept during this analysis.

6.2.3 Results and discussion

The novelty scores $z(\mathbf{x})$ were computed for each day for each patient, by averaging the scores of the observations sets performed on that day. The group mean novelty scores for each day are shown in Figure 6.8 for patients in the normal group in both training and test sets. This is, it presents the physiological status of normal patients as quantified by the novelty score. The novelty scores are displayed for the length of stay up to the 75^{th} percentile for the normal group of patients (Table 6.1). As patients have varying lengths of stay, and due to the slightly different sample sizes considered in each fold of cross-validation (because folds are defined by whole patient time-series), refer to Figure 6.6 for the overall sample size used for each daily summary mean and standard error. We observe

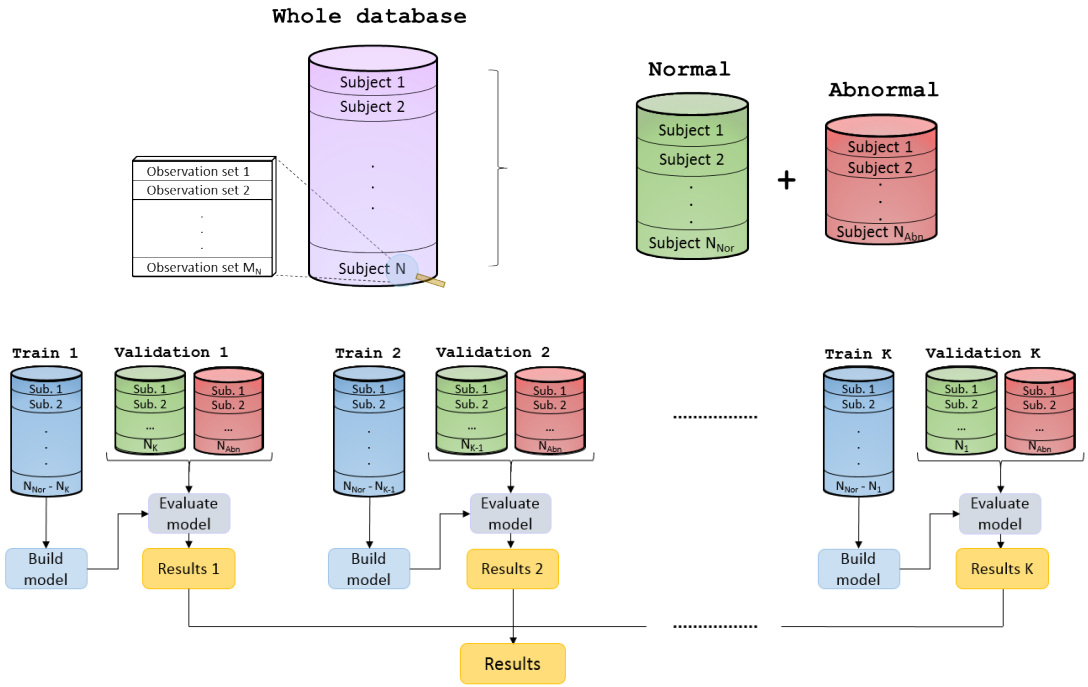


Figure 6.7: Schematic representation of the CALMS-2 data partitioning method used for evaluating the performance of the models studied. We note that this scheme is slightly different from the schemes present in previous chapters, as the Portsmouth dataset is not used in this chapter.

that there is no marked difference between the normal patients included in the training and validation sets.

More significantly, we observe a gradual decrease in the novelty score in the first few days after surgery, which stabilises after day 5. We can see a pronounced decrease in the novelty score in the first 5 days, after which $z(\mathbf{x})$ is approximately constant for $t \geq 6$ days. This provides an interesting summary of the typical physiological trajectory for post-operative patients. Patients have a relative high initial physiological derangement following major surgery. A clear return to normality (decrease in the physiological novelty score) is then exhibited as a result of the patient's recovery on the ward. This trend continues to stabilise after day 5, albeit at a much reduced rate. It could be argued that the majority of these patients are sufficiently stable for early discharge to be considered, or for them to be provided with a lower level of care, should they need to remain in hospital

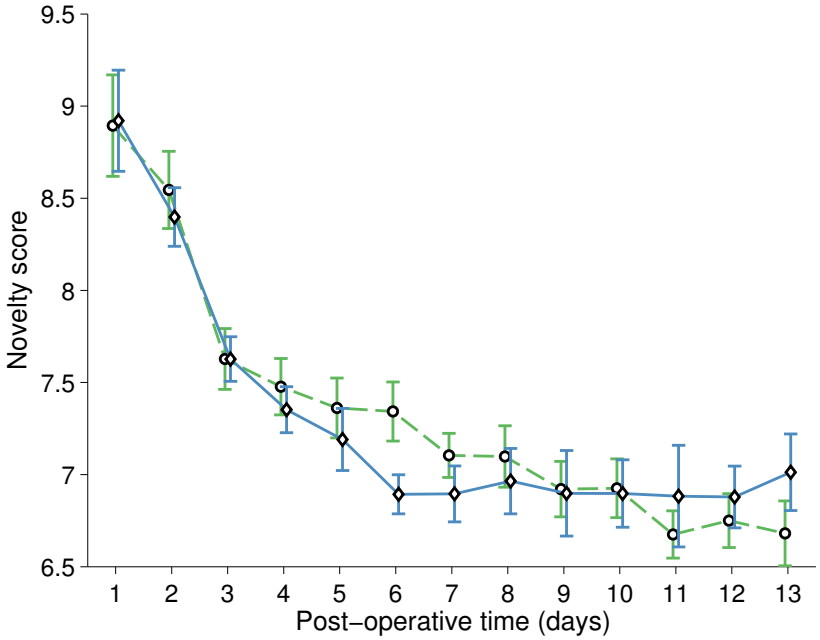
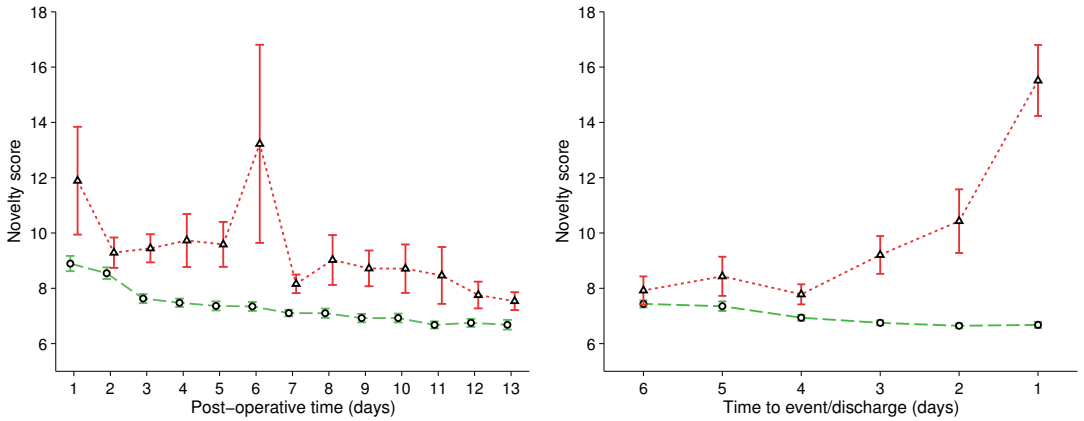


Figure 6.8: Representation of average (per day) of novelty scores $z(\mathbf{x})$ against post-operative time for the normal group of patients included in the training (blue-solid line) and validation (green-dashed line) sets. The results are averaged over five different cross-validation folds: the mean values are shown as the averaged mean values over the five folds, and the error bars denote the averaged standard error of the group mean over the five folds.

for reasons not related to physiological instability.

The group mean novelty scores for each day for patients in both normal and abnormal groups are shown in Figure 6.9. As in the previous case, we display these trajectories aligned according to the day of surgery (Figure 6.9a), and aligned according to the day on which the major adverse event for the abnormal patients took place (Figure 6.9b). The $z(\mathbf{x})$ values for the abnormal group of patients (Figure 6.9), suggest that the physiological trajectory for these patients is remarkably different to that of normal patients, with a sudden increase in novelty in the last 48 hours (Figure 6.9). On day 6 (Figure 6.9a), we observe a peak in the novelty score which is caused by the extremely high HR values of two "abnormal" patients that happen to occur on their sixth post-operative day (as already mentioned in section 6.2). These results suggest that patients' criticality could



(a) (b)

Figure 6.9: **(a)** The averaged values for the novelty scores are shown for the first 13 days post-operatively, for patients in the validation set from the normal group (green '∘'), with the first 10 days of the patients from the abnormal group (red '△'). **(b)** The averaged values for the novelty scores are shown for the last 6 days before a major adverse event for the abnormal group, with the last 6 days on the ward for the normal group as a reference. The results are averaged over five cross-validation folds: the mean values are shown as the averaged mean values over the five folds, and the error bars denote the averaged standard error of the group mean over the five folds.

be assessed by evaluating the change in the distribution of their vital signs after their admission to the post-operative ward, following major surgery. They also suggest that physiological deterioration may be identified up to 48 hours before the occurrence of the major adverse event.

Figure 6.10 shows the averaged physiological trajectories for the "normal" patients (in the validation sets) having removed one of the five vital signs from the model of normality (i.e., $D = 4$). One could argue that the trajectory observed in Figure 6.8 is mainly due to the change of, for example, systolic BP alone. Nevertheless, if one of the vital signs is removed from the model of normality, the overall pattern of the physiological trajectory still holds, as no pronounced difference in the overall pattern of the physiological trajectory is observed for the different models.

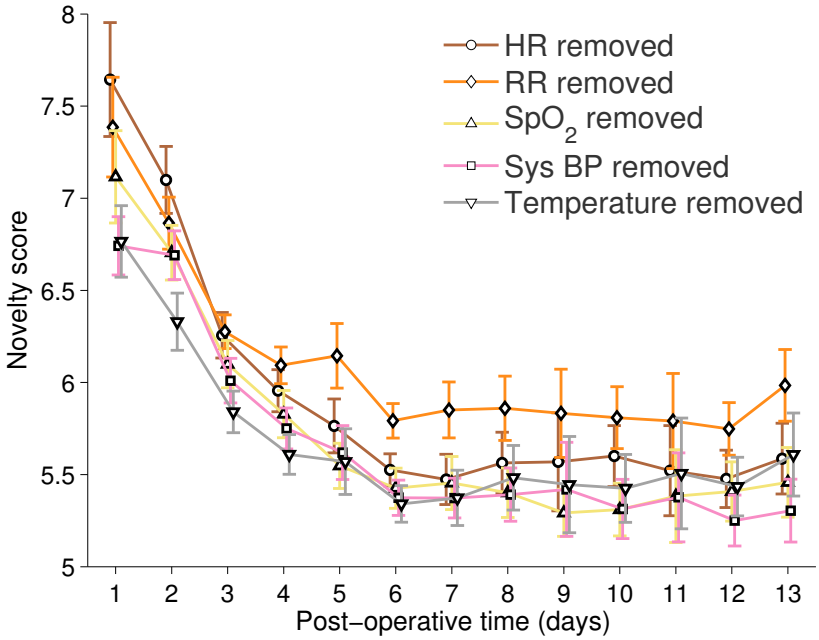


Figure 6.10: Representation of average (per day) of novelty scores $z(\mathbf{x})$ against post-operative time for the normal group of patients, computed for different models of normality by excluding one of the vital signs. The results are averaged over five folds: the mean values are shown as the averaged mean values over the five folds, and the error bars denote the averaged standard error of the group mean over the five folds.

6.3 Physiological variability

As with the majority of current scoring systems, the KDE described in the last section relies on the absolute values of the vital signs as measured by the clinical staff. In Figure 6.5, we showed that the mean values of the vital signs, for each day, are often similar for the two groups of patients (with values that are within the accepted range of physiological normality). Yet, most of the patients in the abnormal group will have shown evidence of abnormal physiology at some point. Anecdotal observations led us to postulate that patients with high variability in their vital signs may have poor outcomes.

Several studies have been conducted on the variability of a specific subset of vital signs. In particular, heart rate variability has been investigated as an indicator of cardiovascular and autonomic system function, and also as a predictor

of patient outcome (Liu et al. [2014, 2011]; Norris et al. [2005, 2008]; Stys and Stys [1998]). Most work in HR variability relies on computing beat-by-beat variability, which require continuous waveform data. Grogan et al. [2005] investigated the statistical variability of HR, which the authors termed *volatility*, in a population of ICU patients, and showed that measures of volatility (which included the standard deviation of HR over the entire patient's stay in the ICU), rather than measures of central tendency (such as the mean or median), were more predictive of hospital mortality. Moreover, most severity of illness scores used in the ICU include some measure of variability of certain vital signs over a given period of time (Mayaud [2014]). Nevertheless, to the best of our knowledge, there are no studies in the literature which explored the variability of vital signs of patients outside the ICU.

In this chapter, we now propose a new measure of variability that is based on routine observations performed by the clinical staff (rather than requiring waveform data), and we investigate the variability of all physiological variables as a potential indicators of physiological deterioration.

6.3.1 Computing the variability index

We hypothesise that abnormal physiology may be characterised by an abnormal variation of the physiological variables about their mean. We define the *variability index* to be the difference between the maximum and minimum values in a 24-hour period, for each physiological variable. Figure 6.11 shows how a 24-hour sliding window is used to compute the variability index Δ at the time of each observation set. For each observation, the corresponding variability index Δ is calculated using all the observations recorded during the 24 hours up to that observation. We require a minimum of four observations in each 24-hour window; if there are fewer than four observations, the variability index for the current observation is taken to the be the variability index of the previous observation (Figure 6.11d).

The computed variability indices for each physiological variable are shown for both abnormal and normal groups of patients in Figure 6.12[1]. It is clear that

[1]The variability indices for each patient can only be computed after 24 hours of data have

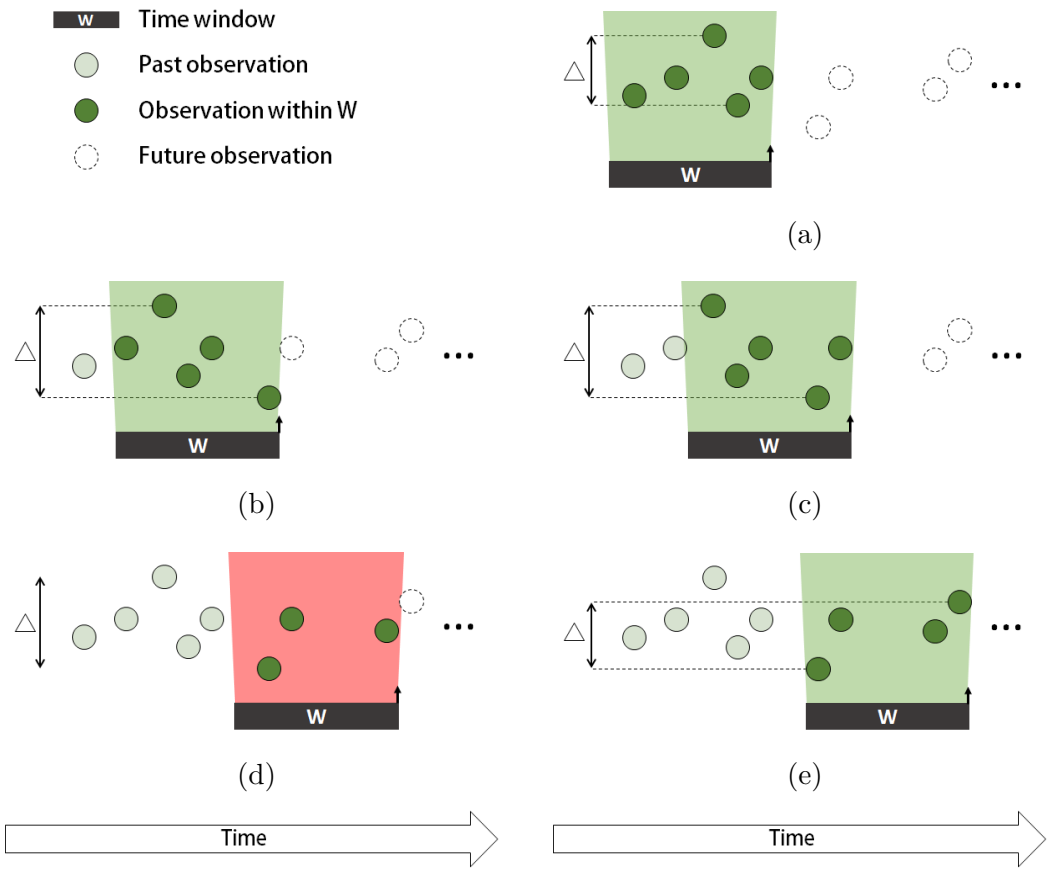


Figure 6.11: Method used to determine the variability index (denoted by Δ) for each vital sign using a sliding window (W). For each observation time (denoted by the small black vertical arrow at the end of the window W), all observations recorded within W (shaded region) are used to determine the variability index for the observation indicated by the small arrow. If only three observations (or fewer) were made during the 24-hour period preceding the current observation, the variability index is taken to be that of the previous observation; e.g., Δ at **(d)** is the same as Δ at **(c)**.

variability is generally higher for normal patients, for all physiological variables, for the first two to four days following major surgery. After this, variability decreases as the process of recovery from surgery takes place. The most relevant plots in Figure 6.12 for the abnormal group of patients are those for RR, HR, temperature, and BP. We observe that the variability index for these variables

been acquired; for graphical representation here, the variability index for the observation sets that were recorded during the first 24 hours for each patient were assumed to be the same as the first variability index calculated for that patient.

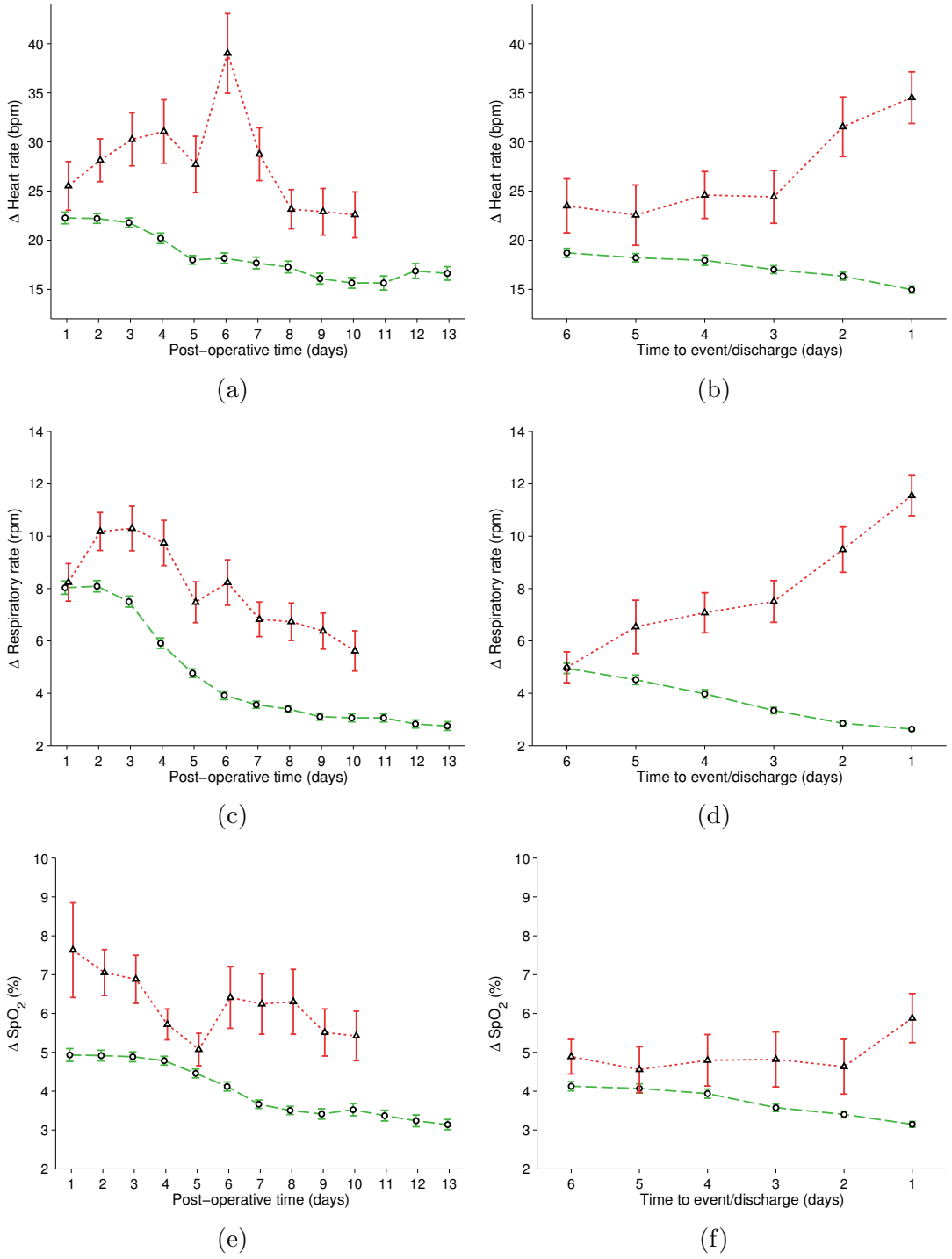


Figure 6.12: On the left, (a)-(c)-(e): the averaged values for the vital sign variability are shown for the first 13 days post-operatively, for patients from the normal group (green '○'), with the first 10 days of the patients from the abnormal group (red '△'); trajectories are aligned according to the date of surgery. On the right, (b)-(d)-(f): the averaged values for the vital sign variability are shown for the last 6 days before a major adverse event for the abnormal group, with the last 6 days on the ward for the normal group as a reference. Error bars denote one standard error of the group mean. (This figure continues on the following page.)

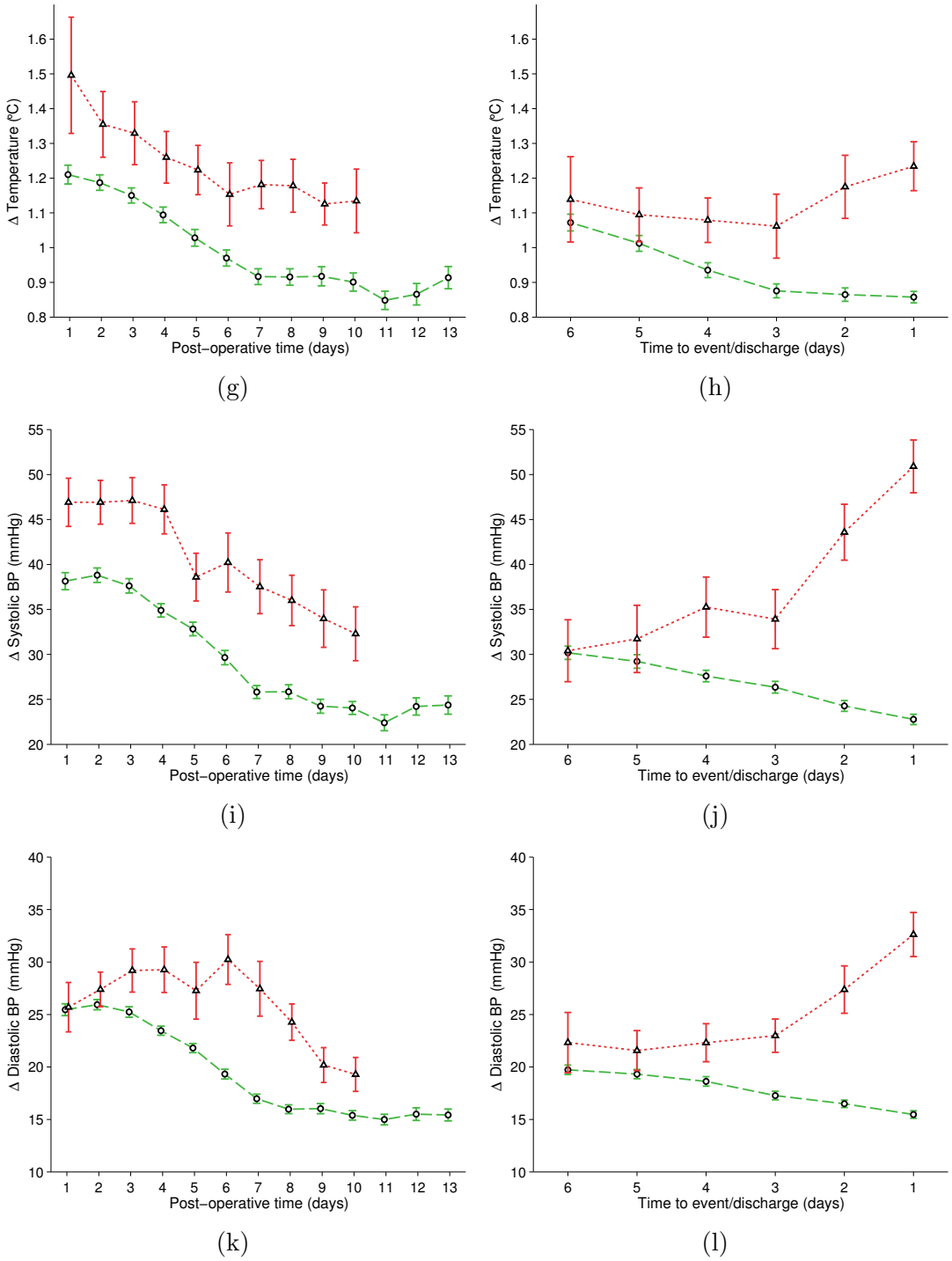


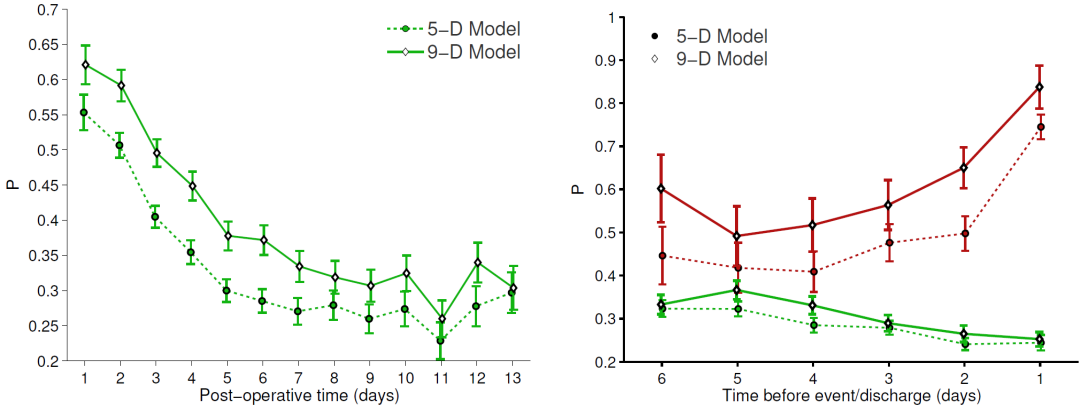
Figure 6.12: (Continuation) On the left, (g)-(i)-(k): the averaged values for the vital sign variability are shown for the first 13 days post-operatively, for patients from the normal group (green '○'), with the first 10 days of the patients from the abnormal group (red '△'); trajectories are aligned according to the date of surgery. On the right, (h)-(j)-(l): the averaged values for the vital sign variability are shown for the last 6 days before a major adverse event for the abnormal group, with the last 6 days on the ward for the normal group as a reference. Error bars denote one standard error of the group mean.

for the latter is consistently higher than that for normal patients. In particular, we observe that in the 48 hours that precede a major adverse event, the daily averaged values of the variability indices are much higher than those for the normal group of patients at any point during their stay. It is difficult to interpret the results for SpO_2 because a substantial proportion of patients in each group will have been on oxygen support. To summarise, the 24-hour variability indices for HR, RR, temperature, and systolic BP during the post-operative period are higher for the abnormal patients, and therefore, we hypothesise that these indices predict major adverse events for post-surgical patients.

In order to study the trajectory of the vital signs combined with information about physiological variability we built on the approach described in the previous section. We have previously considered the construction of a model of normality based on the vital-sign observation sets made on the last day on the ward for each patient in the normal group. We now consider a new KDE, for which the feature vectors are now augmented to include the four variability indices ($D = 9$). We compare this to the model described earlier, which uses only the vital signs ($D = 5$) before discharge.

Likelihoods $p(\mathbf{x})$ depend on the dimensionality of the space over which they are defined, as do novelty scores $z(\mathbf{x})$ based on those likelihoods. Hann [2008] addressed the problem of comparing novelty scores from models defined over spaces of differing dimensionality by proposing a numerical approach. This method finds the probability mass P enclosed by level sets of the likelihood using sampling. This enables us to compare the values of novelty scores from the 5-D and 9-D models as described in Appendix C.

As in the previous section, we determined the likelihood of all observations sets of all patients in the validation set with respect to each model of normality, and quantified the departure from normality using the novelty score $z(\mathbf{x})$. The results are shown here for one representative fold of cross-validation; we conducted experiments for a single fold because no significant variation between the different folds was observed in the previous section.



(a) (b)

Figure 6.13: **(a)** The averaged values for the novelty scores (mapped to P) are shown for the first 13 days post-operatively, for patients from the normal group in the validation set using each model. **(b)** The averaged values of P are shown for the last 6 days before a major adverse event for the abnormal group (in red), with the last 6 days on the ward for the normal group as a reference (in green). Dashed lines correspond to P computed using the 5-D model. The results correspond to those obtained in one of the five cross-validation folds. Error bars denote one standard error from the group mean.

6.3.2 Results and discussion

The novelty scores $z(\mathbf{x})$ computed using the two different models of normality (5-D and 9-D models), after mapping to the probability P scale, are shown in Figure 6.13 for the normal and abnormal patients in the validation set. Values are represented as the group mean for each day.

If we compare the trajectories obtained with each model, we can see that the difference between the normal and abnormal trajectories computed with the 9-D model is generally higher than that computed with the 5-D model. In fact, while the pattern of recovery appears to be equally accentuated when the 24-hour variability indices are included in the model (i.e., there is a significant decrease in P for the normal group with respect to the value on day 1), we observe a very pronounced increase in the novelty score (mapped to P) in the last 48 hours, pointing to "abnormal" variability indices prior to a major adverse event. We also observe that, if we consider the last days prior to an adverse event

(Figure 6.13b), a more clear distinction between the overall trajectory of the two groups of patients is obtained when computed with the 9-D model.

These results suggest that the 24-hour variability indices of RR, HR, temperature and systolic BP, may improve the early identification of patient deterioration.

6.4 Identifying patient deterioration

Knowledge of the typical post-operative patient recovery trajectory promotes the design of early warning systems based on the patient's changing physiology during recovery from surgery on the ward. Alerts may be generated whenever the patient's recovery does not follow the expected trajectory, i.e., the novelty score is higher than expected for that day post-operatively. Nevertheless, the incorporation of the information about this "trend" in the vital signs and the variability of vital signs may not be straightforward. Varying thresholds for each post-operative day could be used, but this would require a much larger dataset for the purpose of cross-validation to set the threshold values for each day. It might be considered useful to include the time of observation as an extra feature in the model. However, such approach has at least one major limitation. If we consider trajectories aligned according to the date of surgery (as before), all patients will contribute with observation sets for the first couple of days, while far fewer patients will contribute with observations for the following days (as they are gradually discharged from the ward). This will create the effect of having a higher score for longer stayers on the ward; e.g., the novelty score on day 7 may be higher than that on day 3, not because of the "abnormal" physiology of the patient, but because of the presence of more observations sets (higher data density) on day 3 during the training of the model. An alternative solution must be applied.

In order to incorporate the information about vital-sign trajectories and variability indices into models of physiological normality, we introduce the concept of *time-based normalisation*.

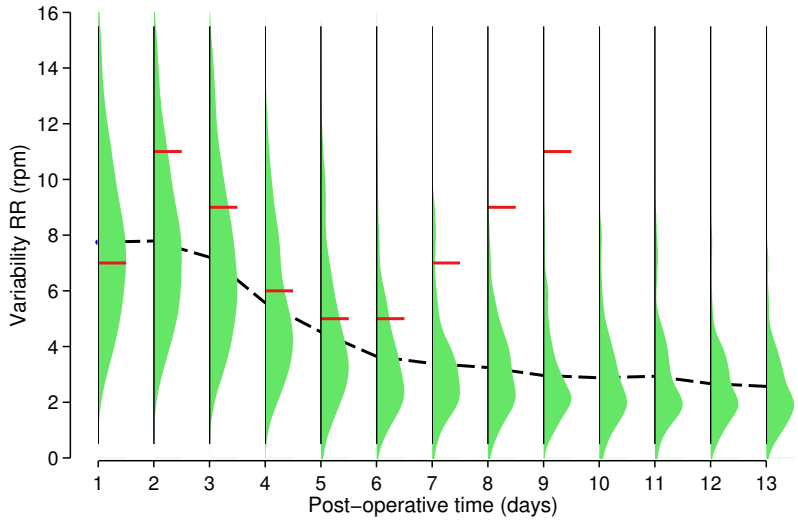


Figure 6.14: Representation of the distributions of the 24-hour variability indices for each post-operative day for the normal group of patients along the vertical axis (green regions), with the mean of each distribution represented by the dark-gray dashed line. The variability indices for RR of one patient from the abnormal group, averaged per day, are shown with horizontal red thick lines.

6.4.1 Time-based normalisation

As with other models described in the previous chapter, we assume *a priori* that each physiological variable (including the four variability indices) has equal importance in the patient model of normality. Therefore, each variable is scaled so that variables with large dynamic ranges do not dominate variables with smaller ranges. Hence, every vital sign x is standardised using a zero-mean unit-variance transformation, as before, using the mean and standard deviation values computed from the training observation sets.

However, from the results presented in the previous sections, we observe that the mean values of the vital signs change with time; for example, most patients exhibit low systolic BP and high temperature immediately after surgery. To a first approximation, steady state for both systolic BP and temperature can be assumed to be reached five days after the operation. More significantly, we observe that normal values of variability change with time. In order to illustrate the importance of this point, consider the example in Figure 6.14. The distributions

of the 24-hour variability of RR for each post-operative day for patients in the normal group are plotted along the vertical axis. The red horizontal lines correspond to the RR variability index for one patient in the abnormal group, averaged per day, who was admitted to the ICU after 9 days on the ward. The value of the variability index for this patient on both days 3 and 8 is 9 rpm While this value looks normal with respect to the distribution for the normal group of patients on day 3, it is in the tail of the distribution for the normal training data on day 8 and hence it is highly "abnormal" for that day. This leads us to propose a time-based normalisation scheme whereby each variable is normalised using a zero-mean unit-variance transformation *for each day*, $x_n^j = (x^j - \mu^j)/SD^j$, where x^j is the value observed on day j, x_n^j is the corresponding normalised value, and μ^j and SD^j are the mean and standard deviation (respectively) for parameter x for day j, computed from the (normal) training data. After day 5, many patients have already been discharged from the ward. Hence, there are not enough data from day 5 onwards to perform day-based normalisation, and so, we group all data from day 5 onwards together for this normalisation process; i.e., $j \in \{1, 2, 3, 4, 5, > 5\}$.

It is important to note that the variability indices can only be computed after 24 hours of data have been recorded. Therefore, before normalisation is carried out, the variability indices corresponding to the observations performed during the first 24 hours for each patient on the ward are assumed to be the group average of the variability indices computed from the training data for the corresponding day. That is, if a patient is admitted on day 1, the variability index for HR, for example, can only be calculated on day 2 (in order to take into account the observations of HR recorded during the previous 24 hours). Hence, the variability index of HR for the observations recorded on day 1 is set to be the mean variability index of HR for day 1 computed from the training data.

6.4.2 Models considered

To test this approach, we constructed models that are based on the machine learning methods used in the previous chapter. In particular, we used the kernel density estimates approach for mixed data (in which the smoothing parameters of the model were optimised by maximising the leave-one-out likelihood of the

training data)[1].

Observational data were pre-processed in order to remove artefactual vital-sign values from the observation sets (as performed in the previous chapters). The variability indices for HR, RR, SpO_2 and systolic BP were calculated using the approach illustrated in Figure 6.11. The variability indices for the first day of each patient on the ward were set to be the mean of the variability indices calculated using the training data for that day. Each variability index was normalised according to the daily-based normalisation approach described above, while the other variables were normalised using the standard (global) zero-mean unit-variance transformation. This model, KDE_{var}, which includes the variability indices, was compared to a model that does not include the variability indices (KDE).

The performance of each model for identifying the derived composite outcome of death, emergency ICU admission and cardiac arrest (within 24 hours after an observation set) was evaluated during a five-fold validation procedure using the same sets as in the previous sections. Additionally, in order to compute different metrics of performance, for each model, a set of thresholds was optimised on the training split, by selecting target values for the false alerting rate; i.e., we selected the threshold based on the proportion of observation sets in the training set wrongly classified as abnormal. The performance metrics used to evaluate each model on the validation set include the false positive rate (i.e., the number of normal observation sets incorrectly classified as abnormal), the true positive rate in terms of patients (i.e., the number of patients in the abnormal group who had, at least, one abnormal observation set in the 24 hours before the major adverse event correctly classified as abnormal), and the time-to-event, which corresponds to the time of the first alert generated by each model in the 36 hours preceding the adverse event (reported as the median value for each validation set).

6.4.3 Results

The results of this experiment are presented in Table 6.3. This shows the different models evaluated (first row), and the values of the various performance metrics

[1]For further details on the optimisation of the parameters of the model, refer to section 5.4.1.

Table 6.3: Performance of the different models tested, together with those given by the ViEWS (Prytherch et al. [2010]) and NEWS (RCP [2012]) scoring systems, for detecting a major adverse event. The results are extracted from five test folds and the values are presented with mean and standard deviation; performance metrics are: AUROC for identification of patient deterioration within 24 hours of the observation set, the false positive rate (FPR) for each threshold selected during the training procedure, the true positive rate (TPR) calculated as the percentage of major adverse events identified by the system during the 36 hours before the event, and the time-to-event (TTE) for those events which were identified.

	ViEWS	**NEWS**	**KDE**	**KDE$_{var}$**
			AUROC	
	0.837 (0.003)	0.829 (0.005)	0.839 (0.004)	0.856 (0.003)
Threshold			**FPR, % of "normal" observation sets**	
25% [5,4]	24.6 (0.9)	30.0 (1.2)	25.5 (2.3)	26.6 (1.6)
20%	-	-	20.5 (1.8)	21.6 (1.3)
15% [6,5]	14.1 (0.6)	16.8 (0.6)	15.0 (1.5)	15.9 (0.9)
10%	-	-	9.5 (0.8)	8.8 (0.1)
5% [7,6]	6.6 (0.3)	8.2 (0.2)	5.2 (0.6)	5.6 (0.2)
			TPR, % of events	
25% [5,4]	91.1 (-)	92.9 (-)	94.6 (0.0)	94.6 (0.0)
20%	-	-	94.6 (0.0)	94.6 (0.0)
15% [6,5]	87.5 (-)	89.3 (-)	92.9 (0.0)	94.6 (0.0)
10%	-	-	88.7 (0.0)	89.3 (1.0)
5% [7,6]	78.6 (-)	82.1 (-)	82.1 (1.0)	83.3 (0.0)
			TTE, hours	
25% [5,4]	32.9 (-)	35.0 (-)	31.7 (2.1)	30.5 (0.7)
20%	-	-	28.9 (0.9)	29.6 (0.6)
15% [6,5]	23.7 (-)	23.8 (-)	25.4 (1.5)	27.9 (0.3)
10%	-	-	21.9 (0.1)	22.9 (0.4)
5% [7,6]	20.7 (-)	20.7 (-)	20.8 (0.1)	24.9 (0.3)

[1]The values within squared brackets, $[A, B]$, correspond to the thresholds for the ViEWS and NEWS scoring systems that gave the closest results to those for the other models in terms of false positive rate in the training set.

estimated on the test sets during the five-fold cross-validation procedure. A small variation between the models obtained in each fold was obtained (as reflected by the small standard deviation values).

6.4.4 Discussion

The results in Table 6.3 show the superiority of the KDE_{var} model with respect to the KDE model considering the various performance metrics. An AUROC value, mean (SD), of 0.856 (0.003) was obtained for this model, which includes the physiological variability indices. We observe that the KDE_{var} model, for the same false alerting rate, is slightly more sensitive than the KDE model, as the true positive rate values produced by the KDE_{var} model are just above those produced by the KDE model (depending on the threshold used). The model built using the kernel density estimates which includes the variability indices as inputs has an averaged true positive rate of 89.3% for a false positive rate of 10%. The equivalent KDE model without the variability indices has a lower true positive rate of 88.7%. Hence, the introduction of variability indices allows more "abnormal" patients to be detected without significantly increasing the false positive rate. Moreover, the KDE_{var} model has higher time-to-event values. Overall, the difference between the time-to-event values generated by the KDE_{var} model and those generated by the KDE model can be as high as 4 hours. This suggests that the first alert occurs earlier with the KDE_{var} model, and therefore, physiological deterioration is detected sooner. These results support the hypothesis that the 24-hour variability indices may help to identify deterioration earlier, and hence, predict a major adverse event earlier (emergency ICU admission or death on the ward).

We also observe that the models built outperform the recently proposed EWS scoring systems. An important observation is that the data-fusion models allow for small changes in vital signs, unlike EWS scoring systems, which provide coarse estimates of vital-sign abnormality, as the scores may only take integer values. The consequence of the latter can be seen in Table 6.3. For example, a score of 7 in ViEWS (Prytherch et al. [2010]) would trigger for 6.6% of the observation sets, and the score immediately below that, a score of 6, would trigger for 14.1% of the

observation sets, which would represent a significant increase in the clinical staff's workload. Furthermore, we note that the performance of the model trained with the CALMS-2 dataset, 0.829 (0.005), is higher than that of the model trained using the same method but with the Portsmouth dataset, 0.788 (0.009), as reported in the previous chapter. This is an expected result as the data used for training and testing the model were acquired from the same patient population; thus, there are no significant effects of dataset shift or domain shift that were discussed in the last chapter.

A few points should be made here regarding the analysis conducted and the overall results obtained in the analysis described in this chapter. In the first place, we considered all observation sets for developing the models. Recent studies have used discrete-time analysis to develop classification models (Churpek et al. [2014]). These methods typically involve separating time into discrete intervals and using the predictor variable values nearest to each time interval cut-off to predict whether the combined outcome occurred within that time-block. While this method may cope with unevenly-sampled data by considering discrete intervals and holding data when certain variables are missing, this is unlikely to produce significant changes in the results obtained here. Moreover, the selection of the discrete-time intervals may be problematic, and the method may discard information that is important for detecting periods of physiological abnormality.

It is also important to note that we did not include the diastolic BP in our models. Most early warning scoring systems do not include this variable, and report that the systolic BP is a more reliable physiological variable to model the change in a patient's condition. As noted in previous studies (Prytherch et al. [2010]; Wong [2011]), systolic and diastolic BP can be combined into a single variable (such as the mean arterial pressure, or pulse pressure), but it is not clear which combination should be used. The simple addition of diastolic BP to the latest models built did not improve the performance metrics calculated, which reflects the results reported in the literature (Prytherch et al. [2010]).

Regarding the computation of the variability index for each physiological variable, we observe that the variability index proposed is computed using the periodic observations performed by the clinical staff on the ward. The period of time over which the variability indices are calculated (24 hours) was not selected

arbitrarily. The rationale for selecting 24 hours was two-fold: (1) the period of time selected includes observations performed during a full post-operative day, which accounts for the expected daily variation of the vital signs (Rocha et al. [2011]); and (2) the number of observations performed within that period should be sufficient to compute a variability index (at least four observations are needed in our models). Shorter periods of time could be considered if data were more frequently recorded on the ward, which may be possible using the continuous data acquired with bedside and telemetry monitors.

Finally, we also note that the computation of the variability indices may become problematic in noisy datasets; i.e., if some of the observations have extremely high (or extremely low) values due to noise or other artefacts, the variability index for the corresponding period will be incorrectly assigned. However, because the dataset used in this study comprises only observational data recorded by nursing staff, data were subjected to a "human filtering" process, which makes it likely that artefactual data have been discarded. In the presence of artefactual data, a more robust metric (such as the variance or the interquartile range) may be used to mitigate the effects of these outliers in the computation of the variability indices.

6.5 Conclusion

This study has shown that data-driven modelling of physiology can effectively quantify patient status during the period when patients are recovering from major surgery. A multivariate model of the distribution of vital-sign data from "normal" patients was constructed using a kernel density estimator, and tested using "abnormal" data from patients who deteriorated after surgery. Important differences were found between the physiological trajectories for "normal" patients and those for "abnormal" patients. We further introduced the concept of 24-hour variability for each vital sign and showed that during the first days after surgery variability indices may help to predict earlier a major adverse event for post-operative patients. There have been several reports on the monitoring of patients post-operatively on surgical wards (such as the studies by De Meester et al. [2013]; Gao et al. [2007]; Ludikhuize et al. [2012]; Paterson et al. [2006]), but

to the best of our knowledge, none has focused on the *variability* of physiological variables. A strategy to incorporate these clinically significant variations in the vital signs within a 24-hour period in the construction of models of normality has been proposed. Compared to current early warning scoring systems, these data-driven strategies would provide earlier identification of instability, which would allow earlier escalation of care, which in turn could lead to improved patient outcomes.

The analysis presented here goes some way towards addressing the lack of clinical evidence for the efficacy of machine learning methods in patient monitoring. It also suggests that the "real-time" implementation of these novelty detection methods may be possible in UK hospitals. Because of the gradual introduction in hospitals of electronic patient records, and of electronic devices to record physiological data, the use of scoring systems which are based on computerised algorithms and data-fusion methods, rather than simple integer scores, is becoming feasible, which will prompt the adoption of more sophisticated approaches.

As a final point, the CALMS-2 dataset used in the analyses described in the last three chapters consists of manual measurements of vital signs acquired periodically (approximately, every four hours) by ward staff. These infrequent patient observations can lead to unnoticed clinical deterioration, including "abnormal" 24-hour variability in the vital signs. A solution to the sparseness of observational data will be the use of patient monitoring systems based on continuous data acquired from patient-worn sensors. The challenges for such an approach are to provide early warning of patient deterioration in a robust manner with low numbers of false alerts.

Chapter 7

Functional characterisation of vital-sign trajectories with Gaussian processes

The task of discovering novel medical knowledge from complex, large-scale and high-dimensional patient data, collected during care episodes, is central to innovation in medicine. The recognition of complex trajectories in multivariate time-series data requires effective models and representations for the analysis and matching of functional data.

In this chapter, we propose a method based on Gaussian processes for exploratory data analysis using the observational physiological time-series data. While our primary motivation comes from clinical data, this approach may be applicable to other time-series domains. Our method focuses on a representation of unevenly-sampled *trajectories* that allows for revealing physiological recovery patterns and identifying unseen, and possibly "abnormal", patterns in the database of vital signs acquired from post-operative patients. We first describe methods that have been proposed in the literature for the same purpose. We then provide a brief summary of Gaussian processes, and describe our proposed approach for performing clustering of patients' trajectories.

7.1 Background

The task of knowledge discovery from time-series data, as shown in the previous chapter, is important for "tracking" the health status of post-operative patients. An enormous amount of work has been devoted to the task of modelling time-series data.

The autoregressive model is a basic means of analysing time-series data, which specifies that the output variable depends linearly on its previous values. Other examples include state-space models, which are based on the notion that there is an unobserved state of the system, or latent state, that evolves through time and which may only be observed indirectly. For example, the health status of a patient can only be observed through "noisy" observations of the patient's physiology and mental status.

The most basic state-space model with a continuous-valued latent state is the linear dynamical system (LDS), which is the discrete-time analogue of a linear differential equation. The hidden Markov model (HMM) (Baker [1975]) is the discrete-state space analogue of an LDS. Quinn et al. [2009] applied an extension of an LDS model to the problem of monitoring the condition of premature infants receiving intensive care. A factorial-switching LDS model (equivalent to a switching Kalman filter) was described and tested with continuous time-series data collected from bedside monitors. This model was developed into a hierarchical factorial switching LDS (Stanculescu et al. [2014]) by adding a set of higher-level variables to model correlations in the physiological factors in order to detect sepsis in ICU patients. Lehman et al. [2013] used a switching vector autoregressive framework to systematically learn and identify continuously-acquired arterial blood pressure data dynamics. These can possibly be recurrent within the same patient and shared across an entire cohort of ICU patients.

Recent work by Willsky et al. [2009a,b] uses Bayesian nonparametric models for capturing the generation of continuous-valued time-series. This method uses a HMM for segmenting time-series data, where the latter are characterised by autoregressive models. Beta processes, which provide prior distributions in the unit interval, are then used to share observation models across several series. Thus, this *BP-AR-HMM* model is used to capture variability between series

by sampling subsets of low-level features that are specific to individual series. Lehman et al. [2012] used this model to discover shared dynamics in ICU patients' continuously-acquired blood pressure time-series data. A different Bayesian non-parametric method for exploratory data analysis and feature construction in continuous time-series has been proposed by Saria et al. [2010]. This method builds on the framework of latent Dirichlet allocation and its extension to hierarchical Dirichlet processes, which allows the characterisation of each series as switching between latent "modes", where each mode is characterised as a distribution over features that specify the series dynamics. The model was applied to heart-rate data collected from premature infants admitted to a neonatal ICU. A different probabilistic model, the *continuous shape template model*, has also been applied for discovering time-series' segments that can repeat within and across different series of continuous heart rate data (Saria et al. [2011]).

Although conceptually sound, it is unclear how such approaches cope with irregularly-sampled data and missing data. As opposed to equally-spaced time-series, on which the methods described above have been applied, irregularly-sampled time-series data are characterised by variable intervals between successive measurements; i.e., the spacing of observation times is not constant. Different time-series typically contain different numbers of observations and the times at which observations were recorded may not be aligned. Furthermore, periods of missing data are common in clinical scenarios. The properties of these data mean that most common machine learning algorithms and models for supervised and unsupervised learning cannot be directly applied.

One solution to these problems is offered by Gaussian processes. Gaussian processes are a Bayesian modelling technique that has been widely used for various machine learning tasks, such as dimensionality reduction, nonlinear classification, and regression (Lawrence [2005]; Rasmussen and Williams [2006]). It is a nonparametric method, informally suggesting that the number of parameters in the model can grow with the number of observed data. Compared to other related techniques, Gaussian process models have the advantage that prior knowledge of the functional behaviour (e.g., periodicity or smoothness) may be easily expressed. The Bayesian nature of its formulation also means that inference is performed within a probabilistic framework, allowing us to reason in the presence

of noise, incompleteness, and artefacts, all of which are characteristic of the data recorded in hospital settings.

Gaussian processes have been used for modelling physiological time-series data. Clifton et al. [2013b] and Wong et al. [2012] used Gaussian process regression to cope with artifactual and missing vital-sign data, and incorporated the Gaussian process posterior in their novelty detection schemes. Stegle et al. [2008] proposed a robust regression model for noisy heart rate data based on Gaussian processes and a preliminary clustering procedure that learns the structure of outliers and noise bursts. In the work described in [Wong, 2011, Chapter 6], trend analysis was performed using dependent Gaussian processes, in which the correlation between two or more physiological variables is used to obtain improved regression results. Clifton et al. [2013a] extended extreme value theory such that a function-wise approach to novelty detection was taken, as opposed to pointwise approaches that are most commonly described in the literature. The method was illustrated using Gaussian process regression, which offers a probabilistic framework in which distributions over a function space are defined. Gaussian process regression has also been used for the ranking of gene expressions (Kalaitzis and Lawrence [2011]).

In this work, we propose a representation of vital-sign trajectories using Gaussian process regression, which may be used for the recognition of "normal" and "abnormal" patterns of physiological trends. Figure 7.1 illustrates the components of our proposed approach. We model the evolution of the unevenly-sampled physiological trajectories using Gaussian process regression, and we introduce a kernel similarity measurement for the comparison of the latent functions based on the likelihoods of the data points in each trajectory. This *patient-to-patient* similarity measurement can then be used for recognising known trajectories and identifying unknown trajectories as would be required for identifying "abnormal" vital-sign time-series.

7.2 Dataset

For the analysis described in this chapter, we included patients who stayed for a minimum of 24 hours on the post-operative ward, and for a period no longer than

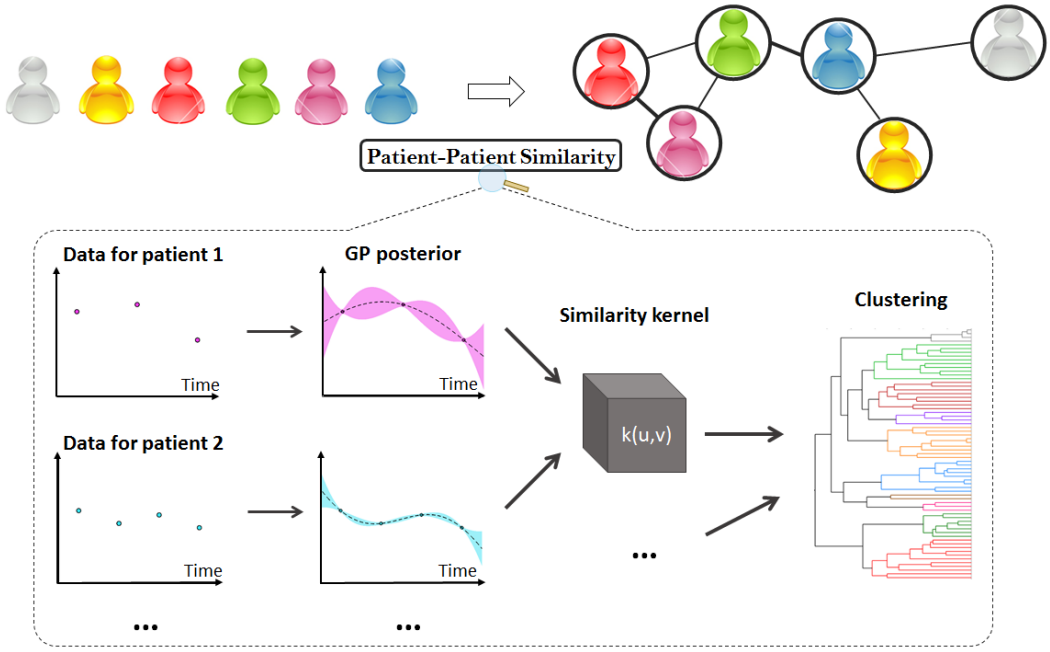


Figure 7.1: Overview of our approach for the functional charaterisation of vital-sign trajectories with Gaussian processes.

20 days (which corresponds approximately to the 95$^{\text{th}}$ quantile for the length of stay on the ward of the entire cohort of patients). The rationale for this was to exclude both very short or very long stayers from our analysis and focus on a more "homogeneous" cohort of patients with regard to length of stay. For patients in the "abnormal" group, we considered only data acquired up to the first major adverse event (death on the ward, emergency ICU admission, or cardiac arrest). In addition, to test the predictive ability of this approach, for each patient in the abnormal group we excluded the last 12 hours before the major adverse event. This resulted in a total of 364 patients (326 from the normal group, plus 38 from the abnormal group) that were included in this analysis. For all patients, the first day of their vital-sign trajectories corresponds to the day on which surgery took place.

Although the proposed approach can be directly applied to multivariate time-series data, given the small size of the dataset we demonstrate the proposed approach using univariate observational data from our cohort of post-operative

patients. We can, however, take into account the contribution of all five vital signs (and not focus only on a single vital sign). For this, we consider the output of the model constructed in section 6.2, which provides a parsimonious representation of the overall physiological trajectories. In short, a multivariate model of normality based on pre-discharge vital-sign data from normal patients $\mathbf{U}$ is constructed using kernel density estimates; then, for each patient, the likelihood $p(\mathbf{u}|\mathbf{U}, \sigma)$ of each observation set $\mathbf{u}$ with respect to this model is computed, and the correspondent novelty score $z(\mathbf{u})$ is finally obtained as before: $z(\mathbf{u}) = -\log p(\mathbf{u}|\mathbf{U}, \theta)$[1]. Thus, for each patient, we obtain a "univariate", unevenly-sampled time-series of novelty score values; i.e., a collection n pairs of $(t, z(\mathbf{u}))$, where t corresponds to the time of the observation set $\mathbf{u}$, and n is the number of observation sets for that patient. The details of how the model is constructed have previously been described (see section 6.2).

7.3 Gaussian processes

We provide a brief summary of Gaussian processes in this section. It therefore makes a rather compressed introduction to the topic. A more thorough introduction is available in Rasmussen and Williams [2006].

When performing a regression task we assume there exists some optimal prediction function $f \in \mathcal{X} \to \mathcal{Y}$, possibly with a noise distribution. In linear regression, we assume that the outputs $\mathbf{y}$ are a linear function of the inputs $\mathbf{X}$, with some parameters $\boldsymbol{\theta}$, usually fewer than the number of training examples $N : |\boldsymbol{\theta}| \ll N$. However, for many real-world datasets a simple parametric form, such as a linear form, is an unrealistic assumption. Therefore, we would like to have models that can learn general functions f. Since the functions may not be summarised by a small (fixed) number of parameters $\boldsymbol{\theta}$, maximum likelihood estimation of the parameters may cause overfitting. In fact, in a Gaussian process, the effective number of parameters is often infinite. Therefore, in order to perform inference we need to place a prior probability distribution on functions. We make predictions using our posterior on an underlying predictive function f given a set of

[1] To account for highly extreme data points of $z(\mathbf{u})$, values were hard limited at the 1^{st} and 99^{th} quantiles of all z($\mathbf{u}$) values computed during training.

training examples in the form of input-output pairs: $\mathcal{D} = \{(\mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R})\}_{i=1}^N$.

Gaussian processes provide a distribution over real-valued functions which is widely used for non-linear regression and classification tasks (Rasmussen and Williams [2006]). By definition, a function $f : \mathcal{X} \to \mathbb{R}$ is distributed according to a Gaussian process if and only if $p(f(\mathbf{x}_1), ..., f(\mathbf{x}_N))$, the density of that function's values at any N points $\mathbf{x}_i \in \mathcal{X}$, is multivariate Gaussian. This allows Gaussian processes to be parameterised tractably by a mean function $m(\mathbf{x})$ and a covariance kernel function $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ specifying the correlations within any finite point set, such that

$$\mathbf{y} = f(\mathbf{x}) \sim \mathcal{GP}\Big(m(\mathbf{x}), \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)\Big), \tag{7.1}$$

with possibly some Gaussian observation noise. Note that the covariance matrix $\mathbf{K}$, or Gram matrix, whose entries $\mathbf{K}_{ij}$ are often thought of as the "similarity" between inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, encodes our prior knowledge concerning the functional behaviour we wish to model. Without loss of generality, the prior mean function is typically set to zero: $m(\mathbf{x}) = 0$. The most commonly used covariance function is the squared-exponential[1],

$$k_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_0^2 \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\ell^2}\right), \tag{7.2}$$

where $\boldsymbol{\theta} = \{\sigma_0, \ell\}$ are hyperparameters modelling the y-scaling and x-scaling (or time-scale if the data are time-series), respectively, and where $\| \cdot \|$ denotes the Euclidean norm. The squared-exponential covariance function is said to be *stationary* because it only depends on the difference between points $\mathbf{x}_i - \mathbf{x}_j$, rather than on their absolute value. In general, covariance functions have to fulfill Mercer's theorem, meaning that $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ has to be symmetric and positive semidefinite, and therefore $k_{SE}(\cdot, \cdot)$ is a valid kernel. Many mathematical operations, such as summation or taking a product, preserve positive definiteness and can therefore be used for combining basic kernels to make more complex kernels. A survey of covariance functions can be found in [Rasmussen and Williams, 2006, Chapter 4].

Given a training set $\mathcal{D}$, using the standard conditioning rules for a Gaussian

[1] It is also known as the exponentiated-quadratic, or the Gaussian kernel function.

distribution, we can obtain the predictive distribution on a new observation y_* at test input $\mathbf{x}_*$:

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\top & \mathbf{K}_{**} \end{bmatrix} \right) \tag{7.3}$$

implying

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) \sim \mathcal{N}(\mu_*, \sigma_*^2), \text{ with} \tag{7.4}$$

$$\mu_* = \mathbf{K}_*^\top \mathbf{K}^{-1} \mathbf{y} \in \mathbb{R}, \tag{7.5}$$

$$\sigma_*^2 = \mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}^{-1} \mathbf{K}_* \in \mathbb{R}^+. \tag{7.6}$$

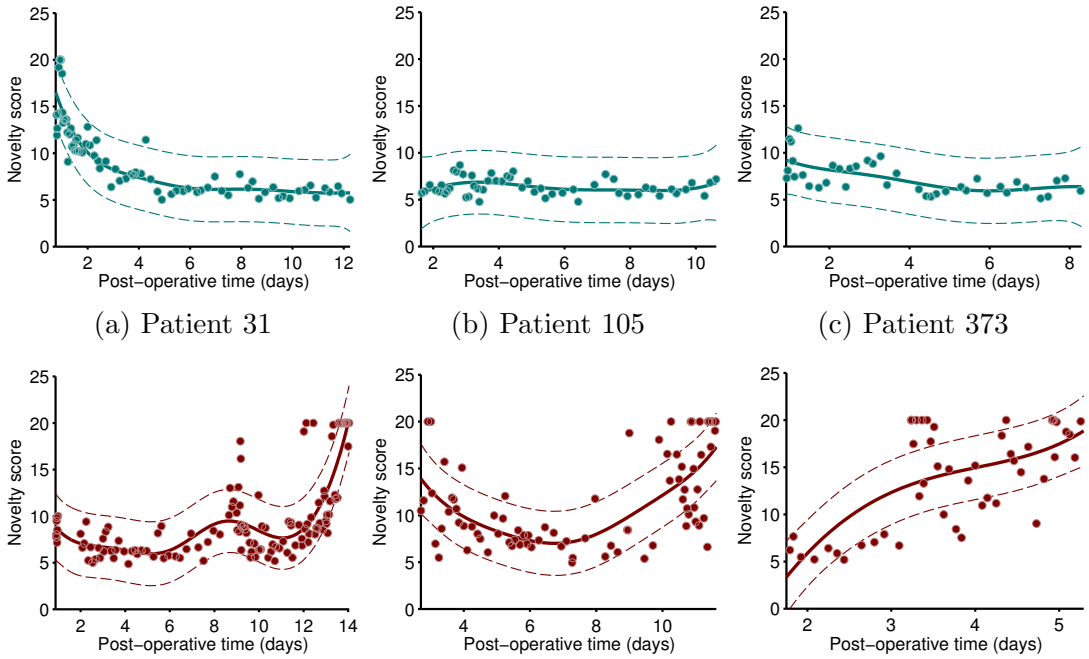
Here, $\mathbf{K}_* = k(\mathbf{X}, \mathbf{x}_*) \in \mathbb{R}^{N+1}$ is the cross-covariance between the test input $\mathbf{x}_*$ and the training inputs $\mathbf{X}$; $\mathbf{K}_{**} = k(\mathbf{x}_*, \mathbf{x}_*) \in \mathbb{R}^+$ is the prior variance of $\mathbf{x}_*$.

The values of the hyperparameters $\boldsymbol{\theta}$ may be optimised by, for example, minimising the negative log marginal likelihood (NLML) which is defined as

$$\begin{aligned} \text{NLML} &= -\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \\ &= \frac{1}{2}\log|\mathbf{K}| + \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y} + \frac{N}{2}\log(2\pi) \end{aligned} \tag{7.7}$$

This is sometimes called the type-II maximum likelihood (if we remove the negative logarithm). Interpreting the NLML as a cost function reveals that the first term penalises model complexity and the second term penalises low data likelihood (i.e., low data fitness). Bias-variance trade-off is therefore performed by minimising the NLML, which is commonly achieved using gradient descent. In a full Bayesian treatment, we should integrate out the hyperparameters. Unfortunately, this cannot be performed analytically in general; e.g., for the input scale. Sampling methods, or other approximations, are usually used to estimate these integrals (Rasmussen and Williams [2006]).

In our experiments, we used a single squared-exponential covariance function and a zero-mean function to capture the overall physiological recovery of post-operative patients. During training, each time-series was centred by removing



(a) Patient 31 (b) Patient 105 (c) Patient 373

(d) Patient 68 (e) Patient 175 (f) Patient 349

Figure 7.2: Examples of the Gaussian process posteriors obtained for the novelty scores of six post-operative patients. Circles correspond to the raw data. Thick lines correspond to the posterior means, and dashed lines mark the 95% confidence area for the computed posterior mean. Coloured areas denote the uncertainty level of the mean (darker areas, uncertainty is lower). Top row shows three patients from the normal group, and the bottom row shows three patients from the "abnormal" group, who had an event 12 hours after the corresponding last observation is shown.

the mean of the time-series data to achieve a zero-mean function. The hyper-parameters $\{\sigma_0^2, \ell\}$ were selected using a grid-search optimiser for minimising the negative log marginal likelihood: $\sigma_o^2 \in [3, 4, 5, ..., 15]$ (in units of $z(\mathbf{x})$) and $\ell \in [2.0, 2.5, 3.0, ..., 5.0]$ (in units of days). We then evaluated the resulting function over a uniform grid of test points $\mathbf{x}_*$ sampled every hour within the range $\mathbf{x}_*^n \in [t_1, t_f]$, where t_1 and t_f correspond to the time of the first and last observations for patient n. Figure 7.2 shows a few examples of the regression results obtained with our dataset using this procedure.

We observe that small (daily) variations of the novelty scores are smoothed by use of this approach. Nevertheless, the model is able to capture the overall trajectory of recovery of the patients. For example, patient 31 exhibits a high

initial physiological derangement following major surgery, and a clear return to normality (decrease in the physiological novelty score), as a result of recovery on the ward. Patient 105, on the other hand, appears to be within the normal range of novelty score values throughout their stay on the ward. Conversely, "abnormal" patients exhibit an increase in novelty score in the last couple of days before the major adverse event, as expected. Although some "abnormal" patients may initially exhibit a recovery "trend" that is similar to that of "normal" patients (such as patients 68 and 175), others may exhibit physiological trajectories that are substantially different, such as that of patient 349, with increasing novelty scores throughout the post-operative period.

7.4 Time-series clustering

In this section we describe our proposed approach for performing clustering of the Gaussian process posteriors over the uniform grid of test points (sampled every hour).

7.4.1 Similarity measurement

To quantify the similarity of time-series we make use of kernels. Kernel-based classifiers, like any other classification scheme, should be robust against invariances and distortions. Dynamic time warping (DTW), a method based on dynamic programming (Sakoe and Chiba [1978]), has been previously combined with kernel methods (Bahlmann et al. [2002]; Shimodaira et al. [2002]).

Let $\mathcal{X}^{\mathbb{N}}$ be the set of discrete-time time-series taking values in an arbitrary space $\mathcal{X}$. One can try to align two time-series $\mathbf{u} = (u_1, ..., u_n)$ and $\mathbf{v} = (v_1, ..., v_m)$ of lengths n and m, respectively, in various ways by distorting them. An alignment $\boldsymbol{\pi}$ of length $|\boldsymbol{\pi}| = p$ between two sequences $\mathbf{u}$ and $\mathbf{v}$ (with $p \leq n+m-1$ since the two series have $n+m$ points and they are matched at least at one point in time) is a pair of increasing integer vectors (π_1, π_2) such that $1 \leq \pi_1(1) \leq ... \leq \pi_1(p) = n$ and $1 \leq \pi_2(1) \leq ... \leq \pi_2(p) = m$, with unitary increments and no simultaneous repetitions (we use the notation of Cuturi [2011]). We write $\mathcal{A}(\mathbf{u}, \mathbf{v})$ for the set of all possible alignments between $\mathbf{u}$ and $\mathbf{v}$, which can be conveniently represented

by paths in an $n \times m$ matrix. Following the well-known DTW metric, the *cost* of the alignment can be defined by means of a distance ϕ that measures the discrepancy between any two points u_i and v_j, such that

$$D_{\mathbf{u},\mathbf{v}}(\boldsymbol{\pi}) = \sum_{i=1}^{|\boldsymbol{\pi}|} \phi(u_{\pi_1(i)}, v_{\pi_2(i)}) \tag{7.8}$$

Dynamic programming algorithms provide an efficient way to compute the optimal path $\boldsymbol{\pi}^*$ which gives the minimum cost among all possible alignments,

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi} \in \mathcal{A}(\mathbf{u},\mathbf{v})}{\arg\min} \frac{1}{|\boldsymbol{\pi}|} D_{\mathbf{u},\mathbf{v}}(\boldsymbol{\pi}) \tag{7.9}$$

Different kernel distances (or scores) ϕ have been proposed in the literature to compute the similarity between time-series based on DTW, such as the negative squared Euclidean distance $\phi(u, v) = -||u - v||^2$ (Bahlmann et al. [2002]),

$$k_{DTW_1}(\mathbf{u}, \mathbf{v}) = \exp\left(- \underset{\boldsymbol{\pi} \in \mathcal{A}(\mathbf{u},\mathbf{v})}{\arg\min} \frac{1}{|\boldsymbol{\pi}|} \sum_{i=1}^{|\boldsymbol{\pi}|} ||u_{\pi_1(i)} - v_{\pi_2(i)}||^2 \right), \tag{7.10}$$

or a Gaussian kernel (Shimodaira et al. [2002]),

$$k_{DTW_2}(\mathbf{u}, \mathbf{v}) = \underset{\boldsymbol{\pi} \in \mathcal{A}(\mathbf{u},\mathbf{v})}{\arg\max} \frac{1}{|\boldsymbol{\pi}|} \sum_{i=1}^{|\boldsymbol{\pi}|} \exp\left(-\frac{1}{\sigma^2} ||u_{\pi_1(i)} - v_{\pi_2(i)}||^2 \right). \tag{7.11}$$

The global alignment (GA) kernel, proposed by Cuturi et al. [2007], assumes that the alignment that gives the minimum cost may be sensitive to peculiarities of the time-series and intends to take advantage of all possible alignments weighted exponentially. Hence, it is defined as the sum of exponentiated costs of the individual alignments, such that

$$k_{GA}(\mathbf{u},\mathbf{v}) = \sum_{\boldsymbol{\pi}\in\mathcal{A}(\mathbf{u},\mathbf{v})} \exp\left(-D_{\mathbf{u},\mathbf{v}}(\boldsymbol{\pi})\right) \tag{7.12}$$

$$= \sum_{\boldsymbol{\pi}\in\mathcal{A}(\mathbf{u},\mathbf{v})} \exp\left(-\sum_{i=1}^{|\boldsymbol{\pi}|}\phi(u_{\pi_1(i)},v_{\pi_2(i)})\right) \tag{7.13}$$

$$= \sum_{\boldsymbol{\pi}\in\mathcal{A}(\mathbf{u},\mathbf{v})} \prod_{i=1}^{|\boldsymbol{\pi}|} k(u_{\pi_1(i)},v_{\pi_2(i)}) \tag{7.14}$$

where $k = \exp -\phi$. It has been argued that k_{GA} runs over the whole spectrum of the costs and leads to a smoother measure than the minimum of the costs, i.e., the DTW distance (Cuturi et al. [2007]).

In our implementation, we use the kernel suggested by Cuturi [2011],

$$k(u,v) = \exp(-\phi_\sigma(u,v)), \tag{7.15}$$

$$\phi_\sigma(u,v) = \frac{1}{2\sigma^2}d(u,v) + \log\left(2 - e^{-\frac{1}{2\sigma^2}d(u,v)}\right) \tag{7.16}$$

where the bandwidth σ of the kernel can be set as a multiple of a simple estimate of the median (Euclidean) distance of different points observed in different time-series of the training set, scaled by the square root of the median length of time-series in the training set[1], as suggested in Cuturi [2011]; $d(u,v)$ corresponds to the distance between any two points of the time-series $\mathbf{u}$ and $\mathbf{v}$. Cuturi et al. [2007] used $d(u,v) = ||u-v||^2$. In our case, as previously described, the time-series or trajectories obtained with the Gaussian process framework are characterised by a mean function and a measure of the uncertainty in the trajectory estimation, which handles the incompleteness, noise and artefacts underlying the observational data considered. That is, because we used a Gaussian likelihood function, each point u_i in a given trajectory $\mathbf{u}$, is defined by $u_i \sim \mathcal{N}(m_{u_i}, \Sigma_{u_i})$. In order to take this into account, we use the 2-Wasserstein distance between two Gaussian distributions (Takatsu [2011]), which is given by

[1] That is, $\hat{\sigma} = \text{median}(||u-v||)\sqrt{L}$, where L corresponds to the median length of the time-series in $\mathcal{X}$.

$$d(u,v) = d\big(\mathcal{N}(m_u,\Sigma_u),\mathcal{N}(m_v,\Sigma_v)\big) = ||m_u - m_v||^2 + ||\Sigma_u^{1/2} - \Sigma_v^{1/2}||_F^2 \quad (7.17)$$

where $||\cdot||_F$ is the *Frobenius* (also called *Hilbert-Schmidt*) norm.

7.4.2 Clustering method

Using the measure of discrepancy (or similarity) described above, classification or clustering of the trajectories may be performed. There are a large number of clustering methods proposed in the literature. In this chapter, we use an *agglomerative hierarchical clustering* method. Other partitioning techniques, such as k-means or model-based clustering, share the property that objects in a dataset are partitioned into a specific number of clusters at a single step. In contrast, hierarchical clustering methods produce a cluster tree; i.e., a series of nested clusters through a series of partitions.

Hierarchical clustering can be either *agglomerative*, with fewer clusters at the higher level (by fusing clusters generated at the lower level), or *divisive*, which separate the n objects into more and finer groups in sequential steps. Agglomerative hierarchical clustering, in particular, starts with n clusters, each of which contains a single object in the dataset. In the second step, the two clusters that have the closest between-cluster distance are fused and are then treated as a single cluster in the next step. As the procedure continues, it results in a single cluster containing all the n objects. Agglomerative methods vary in the ways of defining the distance between two clusters when more than one object is present in either of them. For example, the single linkage method considers the shortest pairwise distance between objects in two different clusters as the distance between the two clusters. In contrast, with the complete linkage method, the distance between two clusters is defined as the distance between the most distant pair of objects. Here, we use average linkage clustering, in which the average of the pairwise distances between all pairs of objects coming from each of two clusters is taken as the distance between the two clusters.

The number of clusters was estimated using the *gap* method described by Tibshirani et al. [2001]. A short review on the estimation of the optimal number

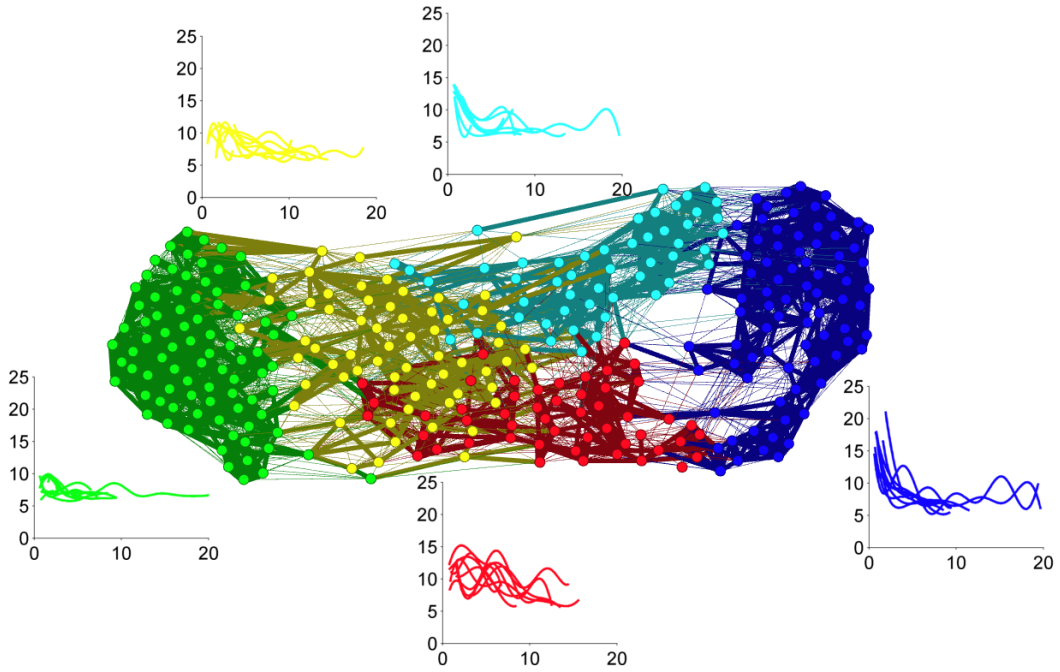


Figure 7.3: Representation of the clusters obtained during training using our patient-to-patient similarity approach: each node of the graph corresponds to a patient, and the edges connecting any two nodes represent the similarity between them. In each sub-plot, 10 random mean trajectories from each cluster are represented.

of clusters is provided in Appendix D.

7.4.3 Characterisation of physiological patterns of recovery from surgery

We firstly applied this method to the trajectories of normal patients in order to find different patterns of physiological recovery from major surgery. For this, the Gaussian process posteriors (over the uniform grid of test points sampled every hour) of the physiological trajectories from all "normal" patients (as computed in section 7.3) were used. Hierarchical clustering was used to group similar trajectories based on the modified global alignment kernel distance introduced above.

The number of clusters obtained, determined using the *gap* method, was 5

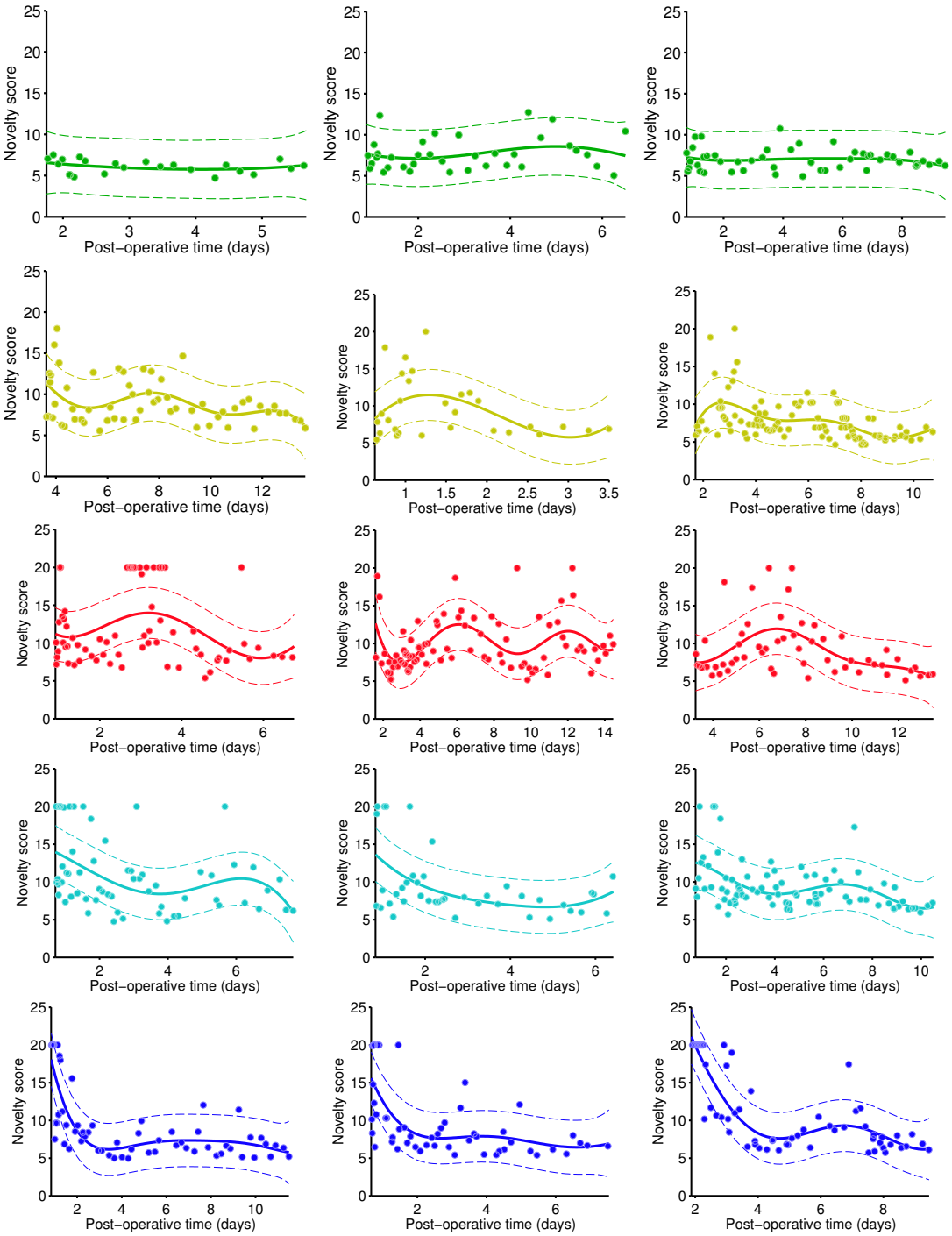


Figure 7.4: Examples of the Gaussian process posteriors for three patients belonging to each cluster (colours correspond to those used in Figure 7.3). Circles correspond to the raw data. Thick lines correspond to the posterior means, and dashed lines indicate the 95% confidence area for the posterior mean obtained. Coloured areas denote the uncertainty level of the mean (in darker areas, uncertainty is lower).

functional clusters. Figure 7.3[1] illustrates the clusters of patients obtained. Figure 7.4 shows examples of trajectories associated with each cluster of patient trajectories, as an overall representation of the results obtained in this experiment. From the 326 patients included in the normal group, 87 (27%) were part of the cluster coloured with green, 79 (24%) were part of the cluster coloured with dark-blue, and the remaining of the patients were part of the other clusters (51 or 16% in the red cluster, 58 or 17% in the yellow cluster, and 51 or 16% in the light-blue cluster).

Different patients may exhibit different physiological trajectories during recovery. Although all patients in the normal group did recover from surgery and were discharged home without any major adverse event in the course of their stay in the hospital, these results suggest that the physiological trajectories (based on the novelty scores) for these patients may be different from one another, as expected. While some patients from the "normal" group exhibit a recovery trend with a pronounced decrease in the novelty score $z(\mathbf{x})$ in the first couple of days after surgery and a constant $z(\mathbf{x})$ for the remainder of their stay (Figure 7.4, bottom row), other patients present a relatively "stable" trajectory, with only small variations of $z(\mathbf{x})$ throughout their stay (e.g., Figure 7.4, top row). For other "normal" patients, a certain variation of $z(\mathbf{x})$ is manifested in their physiological trajectories. Using the similarity metric and clustering procedure proposed, the set of entities that are alike appear (visually) to be assigned to the same cluster, and entities from different clusters are also clearly less alike.

7.5 Abnormal vital-sign trajectory detection

In order to study the ability of the proposed time-series data clustering approach in identifying "abnormal" patient trajectories, we evaluate it using a five-fold cross-validation procedure. As described in the previous chapter, data from four-fifths of the "normal" patients were used for training, and data from the remaining patients were used for testing. For each fold, we define a trajectory from an

[1]The graph was obtained using the freely available software called *Gephi*; for that, the similarity matrix (as computed by the proposed approach) was provided to the software, and the *ForceAtlas 2* algorithm was used to reorganise the layout of the graph, which takes into account the degree of similarity between the nodes and their neighbourhood.

"abnormal" patient to be a true positive (TP) if it is correctly classified as "novel", and a false negative (FN) if it is incorrectly classified as belonging to one of the clusters; we define a trajectory from a "normal" patient to be a true negative (TN) if it is correctly classified as belonging to one of the clusters, and a false positive (FP) if it is classified as "novel".

7.5.1 Methodology

In each fold, during training, the Gaussian process posteriors of the physiological trajectories from "normal" patients (as computed in section 7.3) are clustered using the procedure described above: (1) the modified global alignment kernel distance between any two trajectories is calculated; (2) agglomerative hierarchical clustering is used to group similar trajectories based on the determined similarity metric; and (3) the optimal number of clusters is determined using the *gap* method (where the possible number of clusters is varied from 1 to 30), and, based on this, the maximum similarity score (or minimum distance) for assigning a trajectory to a cluster is determined[1]. In the test phase, the distance between each test Gaussian posterior trajectory and the training Gaussian process posterior trajectories is calculated, and the average of the distances between the test trajectory and the training trajectories belonging to the same cluster is determined. If the minimum average distance is higher than the cutoff distance determined during training, the test trajectory is classified as a "novel" trajectory (that is, it is substantially different from the trajectories computed during training). If the minimum average distance is smaller than the cutoff distance, then the test trajectory is assigned to the cluster for which the average distance to its members is the minimum.

We compared the performance of our approach to that of other two models (described below). For these two models, the input trajectories were set to be the mean of the Gaussian posteriors (over the uniform grid of test points sampled every hour) obtained.

Hidden Markov model (HMM). For this method, described in Rabiner [1989]

[1] That is, the minimum distance at which a cut on the hierarchical clustering tree leaves the determined number of clusters.

Table 7.1: Performance metrics for detecting abnormal patient vital-sign trajectories: optimal number of clusters selected with the proposed method, number of true positives, false negatives, true negatives and false positives. The results are extracted from five test folds and the numbers are presented with mean and standard deviation.

		"Normal" Trajectories ($N = 65$)		"Abnormal" Trajectories ($N = 38$)	
Method	N. Clusters	TN	FP	TP	FN
HMM	-	53 (2)	12 (2)	30 (2)	8 (2)
KDE	-	56 (5)	9 (5)	26 (1)	15 (1)
Our approach	5 (1)	51 (2)	14 (3)	33 (1)	5 (1)

and Song et al. [2013], we used an implementation of Murphy [1998]. We trained HMMs using normal training trajectories, and computed the likelihood of each test trajectory as the normalised negative log-likelihood. Trajectories were classified as "abnormal" if the normalised negative log-likelihood was above 0.5. We varied the number of hidden states and the number of Gaussian mixtures per state to take the values in the intervals $[2, 4, 6, 8]$ and $[1, 2, 3]$, respectively. We report the best performance obtained in each fold.

Kernel density estimates (KDE). This method was used to estimate the pdf of the data using the training trajectories. The width σ of the isotropic Gaussian kernels was found by maximising the leave-one-out likelihood of the training data. A score for each data point in a test trajectory was obtained by computing its negative log-likelihood. Test trajectories were classified as abnormal if the average of the scores was above a certain set of thresholds. The set of thresholds were defined to be the 60^{th}, 70^{th}, 80^{th}, 90^{th} and 95^{th} quantile of the negative log-likelihoods computed from the training trajectories. We report results with the threshold that gave the best performance in each fold.

7.5.2 Results

Table 7.1 shows experimental results on the test splits. We observe that the proposed approach achieved a good performance, which is directly comparable to that obtained with the baseline models. The results show that, using the proposed approach, the majority of normal patients in the test set were assigned to one

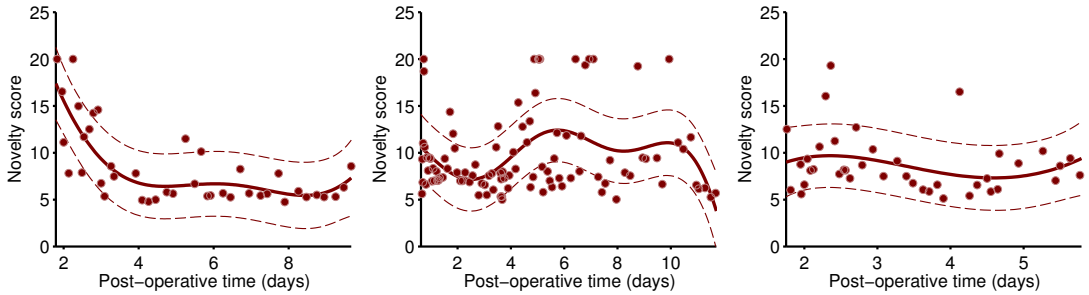
of the clusters determined during training. This suggests that the trajectories included in the training set were representative of the different normal trajectories of patients recovering from major surgery. Some of the trajectories from normal patients included in the test set, however, were classified as "novel" trajectories.

On the other hand, most abnormal trajectories were classified as "novel", which suggests that they are significantly different from normal recovery trajectories, and they would have been identified to be "abnormal" 12 hours before the major adverse event. Examples of misclassified abnormal trajectories are shown in Figure 7.5. It may be seen that these trajectories look very similar to those from normal patients. We note that, for describing the trajectory of patients from the abnormal group, we removed the last 12 hours of data before the major adverse event. Therefore, some events may have been missed due to the fact that the trajectories did not include these data. Also, as observed in the previous chapters, the physiological derangement of some patients before a major adverse event may not be very pronounced, hence, they may exhibit a physiological pattern similar to that of normal patients until very close to the adverse event.

7.5.3 Discussion

We have described a method by which unevenly-sampled time-series data may be analysed to better understand the overall recovery trajectories of post-operative patients. We have also shown that this approach is able to recognise "normal" or previously observed physiological patterns and identify abnormal or "novel" physiological trajectories. As expected, patients may exhibit different recovery patterns during the course of their stay on the ward. There are many possible explanations for the different recovery patterns observed: the type of surgery that the patient underwent, the age of the patient, how fit the patient is at the time of operation, and other possible underlying conditions. No direct and clear associations between the first two factors (surgery type and age) and the clusters of data were found, which may be due to the small number of patients included in this study and the variety of procedures that patients underwent.

A few points should be made here regarding the analysis conducted and the



(a) Patient 17 (b) Patient 62 (c) Patient 126

Figure 7.5: Examples of the Gaussian process posteriors obtained for the novelty scores of three post-operative patients from the "abnormal" group, incorrectly classified as "normal". Circles correspond to the raw data. Thick lines correspond to the posterior means, and dashed lines mark the 95% confidence area for the posterior mean obtained. Coloured areas denote the uncertainty level of the mean (darker areas, uncertainty is lower). Data are shown up to 12 hours before the major adverse event.

results obtained. In the first place, we observe that small (daily) variations of the novelty scores were smoothed out in this analysis (e.g., the trajectory from patient 62 in Figure 7.5). The main goal of this approach was to capture the overall trajectory of recovery of post-operative patients, which motivated the selection of the covariance function (and hyperparameters priors) that was used to model the data with Gaussian processes. In order to also capture short-term variations, one could use a more complex covariance function derived by combining simple covariance functions. For example, the addition of two squared-exponential covariance functions, one to model the short-term variations in the novelty score, and one to model the long-term trends, could be used to provide a better fit to the data. Nevertheless, some additional work would be required to select the set of priors for each hyperparameter. A fully Bayesian approach would be advantageous in this case to better encode the level of uncertainty in the hyperparameters; i.e., rather than using an expensive grid-search optimisation procedure over all possible values for each hyperparameter, prior distributions could be set for each of the hyperparameters, which would be integrated out to obtain the Gaussian process posterior mean and variance.

It is also important to mention that, although we focused on the analysis using trajectories of novelty scores (univariate time-series data), the same approach

could be used for multivariate time-series; for example, by considering all the vital signs, rather than the novelty score that combines them into a single score. As described above, the global alignment kernel distance is able to cope with multivariate time-series data. Nevertheless, the visualisation of the results for evaluating the performance of the method would be more challenging than that for the univariate case. Moreover, due to the increase of degrees of freedom in the multivariate case, a larger sample of data would be needed to derive a more representative set of trajectories for clustering.

We note that the comparison of our method with other approaches proposed in the literature (such as those proposed in Saria et al. [2010, 2011]; Willsky et al. [2009a]) may be difficult due to the characteristics of our observational dataset. In this work, we considered the comparison with two other novelty detection approaches based on HMMs and KDE. Our approach provided a better performance as it includes a direct quantification of the uncertainty in the trajectory estimation (provided by the Gaussian process model), handling incompleteness, noise, and artefact in a robust manner.

Finally, we also observe that a direct comparison of the results obtained in this chapter with those obtained in previous chapters is not possible. Firstly, for the analysis conducted in this chapter, a limited dataset which includes fewer patients and fewer major adverse events was used. Secondly, for describing the trajectory of patients from the abnormal group, we removed the last 12 hours of data before the major adverse event, and therefore, some events may have been missed due to the fact that the trajectories did not include these data. And thirdly, deteriorations that are manifested by very subtle variations of the novelty score may not be captured using our modelling strategy. In the approach presented in this chapter, progressive and long-term deteriorations are captured, as the overall physiological trajectories are substantially different from those of "normal" recoveries from major surgery.

Chapter 8

Conclusion

When a patient deteriorates and becomes acutely unwell whilst in hospital, time is of the essence and a fast and efficient clinical response is required to optimise clinical outcomes (Brady et al. [2011]). Early detection, timeliness of response and competency of the clinical response to deteriorating patients are a triad of determinants of clinical outcome in patients with acute illness. Numerous recent national and international reports on acute clinical care have advocated the use of early warning scores and track-and-trigger systems based on manual observations by clinical staff to identify and respond to patients who present or develop acute illness as efficiently as possible; however, the approach is not standardised.

According to the Royal College of Physicians (London), "this variation in methodology and approach can result in a lack of familiarity with local systems when staff move between clinical areas/hospitals - the various EWS systems are not necessarily equivalent or interchangeable [...] And this can lead to a lack of consistency in the approach to detection and response to acute illness" (RCP [2012]). This lack of "standardisation" also impedes education and training in the monitoring of acutely-ill patients for all grades of healthcare professionals across the NHS. The National Confidential Enquiry into Patient Outcome and Death report (NCEPOD [2012]) also recommended that: "A clear physiological monitoring plan should be created for each patient commensurate with their clinical condition. This should detail what is to be monitored, the desirable [variables] and the frequency of observations. This should be regardless of the type of ward to which the patients are transferred".

While the development of a unique (national) system would deliver some of these objectives, we agree that the evidence provided in this thesis indicates that the use of the same underlying scoring system (based solely on physiological variables) on all hospital wards is not optimal. Although the physiological variables to be measured routinely and included in the scoring system are generally agreed, what weight or score should be given to the magnitude of disturbance to each of these variables? Should a clinical alert be based on an extreme variation of one variable, or an aggregate score for all variables, or a combination of both? At which score should the clinical response be escalated, i.e., how sensitive shall the trigger be? More importantly, should these weights, scores, and triggers be the same for patient populations in different hospital settings? While a unique scoring system is used on general ward patient populations, it may not be suitable for all in-hospital patients. That is, it can be argued that scoring systems need to be developed and validated in specific patient groups and not for large heterogeneous groups of hospital patients (as also argued by Cuthbertson et al. [2007]). For example, major planned surgery results in a substantial physiological insult from which patients are expected to recover to their normal physiological state. These physiological changes expected following major surgery, which are not taken into account by recommended scoring systems, may well not be seen in patients on other wards. Hence, knowledge of these different patterns among hospitalised patients and their incorporation in the monitoring systems can improve early-warning scoring systems used for the identification of physiological deterioration.

Despite recognition that scores need to be developed by taking into account the population in which they will be used, the recommended national scoring system (by the RCP [2012]) has not been developed using specific post-surgical populations.

8.1 Thesis overview

In this thesis, we sought to develop and apply principled approaches based on machine learning to the vital-sign data recorded from patients who are recovering

from elective major surgery, and incorporate knowledge and information about the conditions of these patients and their physiological changes. In chapter 2, we described the CALMS-2 trial, a "before-and-after" prospective trial designed to investigate the clinical use of continuous vital-sign monitoring during the post-operative period. We also introduced the Portsmouth dataset, a large database of vital signs collected from patients admitted to a MAU in another hospital.

A comprehensive review of the performance of current early warning scoring systems in the two cohorts of patients was performed in chapter 3. As one might expect, the systems proposed by Badriyah et al. [2014]; Prytherch et al. [2010]; RCP [2012] had the highest performances for identifying deterioration based on the derived outcome of death within 24 hours after an observation set using the Portsmouth dataset. The same vital signs database (and derived outcome) was used to develop these scoring systems. When applied to the post-operative patients included in the CALMS-2 trial, the performance of these systems for predicting the composite outcome of death, emergency ICU admission and cardiac arrest within 24 hours, was also seen to be superior to the other EWS systems used in the analysis.

In chapter 4, we demonstrated that the method of recording physiological variables on the ward may play a fundamental role in both the design and performance of current early-warning scoring systems. The original CEWS system proposed by Tarassenko et al. [2011] used a large database of continuously-recorded vital-sign data (as opposed to observational data) to obtain the cutoff values for each vital sign, which differ from those of other EWS systems, in particular for respiratory rate. When the cutoff values of this vital sign were adjusted using observational data, the performance of the modified CEWS system improved significantly when tested on both databases. This result is important, as the design of future data-driven early-warning scoring systems should take into account the different sources of data used to build those systems, as well as the different methods of recording the physiological variables for which the systems will be used. Moreover, in the era of electronic medical records and automated methods for acquiring vital-sign data, then the various EWS systems currently in clinical use may have to adjusted in order to cope with the differences between manually-charted and automatically-collected data.

Alternative strategies based on machine learning techniques were explored in chapter 5. In particular, multivariate density estimation (such as kernel density estimates) and boundary-based methods (such as SVMs) were used in the hope that they could capture "normal" physiological patterns of hospitalised patients, so that "abnormal" observation sets could be identified and the occurrence of a major adverse event detected earlier. On the validation set (Portsmouth dataset), the performance of these methods was remarkably similar to those obtained with the scoring systems evaluated in the preceding chapters. On the test set (CALMS-2) the proposed approaches did not perform as well as one would expect. This result can, however, be explained by fundamental differences between the two patient populations included in each database. An important observation from the analysis conducted in the preceding chapters was that there was a clear effect associated with the inclusion of supplemental use of oxygen in the scoring systems. Typically, an extra score of 2 or 3 is added to the overall score if the patient is on oxygen support. This was an expected result, as patients who are more at risk of deterioration will be on oxygen support, while patients who are about to be discharged from the hospital will not be receiving oxygen support. We note, however, that the inclusion of this "marker" on a scoring system may not be straightforward. In the post-operative population studied, patients are generally on oxygen masks for the first couple of days after surgery, which corresponds to the period of time in which a post-operative complication is more likely to occur. In a MAU, oxygen support is provided to patients who are in a state of hypoxia, or are at risk of becoming hypoxic. In the proposed multivariate machine learning methods, we considered the use of oxygen support as an additional (binary) variable for the models. The different protocols used in different hospital settings, may produce fundamental differences in the correlations and relationships between the model variables. As multivariate models have the ability to capture these correlations, they may become biased towards the patient population on which they are trained, and, hence, produce poorer performances when tested on a different population, for which those relationships do not hold. These findings strongly support the view that models need to be developed and validated for *specific patient groups*.

In chapter 6, a slightly different approach was taken in order to account for the

(specific) physiological changes following major surgery. We computed the overall physiological trajectory using a model of normality based on pre-discharge data. This provided an interesting summary of the typical physiological trajectory for post-operative patients. As expected, patients had a relatively high initial physiological derangement following major surgery, and a clear return to normality was then seen as a result of recovery on the ward. A substantial different physiological trajectory was obtained for patients who had a major adverse event, particularly in the last 72 hours before the event. We further introduced the concept of *24-hour variability* for vital signs and proposed a strategy based on *time-based normalisation* for incorporating these clinically significant variations in the vital signs in the construction of models of normality. Results suggest that incorporating the evolution of the variability of physiological variables rather than just their absolute values is beneficial to model performance. A higher AUROC value was obtained, and physiological deterioration would have been identified earlier with the system that included vital-sign variability as input variable (as reflected by the higher time-to-event values).

An approach based on Gaussian processes was then proposed in chapter 7 to explore the physiological trajectories computed for post-operative patients in the preceding chapter. This method focused on a representation of unevenly-sampled trajectories that allowed for revealing physiological recovery patterns and identifying unseen, and possibly "abnormal", patterns in the CALMS-2 database. Using a similarity metric, which is based on the concepts of dynamic time warping and global alignment kernel, and a hierarchical clustering method, different groups of physiological behaviours of recovery from surgery were revealed by the proposed approach. The majority of patients were found to belong to one of two functional clusters: one group of patients who exhibited a recovery trend with a pronounced decrease in the novelty score in the first couple of days after surgery and a constant score for the remainder of their stay on the ward; and a group of patients who presented a relatively "stable" trajectory, with only small variations of the novelty score throughout their stay post-operatively. This approach was also able to recognise "normal" or previously observed physiological patterns and identify physiological trajectories from "abnormal" patients.

The proposed approach may provide a new tool for studying and better un-

derstanding the recovery phase of patients post-operatively, which is known to be heterogeneous. As electronic medical records continue to collect data from other interventions (e.g., elective surgery), there will be a growing need for such tools to refine the characterisation of what constitutes a "normal" and an "abnormal" recovery from a major intervention, and quantify the effects of variability in treatment protocols across individuals in these groups.

A couple of other considerations concerning the overall analyses performed in this thesis should be highlighted. The first point concerns the evaluation of the methods studied. It is difficult to compare between one-class classification methods fairly, as classical evaluation measures may not be relevant for the proposed task, influenced as it is by the nature of the dataset and/or the nature of the domain studied. Indeed, a wide range of measures have been proposed in the literature among which global accuracy, sensitivity, specificity, precision and recall, ROC curves, AUROC, partial AUROC or other averaging methods, all of which aim at summarising the performance of a standard classifier. However, there is no consensus for the performance computation of one-class algorithms nor for their comparison because the standard measures are more or less biased by the uneven ratio between the two classes. In spite of this bias, the results of our analyses were presented in terms of different performance metrics that allowed for an analysis of different trade-offs, and tried to take into account the unbalanced nature of our clinical datasets. We further note that our estimates of the values of the standardized partial AUROC should be interpreted in the context of the range of interest. In particular, using a wider range of interest than that which is of relevance clinically, could lead to overoptimistic estimates of performance in many practical scenarios.

Secondly, we have only used physiological data from patients once their care was transferred from higher dependency areas (such as the ICU) to a standard post-operative ward. The physiology from higher-dependency areas was not included as an underlying principle of high-dependency care is to observe the physiology continuously and to intervene to return patients to the "normal" range. The frequent interventions make the physiological data difficult to interpret.

Finally, it is important to note that all patients included in this study may well have avoided a major adverse event by undergoing review for more minor

deteriorations. This is a problem common to the analysis of performance of all early-warning scores (Tarassenko et al. [2011]). Reducing the alert rate for any population must run the risk of missing deterioration. However, to be useful, early-warning scoring systems must identify a sufficiently small subgroup to justify urgent review. The use of physiological values relevant to both the type of patient and the phase of recovery should therefore be the appropriate next step in developing suitable tools for early recognition of deterioration in post-operative patients, as demonstrated in this thesis. Such improvements may greatly improve the identification of deteriorating surgical patients by reducing alert fatigue, avoiding unnecessary intervention, and recognising physiological deterioration in the relatively stable later phase of recovery. With the gradual introduction in hospitals of electronic patient records and of electronic devices to record physiological data, the introduction of scoring systems that are based on computerised algorithms and data-fusion methods is becoming feasible.

8.2 Future work

The work presented in this thesis provides many opportunities and important directions for future work. We split these into two related research themes, that are described next.

8.2.1 Fusing multi-modal data

The benefits of continuous monitoring systems have been described in previous studies (Clifton et al. [2014]; Taenzer and Blike [2012]; Taenzer et al. [2010, 2011]; Watkinson et al. [2006]). In addition to nurse observations, in the CALMS-2 trial, patients also had their vital signs monitored continuously by both a bedside monitor (in the first couple of days following surgery) and wearable sensors for the remainder of their stay on the ward (see Figure 2.1). Such monitors provide not only continuous (high-temporal resolution) vital-sign data streams, but also the waveforms from which vital signs are obtained, and additional information may be extracted. Some work has already been performed on the extraction of additional

physiological variables from ECG and PPG waveforms acquired from the ECG sensors and pulse oximeters, respectively. Specifically, recent work has focused on the extraction of respiratory rate from these waveforms (Fleming and Tarassenko [2007]; Garde et al. [2014]; Karlen et al. [2013]; Orphanidou et al. [2013]). However, trial implementations of these methods have demonstrated that resulting RR estimates are not robust and cannot be used in clinical practice without further improvement of the estimation algorithms. This improvement can be achieved by including metrics that evaluate the quality of the waveforms (Orphanidou et al. [2013]; Shah [2012]), which in combination with the RR estimates translate these waveforms into additional information that can be used as inputs to scoring systems.

This will generate multiple sources (channels) of the same variable of interest; for example, RR continuously estimated from waveform data, and RR manually collected by nursing staff. This thesis provides additional evidence of the limitations of the current standard of patient monitoring with intermittent manual vital-sign data collection, because of the low temporal resolution that may lead to important indicators of deterioration being missed.

There are a growing number of scenarios in machine learning in which multimodal information is used, and where information from multiple sensors (or sources) needs to be fused to recover the variable of interest. The three main issues in this area can be summarised as follows: data channels are frequently missing, they are usually not sampled at the same rate, and there might be different biases and noise levels associated with the different channels. Kapoor et al. [2005] proposed a mixture of Gaussian processes for addressing these issues and combining multiple modalities within a Bayesian framework. The framework uses a mixture of Gaussian processes, where the classification using each channel is learned via Expectation-Propagation, a technique for approximate Bayesian inference. The resulting posterior over each classification function is a product of Gaussians and can be updated very quickly. A different approach was more recently proposed by Xiao et al. [2013]. In the latter, the authors present a probabilistic model based on Gaussian processes for regression when there are multiple (unreliable) *observers* providing continuous responses. For this problem, each *observer* may correspond to a source of data, and there may exist two or three

observers of variables such as RR estimates. This approach can be used not only to estimate the underlying variable of interest but also make predictions of the underlying variable.

Such approaches could be used to extend the Gaussian process framework so that the intermittent observations performed by clinical staff are optimally combined with the continuous data acquired from the wearable sensors.

8.2.2 Fusing multivariate data

The benefit of continuous monitoring is that a patient's condition can be monitored more frequently, and in principle, deterioration can be detected as soon as it manifests in the physiological variables measured, even if this occurs in the period between nurse's observations. Notwithstanding this, there are key challenges for deploying such continuous data monitoring systems in hospitals: the presence of artefactual data and periods of missing data, and the limited subset of physiological variables captured by continuous data monitors.

An approach for combining multivariate, temporal data has been explored by Wong [2011], who applied multi-task Gaussian processes to vital-sign data collected from the Emergency Department. Multi-task Gaussian processes (or multi-output Gaussian processes or dependent Gaussian processes) extend Gaussian processes to handle multiple correlated outputs simultaneously (as described in Bonilla et al. [2008]). The main advantage of this method is that the model exploits not only the temporal correlation of data from one output but also those of the other outputs. This can be used to improve Gaussian process regression/prediction of an output given the others, thus performing data fusion. This framework may be further extended to yield further improvements in the identification of patient deterioration by considering the through-time behaviour of the vital signs. That is, standard Gaussian process models typically use a stationary covariance function, in which the covariance between any two points is a function of the Euclidean distance between the input points. However, stationary Gaussian processes fail to adapt to variable smoothness in the function of interest (Plagemann et al. [2008]; Rasmussen and Williams [2006]). This is of particular importance in patient monitoring data, in which results presented in this thesis

suggest that the functional behaviour of the vital signs may vary more quickly in some periods of time than in others. The inferred hyperparameters of such Gaussian process models may then be associated with important aspects of physiological behaviour, such as the variability of the vital signs, and may contribute to the identification of patient deterioration.

On a more general note, we observe that very large amounts of clinical data will soon be available for real-time analysis. Novel technological developments in the field of big data will allow the storage and analysis of such data in real-time. In the next generation of patient monitoring systems, static alarm triggers will be combined with smart alarms, which have the ability to identify and track patterns associated with clinical deterioration. Ideally, systems will integrate electronic patient records and bedside monitors to allow the calculation of early-warning scores based on physiological, laboratory, demographic, and comorbidity data, without the need for the user to provide additional input or set alarms and triggers.

8.3 Conclusion

To conclude, starting from the currently-used early-warning scores, different methods based on principled approaches for identifying patient deterioration have been developed and evaluated in two different patient populations. We have demonstrated that the method of recording vital signs on the ward play a fundamental role in both the design and performance of early-warning scoring systems. Machine learning modelling techniques using the manual observations recorded by clinical staff have been implemented; new indicators of physiological deterioration based on the variability of individual vital signs have been introduced; and approaches for incorporating such dynamic information into multivariate models have been explored. The most important model improvement has come from the use of population-specific recovery trends and derived "biomarkers". Finally, we also proposed a dynamic modelling approach that may be used for describing different physiological trajectories during recovery from surgery. While further studies on additional post-surgical populations are needed to validate this ap-

proach, the analysis and results obtained with the methods described in this thesis supports the use of data-fusion models based on machine learning techniques for the effective and robust identification of patient deterioration.

References

N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 504–509. ACM, 2006. 128

M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538, 2006. 111

N. Alam, E. L. Hobbelink, A. J. van Tienhoven, P. M. van de Ven, E. P. Jansma, and P. W. B. Nanayakkara. The impact of the use of the early warning score (EWS) on patient outcomes: A systematic review. *Resuscitation*, 85:587–594, Jan 2014. 2

K. Allen. Recognising and managing adult patients who are critically sick. *Medicine*, 22:244–247, 2004. 64, 77

C. A. Alvarez, C. A. Clark, S. Zhang, E. A. Halm, J. J. Shannon, C. E. Girod, L. Cooper, and R. Amarasingham. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Med Inform Decis Mak*, 13:28, 2013. 137

T. Andrews and H. Waterman. Packaging: a grounded theory of how to report physiological deterioration effectively. *Journal of Advanced Nursing*, 52(5): 473–481, 2005. 64

T. Badriyah, J. S. Briggs, P. Meredith, S. W. Jarvis, P. E. Schmidt, P. I. Featherstone, D. R. Prytherch, and G. B. Smith. Decision-tree early warning score

(DTEWS) validates the design of the National early warning score (NEWS). *Resuscitation*, 85(3):418–423, Mar 2014. 64, 67, 69, 75, 86, 87, 95, 96, 101, 198

C. Bahlmann, B. Haasdonk, and H. Burkhardt. Online handwriting recognition with support vector machines - A kernel approach. In *Proceedings of the IEEE Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 49–54, 2002. 184, 185

Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris. A comparative study for outlier detection techniques in data mining. In *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6, 2006. 111

J. Baker. The DRAGON system - An overview. *IEEE Transactions on Acoustics Speech and Signal Processing*, 23(1):24–29, 1975. 176

C. Ball, M. Kirkby, and S. Williams. Effect of the critical care outreach team on patient survival to discharge from hospital and readmission to critical care: non-randomised population based study. *BMJ*, 327(7422):1014, Nov 2003. 65

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. 107

K. Beaumont, D. Luettel, and R. Thomson. Deterioration in hospital patients: early signs and appropriate actions. *Nurs Stand*, 23(1):43–48, 2008. 8

M. Bekkar, H. K. Djemaa, and T. A. Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3(10):27–38, 2013. 57

M. B. Bell, D. Konrad, F. Granath, A. Ekbom, and C.-R. Martling. Prevalence and sensitivity of MET-criteria in a Scandinavian University Hospital. *Resuscitation*, 70(1):66–73, 2006. 2, 65

R. Bellomo, M. Ackerman, M. Bailey, R. Beale, G. Clancy, V. Danesh, A. Hvarfner, E. Jimenez, D. Konrad, M. Lecardo, K. S. Pattee, J. Ritchie, K. Sherman, P. Tangkau, and Vital Signs to Identify, Target and Assess Level

of Care Study (VITAL Care Study) Investigators. A controlled trial of electronic automated advisory vital signs monitoring in general hospital wards. *Crit Care Med*, 40(8):2349–2361, Aug 2012. 22

Y. Bengio, H. Larochelle, and P. Vincent. Non-local manifold Parzen windows. *Advances in Neural Information Processing Systems*, 18:115–123, 2006. 246, 256

A. Benning, M. Dixon-Woods, U. Nwulu, M. Ghaleb, J. Dawson, N. Barber, B. D. Franklin, A. Girling, K. Hemming, M. Carmalt, G. Rudge, T. Naicker, A. Kotecha, M. C. Derrington, and R. Lilford. Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ*, 342(d199), 2011. 21

G. Berlot, A. Pangher, L. Petrucci, R. Bussani, and U. Lucangelo. Anticipating events of in-hospital cardiac arrest. *Eur J Emerg Med*, 11(1):24–28, Feb 2004. 10

A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406, 1946. 145, 146

W. Bianchi, A. F. Dugas, Y.-H. Hsieh, M. Saheed, P. Hill, C. Lindauer, A. Terzis, and R. E. Rothman. Revitalizing a vital sign: improving detection of tachypnea at primary triage. *Ann Emerg Med*, 61(1):37–43, Jan 2013. 88, 94

Y. U. Bing-Hua. Delayed admission to intensive care unit for critically surgical patients is associated with increased mortality. *Am J Surg*, 208(2):268–274, Jan 2014. 11, 12

C. M. Bishop. Novelty detection and neural network validation. In *Proceedings of the IEEE Conference on Vision, Image and Signal Processing*, volume 141, pages 217–222, 1994. 110, 116, 250

C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer, New York, 2006. 116, 120, 128

O. Boehm, D. R. Hardoon, and L. M. Manevitz. Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms. *International Journal of Machine Learning and Cybernetics*, 2(3): 125–134, 2011. 109

E. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian process prediction. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 153–160, 2008. 204

W. J. Brady, K. K. Gurka, B. Mehring, M. A. Peberdy, R. E. O'Connor, and American Heart Association's Get with the Guidelines (formerly, NRCPR) Investigators. In-hospital cardiac arrest: impact of monitoring and witnessed event on patient survival and neurologic status at hospital discharge. *Resuscitation*, 82(7):845–852, Jul 2011. 196

H. Brown, J. Terrence, P. Vasquez, D. W. Bates, and E. Zimlichman. Continuous monitoring in an inpatient medical-surgical unit: A controlled clinical trial. *The American Journal of Medicine*, 127(3):226–232, 2014. 26

S. E. S. Brown, S. J. Ratcliffe, J. M. Kahn, and S. D. Halpern. The epidemiology of intensive care unit readmissions in the United States. *Am J Respir Crit Care Med*, 185(9):955–964, May 2012. 10

S. E. S. Brown, S. J. Ratcliffe, and S. D. Halpern. An empirical derivation of the optimal time interval for defining ICU readmissions. *Med Care*, 51(8):706–714, Aug 2013. 10

C. P. Bryan and G. E. Smith. *Ancient Egyptian medicine: the papyrus ebers.* Ares, 1974. 6

M. Buist, S. Bernard, T. V. Nguyen, G. Moore, and J. Anderson. Association between clinically abnormal observations and subsequent in-hospital mortality: A prospective study. *Resuscitation*, 62(2):137–141, Aug 2004. 2, 8, 10, 18

M. D. Buist, G. E. Moore, S. A. Bernard, B. P. Waxman, J. N. Anderson, and T. V. Nguyen. Effects of a medical emergency team on reduction of incidence of

and mortality from unexpected cardiac arrests in hospital: preliminary study. *BMJ*, 324(7334):387–390, Feb 2002. 18, 31

G. Bunkenborg, K. Samuelson, I. Poulsen, S. Ladelund, and J. Åkeson. Lower incidence of unexpected in-hospital death after interprofessional implementation of a bedside track-and-trigger system. *Resuscitation*, 85(3):424–430, Mar 2014. 21

J. Caliński, T.and Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 263

P. Calzavacca, E. Licari, A. Tee, M. Egi, M. Haase, A. Haase-Fielitz, and R. Bellomo. A prospective study of factors influencing the outcome of patients after a Medical Emergency Team review. *Intensive Care Med*, 34(11):2112–2116, Nov 2008. 31

P. Calzavacca, E. Licari, A. Tee, M. Egi, A. Downey, J. Quach, A. Haase-Fielitz, M. Haase, and R. Bellomo. The impact of rapid response system on delayed emergency team activation patient characteristics and outcomes - A follow-up study. *Resuscitation*, 81(1):31–35, Jan 2010. 19

L. T. Q. Cardoso, C. M. C. Grion, T. Matsuo, E. H. T. Anami, I. A. M. Kauss, L. Seko, and A. M. Bonametti. Impact of delayed admission to intensive care units on mortality of critically ill patients: A cohort study. *Crit Care*, 15(1): R28, 2011. 11, 12

I. M. Chakravarty, J. D. Roy, and R. G. Laha. Handbook of methods of applied statistics. *Wiley, New York*, 1967. 97, 145

D. B. Chalfin, S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger, and D. E. L. A. Y-E. D. study group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med*, 35(6):1477–1483, Jun 2007. 11, 12

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15:1–58, 2009. 110, 111, 122

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 127

M. T. Chatterjee, J. C. Moon, R. Murphy, and D. McCrea. The "OBS" chart: an evidence based approach to re-design of the patient observation chart in a district general hospital setting. *Postgraduate Medical Journal*, 81(960):663–666, 2005. 64

C. A. Chrusch, K. P. Olafson, P. M. McMillan, D. E. Roberts, and P. R. Gray. High occupancy increases the risk of early death or readmission after transfer from intensive care. *Crit Care Med*, 37(10):2753–2758, Oct 2009. 11

G. Chrysochoou and S. R. Gunn. Demonstrating the benefit of medical emergency teams (MET) proves more difficult than anticipated. *Crit Care*, 10(2):306, 2006. 17

M. M. Churpek, T. C. Yuen, C. Winslow, A. A. Robicsek, D. O. Meltzer, R. D. Gibbons, and D. P. Edelson. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med*, 190(6):649–655, Sep 2014. 137, 172

D. A. Clifton, L. Clifton, L. Tarassenko, P. J. Watkinson, V. S. Barber, and J. Salmon. Patient-specific biomedical condition monitoring in post-operative cancer patients. In *Proceedings of the Sixth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies - CM/MFPT*, 2009. 34, 35

D. A. Clifton, S. Hugueny, and L. Tarassenko. Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems*, 65(3):371–389, 2011a. 114

D. A. Clifton, D. Wong, S. Fleming, S. J. Wilson, R. Way, R. Pullinger, and L. Tarassenko. Novelty detection for identifying deterioration in emergency department patients. In Hujun Yin, Wenjia Wang, and Victor Rayward-Smith,

editors, *Intelligent Data Engineering and Automated Learning (IDEAL)*, volume 6936 of *Lecture Notes in Computer Science*, pages 220–227. Springer Berlin Heidelberg, 2011b. 26

D. A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko. An extreme function theory for novelty detection. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):28–37, 2013a. 178

L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko. Identification of patient deterioration in vital-sign data using one-class support vector machines. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 125–131, 2011c. 114

L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2013b. 178

L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE J Biomed Health Inform*, 18(3):722–730, May 2014. 47, 119, 120, 122, 202

G. Cohen, H. Sax, and A. Geissbuhler. Novelty detection using one-class Parzen density estimator. an application to surveillance of nosocomial infections. *Stud Health Technol Inform*, 136:21–26, 2008. 109

T. Cooksley, E. Kitlowski, and P. Haji-Michael. Effectiveness of modified early warning score in predicting outcomes in oncology patients. *QJM*, 105(11): 1083–1088, Nov 2012. 68

N. Cooper. Patient at risk! *Clinical Medicine*, 1(4):309–311, 2001. 64

M. Cretikos, J. Chen, K. Hillman, R. Bellomo, S. Finfer, A. Flabouris, and MERIT study investigators. The objective medical emergency team activation criteria: A case-control study. *Resuscitation*, 73(1):62–72, Apr 2007. 10

B. H. Cuthbertson. The impact of critical care outreach: is there one? *Crit Care*, 11(6):179, 2007. 19

B. H. Cuthbertson, M. Boroujerdi, L. McKie, L. Aucott, and G. Prescott. Can physiological variables and early warning scoring systems allow early recognition of the deteriorating surgical patient? *Crit Care Med*, 35(2):402–409, Feb 2007. 10, 29, 68, 69, 86, 137, 197

M. Cuturi. Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 929–936, 2011. 184, 186

M. Cuturi, J.-P. Vert, Ø. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 413–416, 2007. 185, 186

K. De Meester, T. Das, K. Hellemans, W. Verbrugghe, P. G. Jorens, G. A. Verpooten, and P. Van Bogaert. Impact of a standardized nurse observation protocol including MEWS after intensive care unit discharge. *Resuscitation*, 84 (2):184–188, Feb 2013. 21, 68, 173

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 120

Department of Health. Comprehensive critical care: a review of adult critical care services. Report (London), 2000. 1

A. Depeursinge, J. Iavindrasana, A. Hidki, G. Cohen, A. Geissbuhler, A. Platon, P.-A. Poletti, and H. Müller. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *J Digit Imaging*, 23(1):18–30, Feb 2010. 109

M. A. DeVita, R. S. Braithwaite, R. Mahidhara, S. Stuart, M. Foraida, R. L. Simmons, and Medical Emergency Response Improvement Team (MERIT). Use of medical emergency team responses to reduce hospital cardiopulmonary arrests. *Qual Saf Health Care*, 13(4):251–254, Aug 2004. 65

M. A. Devita, R. Bellomo, K. Hillman, J. Kellum, A. Rotondi, D. Teres, A. Auerbach, W.-J. Chen, K. Duncan, G. Kenward, M. Bell, M. Buist, J. Chen, J. Bion, A. Kirby, G. Lighthall, J. Ovreveit, R. S. Braithwaite, J. Gosbee, E. Milbrandt,

M. Peberdy, L. Savitz, L. Young, M. Harvey, and S. Galhotra. Findings of the first consensus conference on medical emergency teams. *Crit Care Med*, 34(9): 2463–2478, Sep 2006. 15

M. A. DeVita, G. B. Smith, S. K. Adam, I. Adams-Pizarro, M. Buist, R. Bellomo, R. Bonello, E. Cerchiari, B. Farlow, D. Goldsmith, H. Haskell, K. Hillman, M. Howell, M. Hravnak, E. A. Hunt, A. Hvarfner, J. Kellett, G. K. Lighthall, A. Lippert, F. K. Lippert, R. Mahroof, J. S. Myers, M. Rosen, S. Reynolds, A. Rotondi, F. Rubulotta, and B. Winters. "Identifying the hospitalised patient in crisis" - A consensus conference on the afferent limb of rapid response systems. *Resuscitation*, 81(4):375–382, Apr 2010. 27

F. Dexter, A. Macario, D. H. Penning, and P. Chung. Development of an appropriate list of surgical procedures of a specified maximum anesthetic complexity to be performed at a new ambulatory surgery facility. *Anesth Analg*, 95(1): 78–82, Jul 2002. 45

R. W. Duckitt, R. Buxton-Thomas, J. Walker, E. Cheek, V. Bewick, R. Venn, and L. G. Forni. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. an observational, population-based single-centre study. *Br J Anaesth*, 98(6):769–774, Jun 2007. 2, 64

M. Elliott and A. Coventry. Critical care: the eight vital signs of patient monitoring. *Br J Nurs*, 21(10):621–625, 2012. 41, 42, 43

G. Erdoğan. Outlier detection toolbox in Matlab. 2011. Software available at https://goker.wordpress.com/2011/12/30/outlier-detection-toolbox-in-matlab/. 128

D. Erdogmus, R. Jenssen, Y. N. Rao, and J. C. Principe. Multivariate density estimation with optimal marginal parzen density estimation and gaussianization. In *Proceedings of the 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pages 73–82, 2004. 250

G. J. Escobar, J. C. LaGuardia, B. J. Turk, A. Ragins, P. Kipnis, and D. Draper. Early detection of impending physiologic deterioration among patients who

are not in intensive care: development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine*, 7(5): 388–395, 2012. 137

T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 31:1–38, 2004. 56, 58

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8):861–874, 2006. 56, 57, 82

T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62. ACM, 1999. 60

B. J. Fisher. *'Online' system for patients has a double advantage.* Electronics Weekly, 1968. 7

S. G. Fleming and L. Tarassenko. A comparison of signal processing techniques for the extraction of breathing rate from the photoplethysmogram. *International Journal of Biological and Medical Sciences*, 2(4):232–236, 2007. 203

M. Folke, L. Cernerud, M. Ekström, and B. Hök. Critical review of non-invasive respiratory monitoring in medical care. *Med Biol Eng Comput*, 41(4):377–383, Jul 2003. 43

A. J. Forster, K. Kyeremanteng, J. Hooper, K. G. Shojania, and C. Van Walraven. The impact of adverse events in the intensive care unit on hospital mortality and length of stay. *BMC Health Serv Res*, 8:259, 2008. 13

A. H. Freiman and C. A. Steinberg. The analysis of simulaneously recorded cardiovascular data with the digital computer. *Annals of the New York Academy of Sciences*, 115:1091–1105, Jul 1964. 7

H. Gao, D. A. Harrison, G. J. Parry, K. Daly, C. P. Subbe, and K. Rowan. The impact of the introduction of critical care outreach services in England: a multicentre interrupted time-series analysis. *Crit Care*, 11(5):R113, 2007. 1, 2, 17, 18, 19, 63, 65, 71, 151, 173

A. Garde, W. Karlen, J. M. Ansermino, and G. A. Dumont. Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram. *PloS one*, 9(1):e86427, 2014. 203

J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling. The value of modified early warning score (MEWS) in surgical in-patients: a prospective observational study. *Ann R Coll Surg Engl*, 88(6):571–575, Oct 2006. 64, 68

L. A. Geddes. The acquisition of physiological data. *International Anesthesiology Clinics*, 3(3):379–406, 1965. 7

L. A. Geddes, M. Voelz, C. Combs, D. Reiner, and C. F. Babbs. Characterization of the oscillometric method for measuring indirect blood pressure. *Ann Biomed Eng*, 10(6):271–280, 1982. 42

D. D. Gibbs. The physician's pulse watch. *Medical history*, 15(02):187–190, 1971. 6

D. R. Goldhill. Preventing surgical deaths: critical care and intensive care outreach services in the postoperative period. *Br J Anaesth*, 95(1):88–94, Jul 2005. 1

D. R. Goldhill and A. F. McNarry. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. *Br J Anaesth*, 92(6):882–884, Jun 2004. 9, 10

D. R. Goldhill, S. A. White, and A. Sumner. Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia*, 54(6):529–534, Jun 1999. 9, 18

D. R. Goldhill, A. F. McNarry, G. Mandersloot, and A. McGinley. A physiologically-based early warning score for ward patients: the association between score and outcome. *Anaesthesia*, 60(6):547–553, Jun 2005. 64

H. Graham. *Surgeons all.* Philosophical Library, 1957. 6

K. C. Graham and M. Cvach. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *Am J Crit Care*, 19(1):28–34, Jan 2010. 27

E. L. Grogan, P. R. Norris, T. Speroff, A. Ozdas, D. J. France, P. A. Harris, J. M. Jenkins, R. Stiles, R. S. Dittus, and J. A. Morris Jr. Volatility: a new vital sign identified using a novel bedside monitoring strategy. *J Trauma*, 58 (1):7–12, Jan 2005. 160

K. O. Hajian-Tilaki and J. A. Hanley. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Acad Radiol*, 9(11):1278–1285, Nov 2002. 74, 240

M. A. Hamilton, M. Cecconi, and A. Rhodes. A systematic review and meta-analysis on the use of preemptive hemodynamic intervention to improve postoperative outcomes in moderate and high-risk surgical patients. *Anesth Analg*, 112(6):1392–1402, Jun 2011. 32

H. C. Hancock and L. Durham. Critical care outreach: The need for effective decision-making in clinical practice (Part 1). *Intensive and Critical Care Nursing*, 23(1):15–22, 2007. 64

C. Hands, E. Reid, P. Meredith, G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone. Patterns in the recording of vital signs and early warning scores: compliance with a clinical escalation protocol. *BMJ Qual Saf*, 22(9): 719–726, Sep 2013. 66

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. 57, 74, 240

A. Hann. *Multi-parameter Monitoring for Early Warning of Patient Deterioration*. PhD thesis, University of Oxford, 2008. 71, 109, 114, 127, 164, 257, 259

D. A. Harrison, H. Gao, C. A. Welch, and K. M. Rowan. The effects of critical care outreach services before and after critical care: a matched-cohort analysis. *J Crit Care*, 25(2):196–204, Jun 2010. 19

D. Harrison Jr and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978. 248

T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009. 61

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 108

M. D. Hellier and J. G. Williams. The burden of gastrointestinal disease: implications for the provision of care in the UK. *Gut*, 56(2):165–166, Feb 2007. 31

K. Hillman, J. Chen, M. Cretikos, R. Bellomo, D. Brown, G. Doig, S. Finfer, A. Flabouris, and MERIT study investigators. Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet*, 365(9477):2091–2097, 2005. 16, 20, 21

K. M. Hillman, P. J. Bristow, T. Chey, K. Daffurn, T. Jacques, S. L. Norman, G. F. Bishop, and G. Simmons. Antecedents to hospital deaths. *Intern Med J*, 31(6):343–348, Aug 2001. 1, 9, 10

K. M. 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 Med*, 28(11):1629–1634, Nov 2002. 1

V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. 111

T. J. Hodgetts, G. Kenward, I. G. Vlachonikolis, S. Payne, and N. Castle. The identification of risk factors for cardiac arrest and formulation of activation

criteria to alert a medical emergency team. *Resuscitation*, 54(2):125–131, Aug 2002. 10, 18

H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3): 863–874, 2007. 122

X. Hong, S. Chen, and C. J. Harris. A forward-constrained regression algorithm for sparse kernel density estimation. *IEEE Transactions on Neural Networks*, 19(1):193–198, 2008. 116, 251, 253, 256

M. Hravnak, M. A. DeVita, A. Clontz, L. Edwards, C. Valenta, and M. R. Pinsky. Cardiorespiratory instability before and after implementing an integrated monitoring system. *Crit Care Med*, 39(1):65–72, Jan 2011. 26

L. D. Hudson. Design of the intensive care unit from a monitoring point of view. *Respir Care*, 30(7):549–559, Jul 1985. 5

S. Hugueny. *Novelty detection with extreme value theory in vital-sign monitoring*. PhD thesis, University of Oxford, 2014. 71

Institute for Healthcare Improvement. IHI: Improving health and health care worldwide. [Accessed on 16 April 2014], 2010. Available at `http://www.ihi.org/resources/Pages/Changes/EstablishaRapidResponseTeam.aspx`. 1

I. Irigoien, B. Sierra, and C. Arenas. Towards application of one-class classification methods to medical data. *The Scientific World Journal, Hindawi Publishing Corporation*, 2014, 2014. 131

T. Jacques, G. A. Harrison, M.-L. McLaws, and G. Kilborn. Signs of critical conditions and emergency responses (SOCCER): a model for predicting adverse events in the inpatient setting. *Resuscitation*, 69(2):175–183, May 2006. 10

J. O. Jansen and B. H. Cuthbertson. Detecting critical illness outside the ICU: the role of track and trigger systems. *Curr Opin Crit Care*, 16(3):184–190, Jun 2010. 2

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013. 56, 57, 58

R. E. Jensen, H. Shubin, P. F. Meagher, and M. H. Weil. On-line computer monitoring of the seriously-ill patient. *Medical and Biological Engineering*, 4 (3):265–272, 1966. 7

D. A. Jones, M. A. DeVita, and R. Bellomo. Rapid-response teams. *New England Journal of Medicine*, 365(2):139–146, 2011. 16

A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12(1):180, May 2011. 178

A. Kapoor, H. Ahn, and R. W. Picard. Mixture of Gaussian processes for combining multiple modalities. In *Multiple Classifier Systems*, pages 86–96. Springer, 2005. 203

W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont. Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Transactions on Biomedical Engineering*, 60(7):1946–1953, July 2013. 203

L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 264

J. Kause, G. Smith, D. Prytherch, M. Parr, A. Flabouris, K. Hillman, Intensive Care Society (UK), and Australian and New Zealand Intensive Care Society Clinical Trials Group. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom - the ACADEMIA study. *Resuscitation*, 62(3):275–282, Sep 2004. 1, 9, 10

J. Kellett, B. Deane, and M. Gleeson. Derivation and validation of a score based on hypotension, oxygen saturation, low temperature, ECG changes and loss of independence (HOTEL) that predicts early mortality between 15 min and 24 h

after admission to an acute medical unit. *Resuscitation*, 78(1):52–58, Jul 2008. 19

J. Kellett, S. Woodworth, F. Wang, and W. Huang. Changes and their prognostic implications in the abbreviated VitalPAC™ early warning score (ViEWS) after admission to hospital of 18,853 acutely ill medical patients. *Resuscitation*, 84 (1):13–20, Jan 2013. 66

M. Kemmler, E. Rodner, E.-S. Wacker, and J. Denzler. One-class classification with Gaussian processes. *Pattern Recognition*, 46(12):3507 – 3518, 2013. 122

S. Khan and M. Madden. A survey of recent trends in one class classification. In *Artificial Intelligence and Cognitive Science*, volume 6206 of *Lecture Notes in Computer Science*, pages 188–197. Springer Berlin / Heidelberg, 2010. 111, 131

S. F. Khuri, W. G. Henderson, R. G. DePalma, C. Mosca, N. A. Healey, D. J. Kumbhani, and Participants in the V. A National Surgical Quality Improvement Program. Determinants of long-term survival after major surgery and the adverse effect of postoperative complications. *Ann Surg*, 242(3):326–341, Sep 2005. 32

W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: a severity of disease classification system. *Crit Care Med*, 13(10):818–829, Oct 1985. 39, 107

A. A. Kramer, T. L. Higgins, and J. E. Zimmerman. Intensive care unit readmissions in U.S. hospitals: patient characteristics, risk factors, and outcomes. *Crit Care Med*, 40(1):3–10, Jan 2012. 10

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 145

U. Kyriacos, J. Jelsma, and S. Jordan. Monitoring vital signs using early warning scoring systems: a review of the literature. *J Nurs Manag*, 19(3):311–330, Apr 2011. 19

U. Kyriacos, J. Jelsma, and S. Jordan. Record review to explore the adequacy of post-operative vital signs monitoring using a local modified early warning score (MEWS) chart to evaluate outcomes. *PLoS One*, 9(1):e87320, 2014. 68

T. S. Lam, P. S. K. Mak, W. S. Siu, M. Y. Lam, T. F. Cheung, and T. H. Rainer. Validation of a modified early warning score (MEWS) in emergency department observation ward patients. *Hong Kong J Emerg Med*, 13(1):24–30, 2006. 64

T. S. Larson and W. J. Brady. Electrocardiographic monitoring in the hospitalized patient: a diagnostic intervention of uncertain clinical impact. *Am J Emerg Med*, 26(9):1047–1055, Nov 2008. 25

N. Laurens and T. Dwyer. The impact of medical emergency teams on ICU admission rates, cardiopulmonary arrests and mortality in a regional hospital. *Resuscitation*, 82(6):707–712, Jun 2011. 21

N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, Dec 2005. 177

A. Lee, G. Bishop, K. M. Hillman, and K. Daffurn. The Medical Emergency Team. *Anaesth Intensive Care*, 23(2):183–186, Apr 1995. 65

H.-J. Lee and S. Cho. The novelty detection approach for different degrees of class imbalance. In *Neural Information Processing*, volume 4233 of *Lecture Notes in Computer Science*, pages 21–30. Springer Berlin / Heidelberg, 2006. ISBN 978-3-540-46481-5. 108

L. H. Lehman, S. Nemati, R. P. Adams, and R. G. Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5939–5942, 2012. 177

L. H. Lehman, S. Nemati, R. P. Adams, G. Moody, A. Malhotra, and R. G. Mark. Tracking progression of patient state of health in critical care using inferred shared dynamics in physiological time series. In *Proceedings of the*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 7072–7075, 2013. 113, 176

Q. Li and J. Racine. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266 – 292, 2003. 123, 125

M. von Lilienfeld-Toal, K. Midgley, S. Lieberbach, L. Barnard, A. Glasmacher, M. Gilleece, and G. Cook. Observation-based early warning scores to detect impending critical illness predict in-hospital and overall survival in patients undergoing allogeneic stem cell transplantation. *Biol Blood Marrow Transplant*, 13(5):568–576, May 2007. 64, 82, 241

N. T. Liu, J. B. Holcomb, C. E. Wade, M. I. Darrah, and J. Salinas. Utility of vital signs, heart rate variability and complexity, and machine learning for identifying the need for lifesaving interventions in trauma patients. *Shock*, 42 (2):108–114, Aug 2014. 160

S. Liu, R. L. Carpenter, and J. M. Neal. Epidural anesthesia and analgesia. Their role in postoperative outcome. *Anesthesiology*, 82(6):1474–1506, Jun 1995. 45

V. Liu, P. Kipnis, N. W. Rizk, and G. J. Escobar. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med*, 7(3):224–230, Mar 2012. 11

Y.-H. Liu, Y.-C. Liu, and Y.-Z. Chen. High-speed inline defect detection for TFT-LCD array process using a novel support vector data description. *Expert Systems with Applications*, 38(5):6222–6231, May 2011. 160

M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. 58

P. B. Lovett, J. M. Buchwald, K. Stürmann, and P. Bijur. The vexatious vital: neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage. *Ann Emerg Med*, 45(1): 68–76, Jan 2005. 94

J. Ludikhuize, S. M. Smorenburg, S. E. de Rooij, and E. de Jonge. Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the modified early warning score. *J Crit Care*, 27 (4):424.07–424.13, Aug 2012. 68, 173

L. A. Lynn and J. P. Curry. Patterns of unexpected in-hospital deaths: a root cause analysis. *Patient Saf Surg*, 5(1):3, 2011. 25

H. Ma, A. I. Bandos, H. E. Rockette, and D. Gur. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in medicine*, 32(20):3449–3458, 2013. 59

J. Mackenzie. *Diseases of the Heart*. Oxford University Press, 1925. 6

M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003a. 111

M. Markou and S. Singh. Novelty detection: a review - part 2: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003b. 111

M. Markou and S. Singh. A neural network-based novelty detector for image sequence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1664–1677, 2006. 110

S. Marsland. Novelty detection in learning systems. *Neural Computing Surveys*, 3:157–195, 2003. 111

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975. 57

L. Mayaud. *Prediction of mortality in septic patients with hypotension*. PhD thesis, University of Oxford, 2014. 107, 160

D. K. McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989. 59

J. McGaughey, F. Alderdice, R. Fowler, A. Kapila, A. Mayhew, and M. Moutray. Outreach and early warning systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev*, (3):CD005529, 2007. 16, 19

A. McGinley and R. M. Pearse. A national early warning score for acutely ill patients. *BMJ*, 345:e5310, 2012. 19

H. McGloin, S. K. Adam, and M. Singer. Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? *J R Coll Physicians Lond*, 33(3):255–259, 1999. 1

P. McQuillan, S. Pilkington, A. Allan, B. Taylor, A. Short, G. Morgan, M. Nielsen, D. Barrett, G. Smith, and C. H. Collins. Confidential inquiry into quality of care before admission to intensive care. *BMJ*, 316(7148):1853–1858, Jun 1998. 1, 15

C. E. Metz. Basic principles of roc analysis. *Semin Nucl Med*, 8(4):283–298, Oct 1978. 29, 56, 57

D. Miljkovic. Review of novelty detection methods. In *Proceedings of the IEEE 33rd International Convention (MIPRO)*, pages 593–598, 2010. 122

G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985. 263

A. Moon, J. F. Cosgrove, D. Lea, A. Fairs, and D. M. Cressey. An eight year audit before and after the introduction of modified early warning score (MEWS) charts, of patients admitted to a tertiary referral intensive care unit after CPR. *Resuscitation*, 82(2):150–154, Feb 2011. 20

R. Morgan, F. Williams, and M. Wright. An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8:100, 1997. 17, 18

M. M. Moya, M. W. Koch, and L. D. Hostetler. One-class classifier networks for target recognition applications. In *Proceedings of the World Congress on*

Neural Networks, International Neural Network Society, pages 797–801, 1993. 109

K. Murphy. Hidden Markov model (HMM) toolbox for Matlab. 1998. Software available at http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. 192

National Patient Safety Agency. NPSA - Safer care for the acutely ill patient: learning from serious incidents. Report (London), July 2007. 8, 9

NCEPOD. National confidential enquiry into patient outcome and death: An acute problem? Report (London), 2005. 9, 15, 17

NCEPOD. National confidential enquiry into patient outcome and death: Time to intervene? A review of patients who underwent cardiopulmonary resuscitation as a result of in-hospital cardiopulmonary arrest. Report (London), 2012. 14, 17, 196

NICE. National Institute for Clinical Excellence - Acutely ill patients in hospital: recognition of and response to acute illness in adults in hospital. Guidance/Clinical Guidelines CG50, 2007. 15, 17, 22

H. Nickisch and C. E. Rasmussen. Gaussian mixture modeling with Gaussian process latent variable models. In *Pattern Recognition*, pages 272–282. Springer, 2010. 254, 256

P. R. Norris, J. A. Morris, Jr, A. Ozdas, E. L. Grogan, and A. E. Williams. Heart rate variability predicts trauma patient outcome as early as 12 h: implications for military and civilian triage. *J Surg Res*, 129(1):122–128, Nov 2005. 160

P. R. Norris, P. K. Stein, and J. A. Morris, Jr. Reduced heart rate multiscale entropy predicts death in critical illness: a study of physiologic complexity in 285 trauma patients. *J Crit Care*, 23(3):399–405, Sep 2008. 160

U. Nwulu, D. Westwood, D. Edwards, F. Kelliher, and J. J. Coleman. Adoption of an electronic observation chart with an integrated early warning scoring system on pilot wards: a descriptive report. *Comput Inform Nurs*, 30(7):371–379, Jul 2012. 22

D. J. O'Callaghan, P. Jayia, E. Vaughan-Huxley, M. Gribbon, M. Templeton, J. R. Skipworth, and A. C. Gordon. An observational study to determine the effect of delayed admission to the intensive care unit on patient outcome. *Crit Care*, 16(5):R173, Oct 2012. 11, 12

M. Odell. Commentary: Hillman K., Chen J., et al. (2005). Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Nurs Crit Care*, 12(1):50–51, 2007. 64

M. Odell, A. Forster, K. Rudman, and F. Bass. The critical care outreach service and the early warning system on surgical wards. *Nurs Crit Care*, 7(3):132–135, 2002. 68

A. Oliver, C. Powell, D. Edwards, and B. Mason. Observations and monitoring: routine practices on the ward. *Paediatr Nurs*, 22(4):28–32, May 2010. 21

M. O. Opio, G. Nansubuga, and J. Kellett. Validation of the VitalPAC™ Early Warning Score (ViEWS) in acutely ill medical patients attending a resource-poor hospital in sub-Saharan Africa. *Resuscitation*, 84(6):743–746, Jun 2013. 66

P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2010. 132

C. Orphanidou, S. Fleming, S. A. Shah, and L. Tarassenko. Data fusion for estimating respiratory rate from a single-lead ECG. *Biomedical Signal Processing and Control*, 8(1):98 – 105, 2013. 203

L. K. Ott, M. R. Pinsky, L. A. Hoffman, S. P. Clarke, S. Clark, D. Ren, and M. Hravnak. Medical emergency team calls in the radiology department: patient characteristics and outcomes. *BMJ Qual Saf*, 21(6):509–518, Jun 2012. 21

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 115, 246

R. Paterson, D. C. MacLeod, D. Thetford, A. Beattie, C. Graham, S. Lam, and D. Bell. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clin Med*, 6(3):281–284, 2006. 64, 68, 173

R. M. Pearse, D. A. Harrison, P. James, D. Watson, C. Hinds, A. Rhodes, R. M. Grounds, and E. D. Bennett. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Crit Care*, 10(3):R81, 2006. 31

T. Pedersen, A. Nicholson, K. Hovhannisyan, A. M. Møller, A. F. Smith, and S. R. Lewis. Pulse oximetry for perioperative monitoring. *Cochrane Database Syst Rev*, 3:CD002013, 2014. 25

A. Peris, G. Zagli, N. Maccarrone, S. Batacchi, R. Cammelli, A. Cecchi, L. Perretta, and P. Bechi. The use of modified early warning score may help anesthesists in postoperative level of care selection in emergency abdominal surgery. *Minerva Anestesiol*, 78(9):1034–1038, Sep 2012. 68

J. Phua, W. J. Ngerng, and T. K. Lim. The impact of a delay in intensive care unit admission for community-acquired pneumonia. *Eur Respir J*, 36(4): 826–833, Oct 2010. 11, 12

M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. 111, 113, 122

C. Plagemann, K. Kersting, and W. Burgard. Nonstationary Gaussian process regression using point estimates of local smoothness. In *Machine learning and knowledge discovery in databases*, pages 204–219. Springer, 2008. 204

G. Priestley, W. Watson, A. Rashidian, C. Mozley, D. Russell, J. Wilson, J. Cope, D. Hart, D. Kay, K. Cowley, and J. Pateraki. Introducing Critical Care Outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Med*, 30(7):1398–1404, Jul 2004. 20

D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone. ViEWS - towards a national early warning score for detecting adult inpatient deteriora-

tion. *Resuscitation*, 81(8):932–937, Aug 2010. 19, 51, 58, 62, 63, 64, 66, 69, 71, 72, 74, 77, 85, 86, 87, 96, 137, 170, 171, 172, 198

J. A. Quinn, C. K. I. Williams, and N. McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009. 113, 176

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009. 53

A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze. Improved one-class SVM classifier for sounds classification. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 117–122, 2007. 127

L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 191

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 177, 180, 181, 182, 204

RCP. Royal College of Physicians - National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS. Report of a working party. Report (London), 2012. 64, 67, 69, 71, 74, 75, 76, 77, 83, 99, 170, 196, 197, 198, 241

J. E. Rees and C. Mann. Use of the patient at risk scores in the emergency department: a preliminary study. *Emergency Medicine Journal*, 21(6):698–699, 2004. 64

J. Renton, D. V. Pilcher, J. D. Santamaria, P. Stow, M. Bailey, G. Hart, and G. Duke. Factors associated with increased risk of readmission to intensive care in Australia. *Intensive Care Med*, 37(11):1800–1808, Nov 2011. 10

J. A. Reyes and D. Gilbert. Prediction of protein-protein interactions using one-class classification methods and integrating diverse data. *Journal of Integrative Bioinformatics*, 4(3):77, 2007. 109

A. Rhodes, P. Ferdinande, H. Flaatten, B. Guidet, P. G. Metnitz, and R. P. Moreno. The variability of critical care bed numbers in Europe. *Intensive Care Med*, 38(10):1647–1653, Oct 2012. 11

K. Rich. Inhospital cardiac arrest: pre-event variables and nursing response. *Clin Nurse Spec*, 13(3):147–53, May 1999. 41

B. Riley and R. Faleiro. Critical care outreach: rationale and development. *BJA CEPD Reviews*, 1(5):146–149, 2001. 64

G. Ritter and M. T. Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18(6):525–539, 1997. 110

R. Robert, J. Reignier, C. Tournoux-Facon, T. Boulain, O. Lesieur, V. Gissot, V. Souday, M. Hamrouni, C. Chapon, J.-P. Gouello, and Association des Réanimateurs du Centre Ouest Group. Refusal of intensive care unit admission due to a full unit: impact on mortality. *Am J Respir Crit Care Med*, 185(10): 1081–1087, May 2012. 12

A. S. C. Rocha, M. P. Araújo, A. Campos, R. Costa-Filho, E. T. Mesquita, and M. V. Santos. Circadian rhythm of hospital deaths: comparison between intensive care unit and non-intensive care unit. *Rev Assoc Med Bras*, 57(5): 529–533, 2011. 173

S. Romero-Brufau, J. M. Huddleston, J. M. Naessens, M. G. Johnson, J. Hickman, B. W. Morlan, J. B. Jensen, S. M. Caples, J. L. Elmer, J. A. Schmidt, T. I. Morgenthaler, and P. J. Santrach. Widely used track and trigger scores: Are they ready for automation in practice? *Resuscitation*, 85(4):549–552, Apr 2014. 19

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 26(1):43–49, 1978. 184

J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969. 148

C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*. Springer reference. Springer, 2011. 107

M. Sapo, S. Wu, S. Asgari, N. McNair, F. Buxey, N. Martin, and X. Hu. A comparison of vital signs charted by nurses with automated acquired values using waveform quality indices. *J Clin Monit Comput*, 23(5):263–271, Oct 2009. 94

S. Saria, D. Koller, and A. Penn. Discovering shared and individual latent structure in multiple time series. *ArXiv preprint arXiv:1008.2028*, 2010. 113, 177, 195

S. Saria, A. Duchi, and D. Koller. Discovering deformable motifs in continuous time series data. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 22, page 1465, 2011. 113, 177, 195

A. Schnegelsberg, J. Mackenhauer, M. Pedersen, H. Nibro, and H. Kirkegaard. Delayed admission to the ICU is associated with increased in-hospital mortality in patients with community-acquired severe sepsis or shock. *Critical Care*, 18 (Suppl 1):P241, 2014. 11, 12

B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, 12(3):582–588, 2000. 112, 121, 127

M. J. Schull and D. A. Redelmeier. Continuous electrocardiographic monitoring and cardiac arrest outcomes in 8,932 telemetry ward patients. *Acad Emerg Med*, 7(6):647–652, Jun 2000. 25

Scottish Executive Health Department. Better critical care. Report of short-life working group on ICU and HDU issues (Edinburgh), 2000. 1

E. Seward, E. Greig, S. Preston, R. A. Harris, Z. Borrill, T. D. Wardle, R. Burnham, P. Driscoll, B D W. Harrison, D. C. Lowe, and M. G. Pearson. A confidential study of deaths after emergency medical admission: issues relating to quality of care. *Clin Med*, 3(5):425–434, 2003. 15

S. A. Shah. *Vital sign monitoring and data fusion for paediatric triage*. PhD thesis, University of Oxford, 2012. 203

P. J. Sharek, L. M. Parast, K. Leong, J. Coombs, K. Earnest, J. Sullivan, L. R. 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*, 298 (19):2267–2274, Nov 2007. 21

J. T. Sharpley and J. C. Holden. Introducing an early warning scoring system in a district general hospital. *Nurs Crit Care*, 9(3):98–103, 2004. 18

H. Shimodaira, H. S. K.-I. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. *Advances in neural information processing systems*, 14:921, 2002. 184, 185

G. B. Smith. Have we found the perfect early warning score? A view of ViEWS. *Resuscitation*, 84(6):707–708, Jun 2013. 51, 83

G. B. Smith, D. R. Prytherch, P. Schmidt, P. I. Featherstone, D. Knight, G. Clements, and M. A. Mohammed. Hospital-wide physiological surveillance - a new approach to the early identification and management of the sick patient. *Resuscitation*, 71(1):19–28, 2006a. 1, 13, 64

G. B. Smith, D. R. Prytherch, P. Schmidt, P. I. Featherstone, D. Knight, G. Clements, and M. A. Mohammed. Hospital-wide physiological surveillance - a new approach to the early identification and management of the sick patient. *Resuscitation*, 71(1):19–28, Oct 2006b. 51

G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–179, May 2008a. 19, 39, 43, 51, 63, 65, 66, 71

G. B. Smith, D. R. Prytherch, P. E. Schmidt, P. I. Featherstone, and B. Higgins. A review, and performance evaluation, of single-parameter "track and trigger" systems. *Resuscitation*, 79(1):11–21, Oct 2008b. 19, 63, 65

G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, and P. I. Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, Apr 2013. 29

J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*, pages 261–265, 1988. 248

T. Smith, D. Den Hartog, T. Moerman, P. Patka, E. M. M. Van Lieshout, and N. W. L. Schep. Accuracy of an expanded early warning score for patients in general and trauma surgery wards. *Br J Surg*, 99(2):192–197, Feb 2012. 68

N. V. Sneed and A. D. Hollerbach. Accuracy of heart rate assessment in atrial fibrillation. *Heart Lung*, 21(5):427–433, 1992. 87, 88, 94

Y. Song, Z. Wen, C.-Y. Lin, and R. Davis. One-class conditional random fields for sequential anomaly detection. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 1685–1691. AAAI Press, 2013. 192

I. Stanculescu, C. K. I. Williams, and Y. Freer. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014. 176

O. Stegle, S. V. Fallert, D. J. C. MacKay, and S. Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008. 178

H. T. Stelfox, B. R. Hemmelgarn, S. M. Bagshaw, S. Gao, C. J. Doig, C. Nijssen-Jordan, and B. Manns. Intensive care unit bed availability and outcomes for hospitalized patients with sudden clinical deterioration. *Arch Intern Med*, 172 (6):467–474, Mar 2012. 13

J. S. Stewart. The aim and philosophy of patient monitoring. *Postgrad Med J*, 46(536):339–343, Jun 1970. 5, 6, 7

A. Stys and T. Stys. Current clinical applications of heart rate variability. *Clinical cardiology*, 21(10):719–724, 1998. 160

C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, Oct 2001. 19, 64, 85

C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell. Effect of introducing the modified early warning score on clinical outcomes, cardiopulmonary arrests and intensive care utilisation in acute medical admissions. *Anaesthesia*, 58(8):797–802, Aug 2003. 19, 64, 241

C. P. Subbe, H. Gao, and D. A. Harrison. Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Med*, 33(4):619–624, Apr 2007. 62, 64, 65, 85

A. H. Taenzer and G. T. Blike. Postoperative monitoring - The Darmouth experience. *APSF Newsletter, Spring-Summer*, pages 1–4, 2012. 202

A. H. Taenzer, J. B. Pyke, S. P. McGrath, and G. T. Blike. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: a before-and-after concurrence study. *Anesthesiology*, 112(2):282–287, Feb 2010. 25, 202

A. H. Taenzer, J. B. Pyke, and S. P. McGrath. A review of current and emerging approaches to address failure-to-rescue. *Anesthesiology*, 115(2):421–431, Aug 2011. 23, 28, 202

A. H. Taenzer, J. Pyke, M. D. Herrick, T. M. Dodds, and S. P. McGrath. A comparison of oxygen saturation data in inpatients with low oxygen saturation using automated continuous monitoring and intermittent manual data charting. *Anesth Analg*, 118(2):326–331, Feb 2014. 87, 88, 90, 91, 93, 94, 95

A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011. 186

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th International Conference on Artificial Neural Networks*, pages 442–447, 1995. 109, 110

L. Tarassenko, A. Hann, and D. Young. Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia*, 97(1): 64–68, 2006. 26, 27, 33, 39, 109, 113, 114, 118, 126

L. Tarassenko, D. A. Clifton, M. R. Pinsky, M. T. Hravnak, J. R. Woods, and P. J. Watkinson. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–1018, Aug 2011. 19, 63, 64, 66, 67, 69, 84, 86, 87, 96, 97, 198, 202, 241

L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh. Non-contact video-based vital sign monitoring using ambient light and autoregressive models. *Physiol Meas*, 35(5):807–831, May 2014. 25

D. Tax. *One-class classification: Concept-learning in the absence of counter-examples*. PhD thesis, University of Delft, The Netherlands, 2001. 122

D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11):1191–1199, 1999. 112, 122

G. Teasdale and B. Jennett. Assessment of coma and impaired consciousness. A practical scale. *Lancet*, 2(7872):81–84, Jul 1974. 44

R. Teplick and A. E. Anderson. Rapid response systems: move a bit more slowly. *Crit Care Med*, 34(9):2507–2509, Sep 2006. 19

J. Tibballs and S. Kinney. Reduction of hospital mortality and of preventable cardiac arrest and death on introduction of a pediatric medical emergency team. *Pediatr Crit Care Med*, 10(3):306–312, May 2009. 21

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001. 187, 264, 265

M. E. Tipping and D. Lowe. Shadow targets: a novel algorithm for topographic projections by radial basis functions. *Neurocomputing*, 19(1):211–222, 1998. 148

R. Veldhuis. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002. 145

R. Vert and J.-P. Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *The Journal of Machine Learning Research*, 7:817–854, 2006. 131

I. Villegas, I. C. Arias, A. Botero, and A. Escobar. Evaluation of the technique used by health-care workers for taking blood pressure. *Hypertension*, 26(6): 1204–1206, Dec 1995. 87, 88

P. Vincent and Y. Bengio. Manifold parzen windows. *Advances in Neural Information Processing Systems*, 15:825–832, 2002. 246, 251

A. Vlayen, S. Verelst, G. E. Bekkering, W. Schrooten, J. Hellings, and N. Claes. Incidence and preventability of adverse events requiring intensive care admission: a systematic review. *J Eval Clin Pract*, 18(2):485–497, Apr 2012. 11

A. D. Waller. A demonstration on man of electromotive changes accompanying the heart's beat. *J Physiol*, 8(5):229–234, Oct 1887. 6

P. J. Watkinson, V. S. Barber, J. D. Price, A. Hann, L. Tarassenko, and J. D. Young. A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients. *Anaesthesia*, 61(11):1031–1039, Nov 2006. 26, 202

J. B. Waugh, C. A. Epps, and Y. A. Khodneva. Capnography enhances surveillance of respiratory events during procedural sedation: a meta-analysis. *J Clin Anesth*, 23(3):189–196, May 2011. 25

K. Wilkinson. An age-old problem: care of older people undergoing surgery. *Br J Hosp Med (London)*, 72(3):126–127, Mar 2011. 31

J. G. Williams, S. E. Roberts, M. F. Ali, W. Y. Cheung, D. R. Cohen, G. Demery, A. Edwards, M. Greer, M. D. Hellier, H. A. Hutchings, B. Ip, M. F. Longo, I. T. Russell, H. A. Snooks, and J. C. Williams. Gastroenterology services in the uk. the burden of disease, and the organisation and delivery of services for gastrointestinal and liver disorders: a review of the evidence. *Gut*, 56 Suppl 1: 1–113, Feb 2007. 31

A. S. Willsky, E. B. Sudderth, M. I. Jordan, and E. B. Fox. Sharing features among dynamical systems with Beta processes. In *Advances in Neural Information Processing Systems*, pages 549–557. 2009a. 176, 195

A. S Willsky, E. B Sudderth, M. I. Jordan, and E. B. Fox. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 457–464, 2009b. 176

S. J. Wilson, D. Wong, D. Clifton, S. Fleming, R. Way, R. Pullinger, and L. Tarassenko. Track and trigger in an emergency department: an observational evaluation study. *Emergency Medicine Journal*, 30(3):186–191, 2013. 21

B. D. Winters, S. J. Weaver, E. R. Pfoh, T. Yang, J. C. Pham, and S. M. Dy. Rapid-response systems as a patient safety strategy: a systematic review. *Ann Intern Med*, 158(5):417–425, Mar 2013. 15, 16, 19, 21

D. Wong. *Identifying Vital Sign Abnormality in Acutely-ill Patients*. PhD thesis, University of Oxford, 2011. 41, 71, 109, 114, 119, 120, 128, 133, 172, 178, 204

D. Wong, D. A. Clifton, and L. Tarassenko. Probabilistic detection of vital sign abnormality with Gaussian process regression. In *IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, pages 187–192, 2012. 178

M. M. Wright, C. W. Stenhouse, and R. J. M. Morgan. Early detection of patients at risk (PART). *Anaesthesia*, 55(4):391–392, 2000. 64

M. W. Wukitsch, M. T. Petterson, D. R. Tobler, and J. A. Pologe. Pulse oximetry: analysis of theory, technology, and practice. *J Clin Monit*, 4(4):290–301, Oct 1988. 42

H. Xiao, H. Xiao, and C. Eckert. Learning from multiple observers with unknown expertise. In *Advances in Knowledge Discovery and Data Mining*, pages 595–606. Springer, 2013. 203

M. P. Young, V. J. Gooder, K. McBride, B. James, and E. S. Fisher. Inpatient transfers to the intensive care unit: delays are associated with increased mortality and morbidity. *J Gen Intern Med*, 18(2):77–83, Feb 2003. 11

Appendix A - Performance of EWS systems

The cutoff values for the variables included in example EWS systems (including CEWS) are reperesented in Table 1. Tables 2, 3, 4, and 5 correspond to tables presented in chapters 3 and 4. In these chapters, for ease of reading, we omitted the standard error values associated to each metric. Here, the mean values obtained for each metric are shown together with the associated standard error values. We note that, unless otherwise stated, the standard error of the area under the curve and the other performance measures was estimated directly from the data using the standard error of the Wilcoxon statistic, as described in Hanley and McNeil [1982] and Hajian-Tilaki and Hanley [2002].

Table 1: More examples of early warning score systems.

CEWS (Tarassenko et al. [2011])

	Score						
Variable	3	2	1	0	1	2	3
Heart Rate	≤ 42	$43-49$	$50-53$	$54-104$	$105-112$	$113-127$	≥ 128
Resp. Rate	≤ 7	$8-10$	$11-13$	$14-25$	$26-28$	$29-33$	≥ 34
Temperature	≤ 35.4		$35.5-35.9$	$36.0-37.3$	$37.4-38.3$		≥ 38.4
Systolic BP	≤ 85	$86-96$	$97-101$	$102-154$	$155-164$	$165-184$	≥ 185
SpO_2	≤ 84	$85-90$	$91-93$	≥ 94			
Inspired O_2							
AVPU scale				A	V		P, U

NEWS (RCP [2012])

	Score						
Variable	3	2	1	0	1	2	3
Heart Rate		≤ 40	$41-50$	$51-90$	$91-110$	$111-130$	≥ 131
Resp. Rate	≤ 8		$9-11$	$12-20$		$21-24$	≥ 25
Temperature	≤ 35.0		$35.1-36.0$	$36.1-38.0$	$38.1-39.0$	≥ 39.1	
Systolic BP	≤ 90	$81-100$	$101-110$	$111-219$			≥ 220
SpO_2	≤ 91	$92-93$	$94-95$	≥ 96			
Inspired O_2				Air		Any O_2	
AVPU scale				A			V, P, U

MEWS (Subbe et al. [2003])

	Score						
Variable	3	2	1	0	1	2	3
Heart Rate		≤ 39	$40-49$	$50-100$	$101-110$	$111-130$	≥ 131
Resp. Rate			≤ 8	$9-14$	$15-20$	$21-29$	≥ 30
Temperature			≤ 35.0	$35.1-38.5$		≥ 38.6	
Systolic BP	≤ 69	$70-79$	$80-99$	$100-200$		≥ 201	
SpO_2	≤ 91	$92-93$	$94-95$	≥ 96			
Inspired O_2				Air		Any O_2	
AVPU scale				A	V	P	U

MEWS (Lilienfeld-Toal et al. [2007])

	Score						
Variable	3	2	1	0	1	2	3
Heart Rate		≤ 40	$41-50$	$51-100$	$101-110$	$111-130$	≥ 131
Resp. Rate		≤ 8	$9-11$	$12-20$	$21-25$	$26-30$	≥ 31
Systolic BP	≤ 70	$71-79$	$80-99$	$100-179$		≥ 220	
SpO_2	≤ 85	$86-89$	$90-94$	≥ 95			
Inspired O_2				Air	Any O_2		
AVPU scale				A			V, P, U

Table 2: Performance metrics for the 26 track-and-trigger systems evaluated in the 34,060 patient-episodes in the Portsmouth dataset (with the outcome being death within 24 hours). The results are presented in descending order of AUROC, and the best values for each performing metric are underlined. Values are presented with mean and standard error of the Wilcoxon statistic. (Corresponds to Table 3.7, on page 75)

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[26]	0.881 (0.006)	0.827 (0.006)	0.780 (0.010)	0.834 (0.001)	0.037 (0.001)	0.146
[25]	0.877 (0.006)	0.854 (0.006)	0.841 (0.009)	0.760 (0.001)	0.028 (0.001)	0.125
[23]	0.877 (0.006)	0.862 (0.006)	0.793 (0.010)	0.815 (0.001)	0.034 (0.001)	0.139
[21]	0.854 (0.006)	0.777 (0.007)	0.721 (0.011)	0.826 (0.001)	0.033 (0.001)	0.128
[19]	0.836 (0.006)	0.728 (0.007)	0.772 (0.011)	0.758 (0.001)	0.025 (0.001)	0.110
[3]	0.835 (0.006)	0.712 (0.007)	0.771 (0.011)	0.771 (0.001)	0.027 (0.001)	0.115
[9]	0.827 (0.006)	0.752 (0.007)	0.780 (0.010)	0.719 (0.001)	0.022 (0.001)	0.099
[20]	0.827 (0.006)	0.744 (0.007)	0.780 (0.010)	0.728 (0.001)	0.023 (0.001)	0.102
[12]	0.826 (0.006)	0.680 (0.007)	0.703 (0.011)	0.816 (0.001)	0.030 (0.001)	0.119
[11]	0.815 (0.007)	0.799 (0.007)	0.835 (0.009)	0.656 (0.001)	0.019 (0.001)	0.092
[15]	0.814 (0.007)	0.798 (0.007)	0.835 (0.009)	0.657 (0.001)	0.019 (0.001)	0.092
[14]	0.814 (0.007)	0.798 (0.007)	0.835 (0.009)	0.657 (0.001)	0.019 (0.001)	0.092
[5]	0.812 (0.007)	0.764 (0.007)	0.806 (0.010)	0.695 (0.001)	0.021 (0.001)	0.097
[10]	0.811 (0.007)	0.748 (0.007)	0.765 (0.011)	0.715 (0.001)	0.021 (0.001)	0.095
[22]	0.811 (0.007)	0.739 (0.007)	0.766 (0.011)	0.723 (0.001)	0.022 (0.001)	0.097
[18]	0.810 (0.007)	0.829 (0.006)	0.647 (0.012)	0.838 (0.001)	0.032 (0.001)	0.117
[6]	0.809 (0.007)	0.779 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[13]	0.809 (0.007)	0.778 (0.007)	0.813 (0.010)	0.679 (0.001)	0.020 (0.001)	0.094
[1]	0.809 (0.007)	0.779 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[2]	0.809 (0.007)	0.778 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[4]	0.809 (0.007)	0.785 (0.007)	0.821 (0.010)	0.668 (0.001)	0.020 (0.001)	0.093
[17]	0.808 (0.007)	0.776 (0.007)	0.814 (0.010)	0.680 (0.001)	0.020 (0.001)	0.094
[7]	0.808 (0.007)	0.775 (0.007)	0.814 (0.010)	0.680 (0.001)	0.020 (0.001)	0.094
[24]	0.796 (0.007)	0.746 (0.007)	0.750 (0.011)	0.711 (0.001)	0.021 (0.001)	0.090
[16]	0.780 (0.007)	0.749 (0.007)	0.751 (0.011)	0.685 (0.001)	0.019 (0.001)	0.084
[8]	0.779 (0.007)	0.668 (0.008)	0.640 (0.012)	0.815 (0.001)	0.027 (0.001)	0.104

Table 3: Performance metrics for the twenty-six track-and-trigger systems evaluated in the 407 patients in the CALMS-2 dataset (for the combined outcome of cardiac arrest, unanticipated admission to ICU and death occurring within 24 hours of a given observation set). The results are presented in descending order of AUROC, and the best values for each performing metric are underlined. Values are presented with mean and standard error of the Wilcoxon statistic. (Corresponds to Table 3.10, on page 80)

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[23]	0.841 (0.008)	0.826 (0.008)	0.801 (0.014)	0.765 (0.002)	0.084 (0.003)	0.209
[25]	0.835 (0.008)	0.747 (0.010)	0.829 (0.013)	0.715 (0.003)	0.073 (0.003)	0.190
[21]	0.833 (0.008)	0.753 (0.009)	0.639 (0.016)	0.886 (0.002)	0.131 (0.005)	0.251
[26]	0.829 (0.008)	0.816 (0.009)	0.773 (0.014)	0.768 (0.002)	0.083 (0.003)	0.201
[22]	0.791 (0.009)	0.807 (0.009)	0.711 (0.015)	0.791 (0.002)	0.084 (0.003)	0.193
[10]	0.786 (0.009)	0.677 (0.010)	0.700 (0.016)	0.784 (0.002)	0.081 (0.003)	0.185
[12]	0.784 (0.009)	0.825 (0.008)	0.622 (0.016)	0.872 (0.002)	0.116 (0.005)	0.227
[9]	0.782 (0.009)	0.680 (0.010)	0.697 (0.016)	0.777 (0.002)	0.078 (0.003)	0.179
[4]	0.782 (0.009)	0.720 (0.010)	0.778 (0.014)	0.730 (0.002)	0.072 (0.003)	0.180
[20]	0.781 (0.009)	0.678 (0.010)	0.697 (0.016)	0.779 (0.002)	0.078 (0.003)	0.180
[5]	0.779 (0.009)	0.703 (0.010)	0.752 (0.015)	0.751 (0.002)	0.075 (0.003)	0.183
[6]	0.777 (0.009)	0.713 (0.010)	0.762 (0.014)	0.738 (0.002)	0.073 (0.003)	0.179
[17]	0.777 (0.009)	0.712 (0.010)	0.757 (0.015)	0.742 (0.002)	0.073 (0.003)	0.180
[1]	0.776 (0.009)	0.714 (0.010)	0.762 (0.014)	0.736 (0.002)	0.072 (0.003)	0.178
[7]	0.776 (0.009)	0.711 (0.010)	0.757 (0.015)	0.742 (0.002)	0.073 (0.003)	0.180
[2]	0.776 (0.009)	0.712 (0.010)	0.758 (0.015)	0.739 (0.002)	0.073 (0.003)	0.178
[8]	0.776 (0.009)	0.618 (0.010)	0.603 (0.017)	0.875 (0.002)	0.115 (0.005)	0.221
[13]	0.770 (0.009)	0.710 (0.010)	0.752 (0.015)	0.736 (0.002)	0.072 (0.003)	0.175
[18]	0.766 (0.009)	0.623 (0.010)	0.593 (0.017)	0.861 (0.002)	0.103 (0.004)	0.203
[19]	0.764 (0.009)	0.782 (0.009)	0.570 (0.017)	0.845 (0.002)	0.090 (0.004)	0.179
[15]	0.764 (0.009)	0.802 (0.009)	0.565 (0.017)	0.885 (0.002)	0.117 (0.005)	0.216
[11]	0.764 (0.009)	0.803 (0.009)	0.801 (0.014)	0.612 (0.003)	0.053 (0.002)	0.135
[14]	0.763 (0.009)	0.802 (0.009)	0.801 (0.014)	0.613 (0.003)	0.053 (0.002)	0.135
[24]	0.759 (0.009)	0.693 (0.010)	0.702 (0.016)	0.748 (0.002)	0.070 (0.003)	0.163
[16]	0.744 (0.010)	0.609 (0.010)	0.522 (0.017)	0.868 (0.002)	0.096 (0.004)	0.178
[3]	0.717 (0.010)	0.760 (0.009)	0.501 (0.017)	0.898 (0.002)	0.117 (0.005)	0.202

Table 4: Performance metrics for the EWS systems studied, including the two modified versions of CEWS (system [24]), evaluated using the Portsmouth dataset. The results are presented in descending order of AUROC, and the best values for each performing metric are underlined. Values are presented with mean and standard error of the Wilcoxon statistic. (Corresponds to Table 4.3, on page 102)

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[26]	0.881 (0.006)	0.827 (0.006)	0.780 (0.010)	0.834 (0.001)	0.037 (0.001)	0.146
[25]	0.877 (0.006)	0.854 (0.006)	0.841 (0.009)	0.760 (0.001)	0.028 (0.001)	0.125
[23]	0.877 (0.006)	0.862 (0.006)	0.793 (0.010)	0.815 (0.001)	0.034 (0.001)	0.139
[24]^3	0.856 (0.009)	0.855 (0.012)	0.731 (0.001)	0.835 (0.002)	0.030 (0.0003)	0.133
[21]	0.854 (0.006)	0.777 (0.007)	0.721 (0.011)	0.826 (0.001)	0.033 (0.001)	0.128
[19]	0.836 (0.006)	0.728 (0.007)	0.772 (0.011)	0.758 (0.001)	0.025 (0.001)	0.110
[3]	0.835 (0.006)	0.712 (0.007)	0.771 (0.011)	0.771 (0.001)	0.027 (0.001)	0.115
[9]	0.827 (0.006)	0.752 (0.007)	0.780 (0.010)	0.719 (0.001)	0.022 (0.001)	0.099
[20]	0.827 (0.006)	0.744 (0.007)	0.780 (0.010)	0.728 (0.001)	0.023 (0.001)	0.102
[12]	0.826 (0.006)	0.680 (0.007)	0.703 (0.011)	0.816 (0.001)	0.030 (0.001)	0.119
[24]^2	0.817 (0.008)	0.790 (0.020)	0.684 (0.001)	0.812 (0.002)	0.023 (0.002)	0.115
[11]	0.815 (0.007)	0.799 (0.007)	0.835 (0.009)	0.656 (0.001)	0.019 (0.001)	0.092
[15]	0.814 (0.007)	0.798 (0.007)	0.835 (0.009)	0.657 (0.001)	0.019 (0.001)	0.092
[14]	0.814 (0.007)	0.798 (0.007)	0.835 (0.009)	0.657 (0.001)	0.019 (0.001)	0.092
[5]	0.812 (0.007)	0.764 (0.007)	0.806 (0.010)	0.695 (0.001)	0.021 (0.001)	0.097
[10]	0.811 (0.007)	0.748 (0.007)	0.765 (0.011)	0.715 (0.001)	0.021 (0.001)	0.095
[22]	0.811 (0.007)	0.739 (0.007)	0.766 (0.011)	0.723 (0.001)	0.022 (0.001)	0.097
[18]	0.810 (0.007)	0.829 (0.006)	0.647 (0.012)	0.838 (0.001)	0.032 (0.001)	0.117
[6]	0.809 (0.007)	0.779 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[13]	0.809 (0.007)	0.778 (0.007)	0.813 (0.010)	0.679 (0.001)	0.020 (0.001)	0.094
[1]	0.809 (0.007)	0.779 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[2]	0.809 (0.007)	0.778 (0.007)	0.818 (0.010)	0.677 (0.001)	0.020 (0.001)	0.094
[4]	0.809 (0.007)	0.785 (0.007)	0.821 (0.010)	0.668 (0.001)	0.020 (0.001)	0.093
[17]	0.808 (0.007)	0.776 (0.007)	0.814 (0.010)	0.680 (0.001)	0.020 (0.001)	0.094
[7]	0.808 (0.007)	0.775 (0.007)	0.814 (0.010)	0.680 (0.001)	0.020 (0.001)	0.094
[24]^1	0.796 (0.007)	0.746 (0.007)	0.750 (0.011)	0.711 (0.001)	0.021 (0.001)	0.090
[16]	0.780 (0.007)	0.749 (0.007)	0.751 (0.011)	0.685 (0.001)	0.019 (0.001)	0.084
[8]	0.779 (0.007)	0.668 (0.008)	0.640 (0.012)	0.815 (0.001)	0.027 (0.001)	0.104

[24]^1 Original CEWS system; [24]^2 CEWS system with modified limits for RR; [24]^3 CEWS system with modified limits for RR and with an additional score of 2 for any oxygen support.

Table 5: Performance metrics for the EWS systems studied, including the two modified versions of CEWS (system [24]), evaluated using the CALMS-2 dataset. Results are presented in descending order of AUROC, and the best values for each performing metric are underlined. Values are presented with mean and standard error of the Wilcoxon statistic. (Corresponds to Table 4.4, on page 103)

No.	AUROC	pAUROC	Sens.	Spec.	PPV	MCC
[23]	0.841 (0.008)	0.826 (0.008)	0.801 (0.014)	0.765 (0.002)	0.084 (0.003)	0.209
[24]^3	0.839 (0.008)	0.830 (0.009)	0.789 (0.014)	0.797 (0.002)	0.132 (0.003)	0.251
[25]	0.835 (0.008)	0.747 (0.010)	0.829 (0.013)	0.715 (0.003)	0.073 (0.003)	0.190
[21]	0.833 (0.008)	0.753 (0.009)	0.639 (0.016)	0.886 (0.002)	0.131 (0.005)	0.251
[26]	0.829 (0.008)	0.816 (0.009)	0.773 (0.014)	0.768 (0.002)	0.083 (0.003)	0.201
[24]^2	0.796 (0.009)	0.737 (0.010)	0.653 (0.016)	0.860 (0.002)	0.112 (0.004)	0.228
[22]	0.791 (0.009)	0.807 (0.009)	0.711 (0.015)	0.791 (0.002)	0.084 (0.003)	0.193
[10]	0.786 (0.009)	0.677 (0.010)	0.700 (0.016)	0.784 (0.002)	0.081 (0.003)	0.185
[12]	0.784 (0.009)	0.825 (0.008)	0.622 (0.016)	0.872 (0.002)	0.116 (0.005)	0.227
[9]	0.782 (0.009)	0.680 (0.010)	0.697 (0.016)	0.777 (0.002)	0.078 (0.003)	0.179
[4]	0.782 (0.009)	0.720 (0.010)	0.778 (0.014)	0.730 (0.002)	0.072 (0.003)	0.180
[20]	0.781 (0.009)	0.678 (0.010)	0.697 (0.016)	0.779 (0.002)	0.078 (0.003)	0.180
[5]	0.779 (0.009)	0.703 (0.010)	0.752 (0.015)	0.751 (0.002)	0.075 (0.003)	0.183
[6]	0.777 (0.009)	0.713 (0.010)	0.762 (0.014)	0.738 (0.002)	0.073 (0.003)	0.179
[17]	0.777 (0.009)	0.712 (0.010)	0.757 (0.015)	0.742 (0.002)	0.073 (0.003)	0.180
[1]	0.776 (0.009)	0.714 (0.010)	0.762 (0.014)	0.736 (0.002)	0.072 (0.003)	0.178
[7]	0.776 (0.009)	0.711 (0.010)	0.757 (0.015)	0.742 (0.002)	0.073 (0.003)	0.180
[2]	0.776 (0.009)	0.712 (0.010)	0.758 (0.015)	0.739 (0.002)	0.073 (0.003)	0.178
[8]	0.776 (0.009)	0.618 (0.010)	0.603 (0.017)	0.875 (0.002)	0.115 (0.005)	0.221
[13]	0.770 (0.009)	0.710 (0.010)	0.752 (0.015)	0.736 (0.002)	0.072 (0.003)	0.175
[18]	0.766 (0.009)	0.623 (0.010)	0.593 (0.017)	0.861 (0.002)	0.103 (0.004)	0.203
[19]	0.764 (0.009)	0.782 (0.009)	0.570 (0.017)	0.845 (0.002)	0.090 (0.004)	0.179
[15]	0.764 (0.009)	0.802 (0.009)	0.565 (0.017)	0.885 (0.002)	0.117 (0.005)	0.216
[11]	0.764 (0.009)	0.803 (0.009)	0.801 (0.014)	0.612 (0.003)	0.053 (0.002)	0.135
[14]	0.763 (0.009)	0.802 (0.009)	0.801 (0.014)	0.613 (0.003)	0.053 (0.002)	0.135
[24]^1	0.759 (0.009)	0.693 (0.010)	0.702 (0.016)	0.748 (0.002)	0.070 (0.003)	0.163
[16]	0.744 (0.010)	0.609 (0.010)	0.522 (0.017)	0.868 (0.002)	0.096 (0.004)	0.178
[3]	0.717 (0.010)	0.760 (0.009)	0.501 (0.017)	0.898 (0.002)	0.117 (0.005)	0.202

[24]^1 Original CEWS system; [24]^2 CEWS system with modified limits for RR; [24]^3 CEWS system with modified limits for RR and with an additional score of 2 for any oxygen support.

Appendix B - Density estimation: optimising the kernel size

Kernel density estimates (referred to as Parzen windowing) is a generalization of the histogram technique, in which smoother membership functions are used instead of the rectangular and sharp volumes typically used in histogram binning (Parzen [1962]). Although asymptotically Parzen windowing can yield unbiased and consistent estimators, in the finite sample case, selecting the kernel function and the kernel size become a challenging problem. Particularly in multidimensional density estimation, the full covariance matrix of the kernel must be optimised (assuming elliptically symmetric kernels). Some methods go even further by incorporating the nearest-neighbour density estimation approach and try to optimise the kernel size (or covariance) for each sample based on its nearest neighbours' distances (Bengio et al. [2006]; Vincent and Bengio [2002]). Unfortunatelly, some of these methods become intractable and ineffective when it comes to adaptive learning and signal processing, due to increasing computational complexity, as well as discontinuities in gradients introduced by switching neighbours.

In this experiment, we are interested in a fair assessment of how competitive are these different approaches proposed in the literature to determine the kernel size for estimating the density of the data. Different methods are compared using different datasets, that are described next.

Table 6: Sources of the non-synthetic datasets used: we summarise the original source of each dataset.

Name	Source
Birth-weight	http://people.reed.edu/~jones/141/BirthWgt.html
Diabetes	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
Housing	https://archive.ics.uci.edu/ml/datasets/Housing

Datasets

We considered six datasets (three synthetic and three real datasets[1]) frequently used in machine learning (Table 6). The datasets differ in their domain of application, their dimensionality D and their number of instances N. In the following experiment, we do not use the labels.

Sine-wave The first dataset (represented in Figure 1a) corresponds to data points generated from the following distribution of two-dimensional (x, y) points:

$$x = t, \;\; y = 3\sin(t) + \varepsilon_y$$

where $t \sim \mathcal{U}(1, 12)$, $\varepsilon_y \sim \mathcal{N}(0, 0.2^2)$, $\mathcal{U}(a, b)$ is uniform in the interval (a, b), and $\mathcal{N}(\mu, \sigma^2)$ is a normal density.

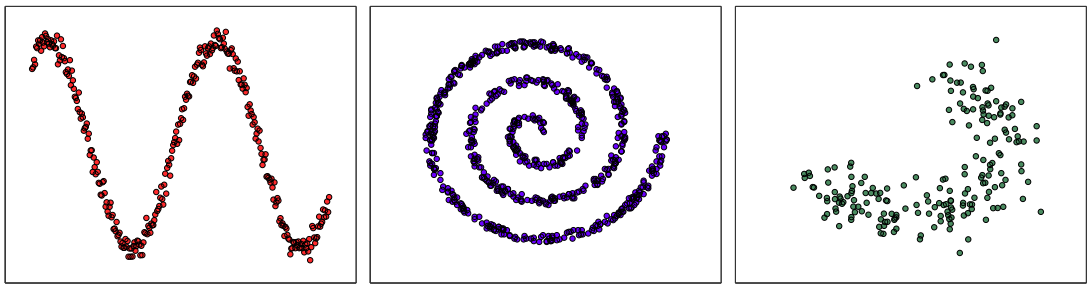
Spiral The second artificial dataset contains a uniform sampling of a single spiral (Figure 1b), with added Gaussian noise proportional to the radial distance. For that, we use the following distribution of two-dimensional (x, y) points to generate the data:

$$x = 0.04t\sin(t) + \varepsilon_x, \;\; y = 0.04t\cos(t) + \varepsilon_y$$

where $t \sim \mathcal{U}(3, 15)$ and $\varepsilon_x, \varepsilon_y \sim \mathcal{N}(0, 0.01^2)$.

Moon The third synthetic dataset (Figure 1c) was generated by sampling uniformly from one arc of circumference (one single cluster), with Gaussian noise of a fixed amplitude added to the radial direction, according to the following

[1] UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html)



(a) *Sine-wave*, N = 300 (b) *Spiral*, N = 700 (c) *Moon*, N = 200

Figure 1: Representation of the three synthetic datasets and correspondent size (N).

two-dimensional distribution:

$$x = 5\sin(1.25\pi t) + \epsilon_x, \;\; y = 5\cos(1.25\pi t)$$

where $t \sim \mathcal{U}(1, 10)$ and $\varepsilon_x \sim \mathcal{N}(0, 1)$.

Birth-weight This dataset contains various attributes of pregnant women that are associated with low birth-weight (less than 2500 grams), and is typically used to create a model which predicts whether or not a mother will give birth to an under-weight baby. The dataset attributes include the age of the mother, the mother's smoking status during pregnancy, number of premature labours, among others. For this experiment, three numerical attributes were selected.

Diabetes We also consider the *Pima Indians Diabetes* dataset which includes attributes of female patients of Pima Indian heritage. This dataset has been used for forecasting the onset of diabetes within five years in Pima Indian women (Smith et al. [1988]). The attributes contain variables that have been found to be significant risk factors for diabetes among different populations, such as diastolic blood pressure, body mass index and glucose concentration in the blood. The six numerical features in the dataset were used in this experiment.

Housing This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University (Harrison Jr and Rubinfeld [1978]), and concerns housing values in suburbs of Boston, USA. For the experiment, ten continuous attributes (excluding the class attribute) were used.

Baseline approaches

Given a finite dataset consisting of N data samples $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}$, where the feature vector variable $\mathbf{x}_i \in \mathbb{R}^D$ follows an unknown pdf $p(\mathbf{x})$, the problem under study is to estimate $p(\mathbf{x})$ based on $\mathbf{X}$.

A general kernel-based density estimate of $p(\mathbf{x})$ is given by

$$\hat{p}(\mathbf{x}|\mathbf{g}, \boldsymbol{\sigma}) = \sum_{i=1}^{N} g_i K(\mathbf{x}, \mathbf{x}_i|\sigma_i) \tag{1}$$

where g_j's are the kernel weights, subject to $g_i \geq 0$ for $i = 1, ..., N$ and $\mathbf{g} = [g_1, ..., g_N]$ with $\mathbf{g}^T\mathbf{1} = 1$, in which $\mathbf{1}$ is a vector with an appropriate dimension and all elements as ones; $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_N]$ is the kernel width vector; $K(\mathbf{x}, \mathbf{x}_i, \sigma_i)$ is a chosen kernel function with kernel width σ_i. A Gaussian kernel

$$K(\mathbf{x}, \mathbf{x}_i|\sigma_i) = \frac{1}{(2\pi\sigma_i^2)^{D/2}} \exp\left(-\frac{||\mathbf{x} - \mathbf{x_i}||^2}{2\sigma_i^2}\right) \tag{2}$$

is typically used. Let the well-known Parzen window estimator be denoted by $\hat{p}(\mathbf{x}|\mathbf{g}, \boldsymbol{\sigma})$, where $\mathbf{g} = [\mathrm{g}_1, ..., \mathrm{g}_N]$ and $\mathrm{g}_i = 1/N, \forall i$.

Most approaches proposed in the literature consider an isotropic kernel (with a single *bandwidth* for all dimensions of the data) for all data points to estimate the density of the data (i.e., $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_N]$ and $\sigma_i = \sigma, \forall i.$). Here, we briefly introduce different methods proposed in the literature to estimate σ and estimate the density of the data.

Maximum leave-one-out likelihood The log-likelihood of the observed data, $E_{\mathbf{x}}[\log \hat{p}(\mathbf{x}|g, \sigma)]$, may be approximated by the sample mean, resulting in

$$J(\sigma) = \frac{1}{N} \sum_{i=1}^{N} \log \hat{p}(\mathbf{x}|g, \sigma) \tag{3}$$

For Parzen windowing this becomes

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{1}{N}\sum_{i=1}^{N}K(\mathbf{x}_j,\mathbf{x}_i|\sigma)\right) \tag{4}$$

If a symmetric kernel function (such as the Gaussian kernel) is used, this criterion exhibits an undesirable global maximum at the null kernel size, since as σ approaches zero, the kernel approaches a Dirac$-\delta$ function and the criterion attains a value of infinity (Erdogmus et al. [2004]). To avoid this situation, the criterion needs to be modified in accordance with the leave-one-out technique. This yields

$$J(\sigma) \approx \frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{1}{N-1}\sum_{i=1,i\neq j}^{N}K(\mathbf{x}_j,\mathbf{x}_i|\sigma)\right) \tag{5}$$

The value of σ can then be chosen to maximise $J(\sigma)$. We use a simple gradient descent method to find the best parameter $\sigma \in \mathbb{R}_+$.

Average nearest-neighbour The global kernel width parameter may also be set using the heuristic suggested by Bishop [1994], who calculates the mean of the local estimates of the variance at each data point location, as such:

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m}\sum_{j\in Q_i}||\mathbf{x}_i-\mathbf{x}_j||\right) \tag{6}$$

where Q_i contains the set of the m nearest neighbours to data point $\mathbf{x}_i$. In the original paper and recent publications in which this heuristic approach was used, m was set to 10.

Mean integrated square error Based on the principle of minimising the mean integrated square error (MISE), σ can be found so as to minimise the least

squares cross-validation criterion M given by

$$M(\sigma) = \frac{1}{N^2}\sum_{i,j=1}^{N} K(\mathbf{x}_i, \mathbf{x}_j|\sqrt{2}\sigma) - \frac{2}{N(N-1)}\sum_{i,j=1;j\neq i}^{N} K(\mathbf{x}_i, \mathbf{x}_j|\sigma)$$
$$\approx \frac{1}{N^2}\sum_{i,j=1}^{N} K^*(\mathbf{x}_i, \mathbf{x}_j|\sigma) + \frac{2}{(2\pi\sigma^2)^{d/2}} \tag{7}$$

where $K^*(\mathbf{x}_i, \mathbf{x}_j|\sigma) = K(\mathbf{x}_i, \mathbf{x}_j|\sqrt{2}\sigma) - 2K(\mathbf{x}_i, \mathbf{x}_j|\sigma)$. This method has been described in Hong et al. [2008]. We employ gradient descent to find the parameter that minimises $M(\sigma)$.

Diagonal KDE The kernel density estimation procedure fits a mixture model by centring one mixture component at each data point $\mathbf{x}_i$. Indenpendent multivariate Gaussians can be used, where the diagonal widths $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_d\}$ of the kernel can be set to maximise the leave-one-out likelihood as in eq. 5,

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{N-1}\sum_{i=1,i\neq j}^{N} K(\mathbf{x}_j, \mathbf{x}_i|\boldsymbol{\sigma})\right) \tag{8}$$

We employ a Newton-scheme to find the best parameters $\boldsymbol{\sigma} \in \mathbb{R}_+^D$.

Manifold Parzen window The manifold Parzen window estimator (Vincent and Bengio [2002]) attempts to capture locality by means of a Gaussian kernel k. It is a mixture of N full Gaussians (no longer isotropic) where the covariance $\boldsymbol{\Sigma}_i = \omega\mathbf{I} + (\sum_{i\neq j} k(\mathbf{x}_i, \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top)/(\sum_{i\neq j} k(\mathbf{x}_i, \mathbf{x}_j))$ of each data point is only computed based on neighbouring data points (Figure 2). As proposed by the authors, we use the r-nearest neighbours-based kernel and do not store full covariance matrices $\boldsymbol{\Sigma}_i$ but a low rank approximation $\boldsymbol{\Sigma}_i \approx \omega\mathbf{I} + \mathbf{V}\mathbf{V}^\top$ with $\mathbf{V} \in \mathbb{R}^{D\times d}$ $(d < D)$. The ridge parameter ω was set to maximise the leave-one-out likelihood (as in the first method described). For each dataset of size N, the number of latent dimensions d, chosen from $d = [D \cdot \{10, 20, 30, 40\}/100]$, and

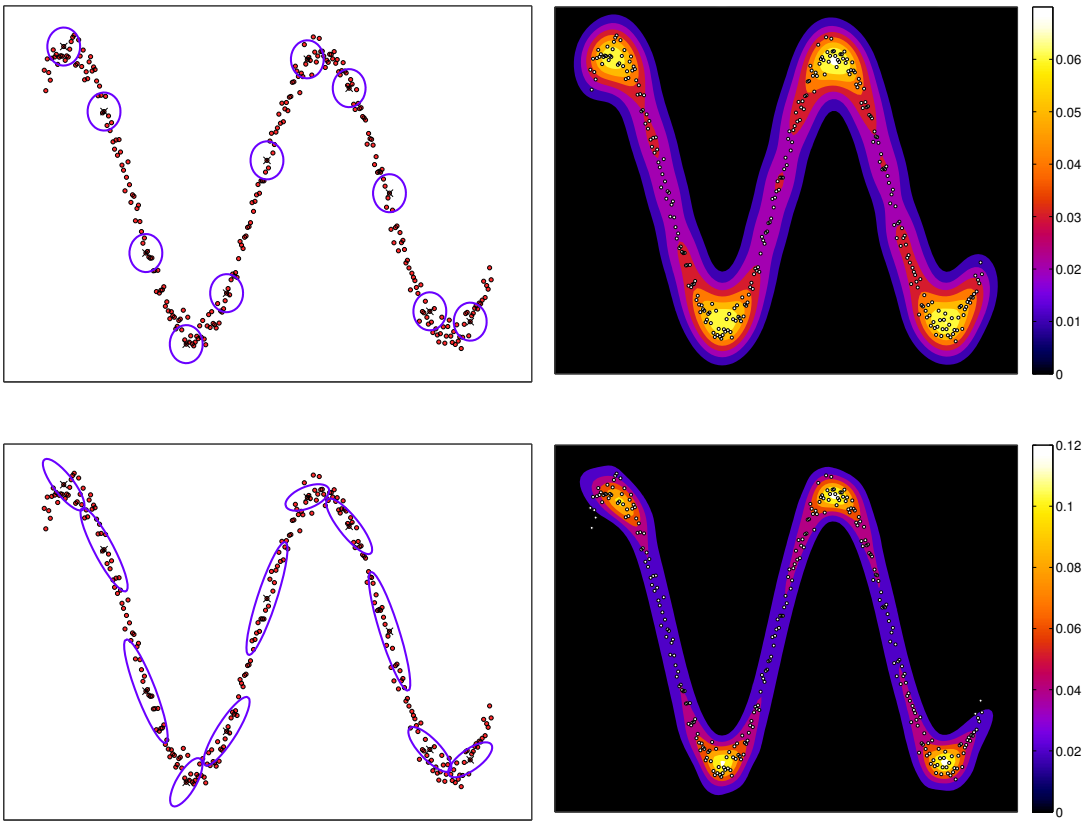


Figure 2: Comparison between the isotropic Parzen windowing method (in this case, σ was optimised using the maximum leave-one-out likelihood method), in the top, and the manifold Parzen window method, in the bottom. In the left, data are represented (red dots) together with the kernels for 11 randomly selected data points: the center of the kernel is marked with a cross '$\times$', and the magenta line denotes the kernel size. Note that in the manifold Parzen window, data points have different kernels which adapt to the directionality of the data (according to the neighbourhood of each data point). In the right, the data density estimated with each method is shown. The colour code denoting high and low-density regions are shown in the bars in the right: black corresponds to low-density areas, and white corresponds to highly dense regions.

the neighbourhood size r, chosen from $r = [N \cdot \{5, 10, 15, 20, 25, 30\}/100]$, were also optimised according to the maximum likelihood of the training data using a grid-search strategy.

Sparse KDE A disadvantage associated with the conventional Parzen window method is its high computational cost of the point density estimate for a future

data sample in the cases whereby the training dataset is very large. Clearly, by taking a much smaller number of mixture components, the finite mixture model can be regarded as a condensed representation of data (Hong et al. [2008]). In order to speed up computation, the training dataset may be partitioned into n disjoint subsets using the K-means algorithm. The resulting centroids $\mathbf{z}_j$, with $j = 1, ..., n$, can then be used to estimate the density of the data, such that

$$\hat{p}(\mathbf{x}) \approx \hat{p}(\mathbf{z}|\sigma) = \frac{1}{n(2\pi\sigma^2)^{D/2}} \sum_{i=1}^{n} \exp\left(-\frac{||\mathbf{z} - \mathbf{z}_i||^2}{2\sigma^2}\right) \tag{9}$$

In this experiment, we compute sparse kernel density estimates using isotropic kernels, in which the kernel width is set to maximise the leave-one-out likelihood using the subsets' centroids. The number of subsets n (also called number of prototype data points or vectors), chosen from $n = [N \cdot \{1, 5, 10, 15, 20\}/100]$, was also optimised according to the maximum likelihood of the training data using a grid-search strategy.

Weighted-sparse KDE The weighted-sparse kernel density estimates method is identical to the previous method. Nevertheless, it fits a penalised Gaussian kernel to each subset's centroid and combines them using the relative cluster (subset) size as a weight, such that the density of the data is given by:

$$\hat{p}(\mathbf{x}) \approx \hat{p}(\mathbf{z}|\sigma) = \frac{1}{(2\pi\sigma^2)^{D/2}} \sum_{i=1}^{n} w_i \exp\left(-\frac{||\mathbf{z} - \mathbf{z_i}||^\mathbf{2}}{2\sigma^2}\right) \tag{10}$$

where $w_i = k_i/N$ corresponds to the weight of each centroid, and k_i corresponds to the relative size of each cluster (i.e., k_i is the number of data points assigned to cluster i, with $i = 1, ..., n$). The number of subsets was optimised as described in the previous method.

Mixture KDE This method uses a similar weighting approach as the previous method, but all data points of the dataset (N) are used to compute the final

density of the data (Nickisch and Rasmussen [2010]). For each subset (obtained using K-means), the kernel's bandwidth is found using the data points contained in that susbset (i.e., $\boldsymbol{\sigma} = [\sigma_1, ...\sigma_n]$, with n being the number of subsets). The final density can then be computed as:

$$\hat{p}(\mathbf{x}|\boldsymbol{\sigma}) = \sum_{j=1}^{n} \frac{k_j}{N(2\pi\sigma_j^2)^{D/2}} \sum_{i=1}^{k_j} \exp\left(-\frac{||\mathbf{x} - \mathbf{x}_{ji}||^2}{2\sigma_j^2}\right) \tag{11}$$

where $\mathbf{x}_{ji}$ corresponds to the data point i assigned to cluster j, k_j corresponds to the relative size of each cluster, and $j = 1, ..., n$. As in the two previous methods, the kernel bandwidth was set to maximise the leave-one-out likelihood.

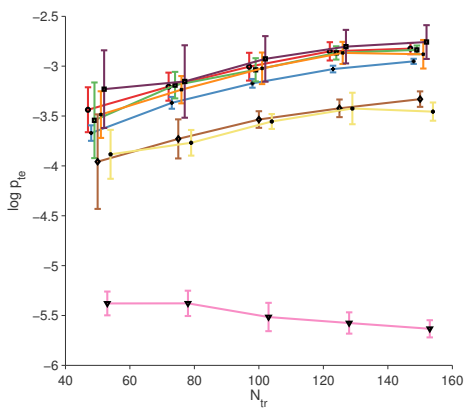
Experimental setting

The experiments were conducted after pre-processing the raw data. Instances with missing variables were removed from the dataset. Then, each variable (or attribute) in each dataset, x, was normalized so that $x_{normalised} = \frac{x-\mu}{SD}$, where μ and SD denote the mean and the standard deviation, respectively, computed from the training data points for the given variable.

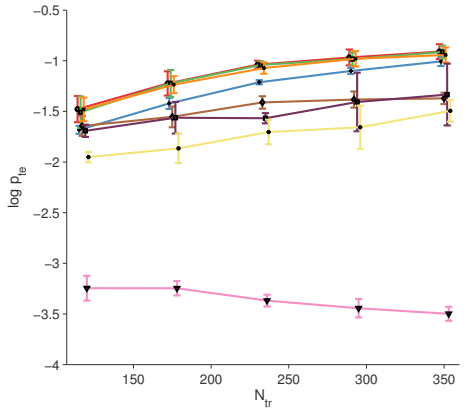
We ran the algorithms using the six different datasets. In each dataset, we split the data in training and test sets. Training data were used to determine the kernel size and obtain the density of the data. Test data were used to evaluate the performance of each density estimator. The training set size was varied to a maximum of $N_{tr} = N/2$. For each partition size considered, 10 random splits of the data were performed, and the averaged test log-likelihood (over the 10 folds) was calculated.

Results and Conclusion

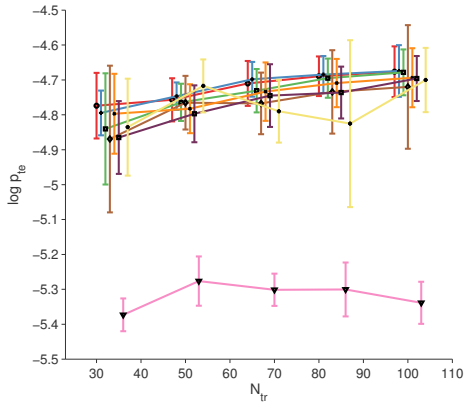
The results of the different density estimators described above can be found in Figure 3. They clearly show five things: (i) more data, in general, yields better perfor-



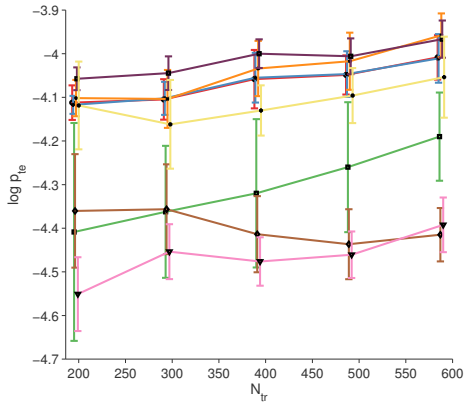
(a) *Sine-wave* data, $(N, D) = (300, 2)$



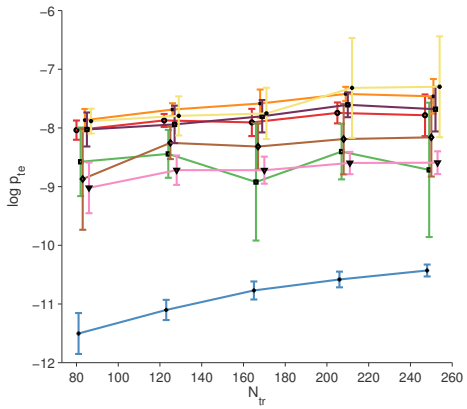
(b) *Spiral* data, $(N, D) = (700, 2)$



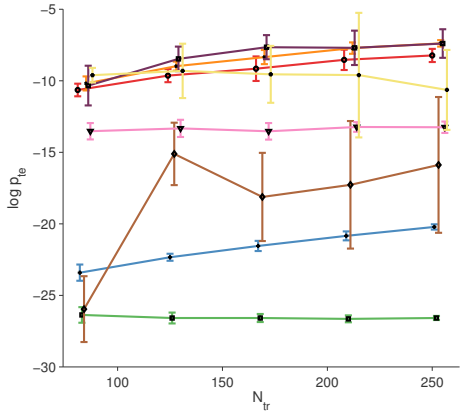
(c) *Moon* data, $(N, D) = (200, 2)$



(d) *Birth-weight* data, $(N, D) = (1175, 3)$



(e) *Diabetes* data, $(N, D) = (500, 10)$



(f) *Housing* data, $(N, D) = (506, 10)$

Figure 3: Each panel shows the log test density ($\log p_{te}$) averaged over 10 random splits for the different methods evaluated as a function of the number of training data points (N_{tr}) in the described dataset (with D dimensions). For each dataset of size N, $N_{tr} = N \cdot \{2, 3, 4, 5, 6\}/12$.

mance; (ii) the maximum leave-one-out likelihood approach is clearly and consistently among the best performing estimators across the different datasets; (iii) the manifold Parzen windows method offer only little benefit; (iv) the performance of the average nearest-neighbour approach is poor for datasets of higher dimension; and (v) in order to perform sparse kernel density estimation, a weighting-scheme should be used in order to penalise the contribution of each cluster (according to their size), as its performance is substantially superior to that of the simple sparse kernel density estimates across all datasets.

Other methods that have recently been proposed (such as those explored in Bengio et al. [2006]; Hong et al. [2008]; Nickisch and Rasmussen [2010]), were not considered in this analysis. Although some of these methods have attractive properties, a few problems remain such as initialisation and parametrisation of the learning algorithms, the computational demand, they do not generalise well, and may only convey little benefit (Nickisch and Rasmussen [2010]). Among all methods tested in this experiment, the maximum leave-one-out likelihood approach appears to provide a simple and reliable way of determining the kernel size, and thus, estimating the density of the data.

Appendix C - Comparing models of different dimensionality: a numerical approach

A consequence of having different values of $p(\mathbf{x})$ arising from the use of models with different dimensions is that the value of a novelty score $z(\mathbf{x})$ (with $z(\mathbf{x}) = -\log p(\mathbf{x})$) for one model will not be the same as the novelty score for a model with a higher or lower number of dimensions. Hann [2008] addressed this problem by proposing a numerical approach which uses probability to relate the probability density or novelty values calculated using a kernel density estimates model with a lower or higher dimensionality. Here, we briefly recover the method proposed, and present the analysis conducted with the models built in chapter 6.

Determining Probability

An intuitive method to relate values of probability density generated by models of different dimensionality is to use probability P, where $0 \leq P \leq 1$. The probability that x lies within the interval (a, b) is given by:

$$P(x \in (a, b)) = \int_a^b p(x)dx \tag{12}$$

If we consider that p($\mathbf{x}$) denotes a model of physiological normality, the peak or maximum of the probability density κ, with

$$\kappa = \max_{\mathbf{x}} p(\mathbf{x}), \tag{13}$$

can be interpreted as the point of "maximum normality", or conversely, the point of miminum novelty $z(\mathbf{x})$. The "boundary of normality" is defined as the surface of all points for which the probability density is equal to a threshold T:

$$\tau = \{\mathbf{x} : p(\mathbf{x}) = T\} \tag{14}$$

The probability of a vector lying within the normal region bounded by the threshold on probability density can then be expressed as:

$$P(\{\mathbf{x} : p(\mathbf{x}) \leq T\}) = \int_{\kappa}^{\tau} p(\mathbf{x}) d\mathbf{x} \tag{15}$$

If we consider the kernel density estimates (or Parzen windows model) probability density, this equation is not analytically trackable; i.e., the probability cannot be solved (computed) in closed-form. Therefore, a numerical approach for solving the problem of estimating the probability, as defined in Eq. (15), must be applied.

Numerical estimation of probability

If we assume that N_s samples, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_s}$, are drawn from $p(\mathbf{x})$, the probability density at every sample may be calculated using the Parzen windows model, giving $p(\mathbf{x}_1), p(\mathbf{x}_2), ..., p(\mathbf{x}_{N_s})$. The probability P of a vector $\mathbf{x}$ lying within the region(s) bounded by τ may then be estimated as the fraction of all samples that

have a probability density less than or equal to the threshold T.

To generate a random sample from the kernel density estimates-based distribution, a kernel is randomly selected (as every kernel has equal probability), and then a sample is drawn from the D-dimensional Gaussian distribution centred on that kernel. The samples are drawn from an unbounded (not-constrained) multivariate distribution. However, the artefact rejection applied to the inputs of the model of normality removes any values outside the ranges specified (see section 3.3.1 on page 70). The generated samples are therefore "pruned" by removing all patterns containing values outside the normalised artefact rejection limits, which ensures that the probability estimate will have an integral close to 1 over the volume of possible normalised parameter values. After the pruning procedure, the probability density at every sample is calculated.

The aim of the overall method is to produce a mapping between probability density, or novelty, and probability. This is achieved by finding the fraction of samples that have a probability density higher than a series of successively lower thresholds; or equivalently, a novelty score lower than a series of increasing thresholds ($z(\mathbf{x}) \leq \tau$, with varying τ), as shown in Figure 4. Hann [2008] noted that it is simplest to test a series of novelty thresholds, rather than probability density thresholds, as the logarithmic scaling of novelty will automatically provide higher resolution in the regions of greatest interest, namely the regions of low probability density.

Experimental setting

The experiments were conducted considering the two models built in chapter 6 that were used to study the trajectories of vital signs:

5-D model This model was constructed using data from the last day on the ward (pre-discharge) of patients in the normal group, which comprise the five vital signs (HR, RR, SpO_2, temperature, and systolic BP);

9-D model As the previous model, this model contains pre-discharge data of patients in the normal group, and the data comprise not only the same five vital

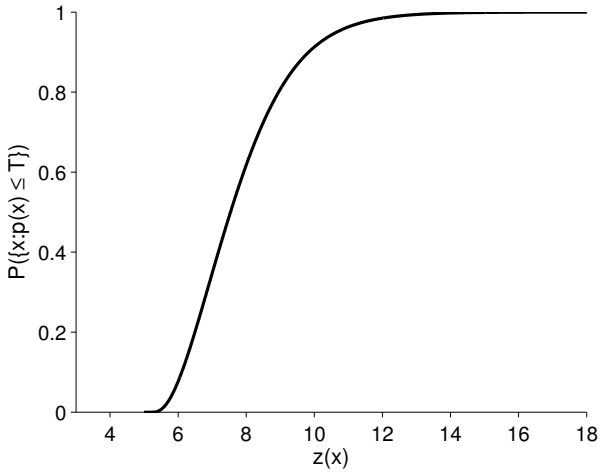


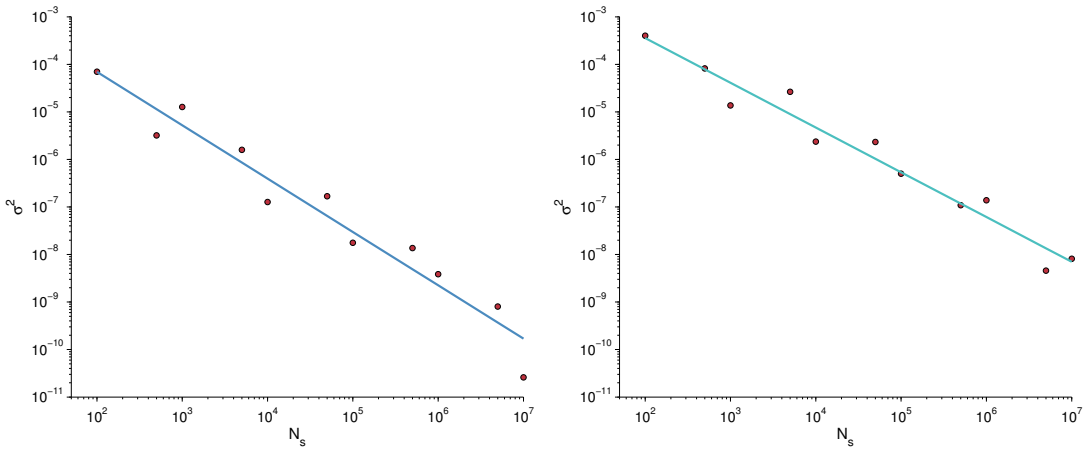
Figure 4: Estimation of the probability of a sample falling in the region bounded by a threshold on $P(\mathbf{x})$ (using the 5-D model): relationship between $z(\mathbf{x}) \leq T$ and $P(\{\mathbf{x} : \hat{p}(\mathbf{x})\}$ using 10^7 samples. The horizontal axis is scaled to show the section of the curve that is of primary interest.

signs, but also the variability indices of HR, RR, temperature, and systolic BP.

In the specific case of our work, we are interested in comparing the novelty scores computed using the 5-D model with those using the 9-D model. As the method described above relies on random sampling, each time it is carried out a different result will be obtained. It is therefore necessary to choose a value of N_s (number of samples) that provides solutions that can be reproduced with sufficient accuracy, and that allow the comparison of both models without a significant variance between the solutions achieved with this sampling strategy.

On each model, we ran the sampling process five times for each value of N_s, for values of N_s in the range from 10^2 to 10^7. The "pruning" procedure was applied to ensure that the probability estimate of the generated samples has an integral close to 1 over the volume of possible normalised parameter values, after artefact rejection[1]. The variance σ^2 of the estimates of the probability (over the

[1]The (non-normalised) rejection limits, or the range of possible values, for the variability index of a given variable are drawn from the rejection limits for that variable; e.g., the lower and upper cutoff values for heart rate are 30 and 300 beats per min, respectively; hence, the lower and upper cutoff values for the corresponding variability index are 0 and 270 beats per min.



(a) (b)

Figure 5: **(a)** Variance of the estimate of $P(\{\mathbf{x} : \hat{p}(\mathbf{x}) \leq \exp(-\tau_{5D})\}$ as the number of samples drawn from the 5-D model is increased from 100 to 10^7. **(b)** Variance of the estimate of $P(\{\mathbf{x} : \hat{p}(\mathbf{x}) \leq \exp(-\tau_{9D})\}$ as the number of samples drawn from the 9-D model is increased from 100 to 10^7. In both cases, both axes use logarithmic scales, and the line fitted has a slope of $(10N_s)^{-1}$.

five sampling runs) associated with a novelty less than or equal to τ_{5D} and τ_{9D} was then computed for the 5-D and 9-D models, respectively. The value of τ_m, with $m \in \{5D, 9D\}$, was computed by evaluating the probability density of all data points from the entire trajectory of the patients selected for the training the model m; i.e., all the data from the patients in the "normal" group in the training set, from which the pre-discharge data were used for training the model m, were used to compute the probability density. The value of τ_m was then set to the 98^{th} percentile of the probability density evaluated at those data points for model m.

Results

The results are shown in Figure 5. Both axes of both plots use logarithmic scales. As expected, we observe that as N_s increases (as more samples are generated from the model), the variance over the five sampling runs decreases exponentially (or linearly in the logarithm scale), in both models. More importantly, we ob-

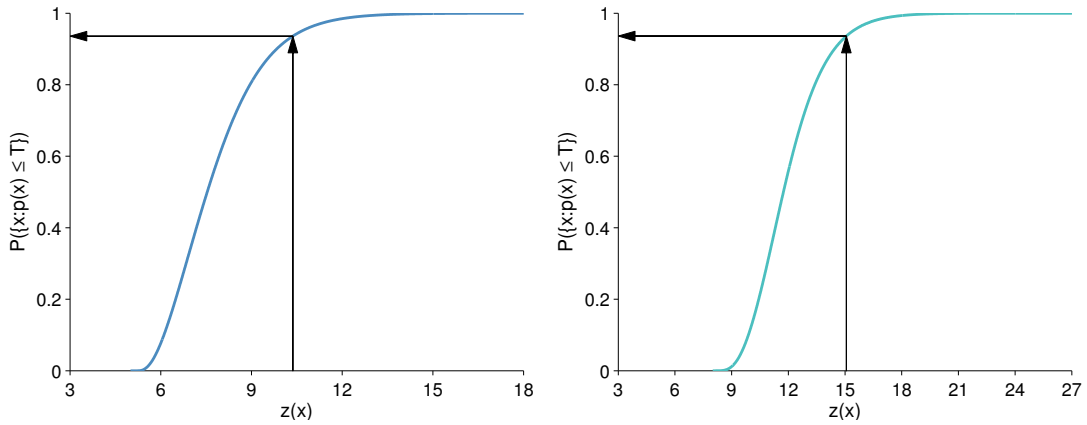


Figure 6: Curves for mapping from novelty scores to probability for **(a)** the 5-D model, and **(b)** the 9-D model. The black arrows indicate how a novelty score computed using each model can be converted to a probability P; e.g., a score of approximately 15 using the 9-D model corresponds to a score of around 10.5 using the 5-D model.

serve that for the 9-D model, a variance of approximately 10^{-8} is obtained using $N_s = 10^7$, which corresponds to a difference between the maximum and minimum estimate for the probability corresponding to a novelty of $\tau_{9D} = 18$, of 0.009% of the mean of the five probability estimates. For the same N_s, the difference between the maximum and minimum estimate for the probability corresponding to a novelty of $\tau_{5D} = 13$, was found to be 0.001% of the mean of the five probability estimates. This level of accuracy is acceptable for comparing both models, and it is therefore suggested that $N_s = 10^7$ is used when sampling both 5-D and 9-D models.

The curve in Figure 4 (reproduced in Figure 6a) corresponding to sampling the 5-D model, may then be compared to the curve corresponding to sampling the 9-D model (Figure 6b). That is, using probability P, the novelty scores generated from the two models may then be compared using the same scale.

Appendix D - Estimating the number of clusters

A fundamental problem in clustering analysis is to determine the number of clusters, which is usually taken as a prior in most clustering algorithms (but not in hierarchical clustering methods). Clustering solutions may vary as different numbers of clusters are specified. A number of strategies for estimating the optimal number of clusters have been proposed. A very extensive comparative evaluation was conducted by Milligan and Cooper [1985], where the authors compared 30 proposed methods for estimating the number of clusters when applying hierarchical clustering algorithms to simulated data with well-separated clusters. Among the methods examined in their work, the method proposed by Caliński [1974] outperformed the others. This approach determines the number of clusters g by maximising the Calinski and Harabasz's index, $CH(g)$ over g, given by

$$CH(g) = \frac{B(g)/(g-1)}{W(g)/(n-g)}, \tag{16}$$

and $B(g)$ and $W(g)$ are the between- and within-cluster sum of squared errors, calculated as the trace of matrices B (Eq. 17) and W (Eq. 20), respectively, which are defined by

$$B = \sum_{m=1}^{g} n_m(\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})', \text{and} \tag{17}$$

$$W = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_{ml} - \bar{x}_m)(x_{ml} - \bar{x}_m)', \tag{18}$$

where $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. Other methods proposed in the literature more recently include the *silhouette* index (described in Kaufman and Rousseeuw [2009]), and the *gap* method (Tibshirani et al. [2001]).

The *silhouette* statistic for a given object i is defined by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{19}$$

where $a(i)$ is the average distance to other points in its cluster, and $b(i)$ is the average distance to points in the nearest cluster to its own cluster. The *silhouette* index, denoted by $\bar{s}(g)$, is defined as the average of the $s(i)$ for all objects in the data. The optimum value of g is chosen such that $\bar{s}(g)$ is maximised over all values of g.

The idea behind the gap statistic is to compare the change in $W(g)$ as g increases for the original data with that expected for the data generated from a suitable reference null distribution; i.e., a distribution with no obvious clustering (Tibshirani et al. [2001]). The estimate for the optimal number of clusters g is the value for which $\log W(g)$ falls the farthest below its expected curve. This information is contained in the following expression for the gap statistic:

$$Gap_n(g) = E_n^*\{\log W(g)\} - \log W(g), \tag{20}$$

where $E_n^*\{\log W(g)\}$ indicates the expected value of $\log W(g)$ under the null distribution. Tibshirani et al. [2001] showed that for univariate cases, the uniform distribution should be used as the reference null distribution. In detail, the computation steps of the gap method are:

- Cluster the data under investigation for fixed number of clusters, g, where $g = 1, 2, ...,$ and compute $W(g)$ for all values of g;
- Generate C reference data points in the way described above; cluster each of the C reference data points and calculate $W_c^*(g), c = 1, 2, ..., C$ for each value of g; and compute the gap statistic $Gap(g) = (1/C)\sum_c \log W_c^*(g) - \log W(g)$;
- With $\bar{w} = (1/C)\sum_c \log W_c^*(g)$, compute the variance (given by $var(g) = C^{-1}\sum_c [\log W_c^*(g) - \bar{w}]^2$), and define $s(g) = \sqrt{1 + 1/C}\sqrt{var(g)}$;
- Choose the number of clusters as the smallest g such that $Gap(g) \geq Gap(g+1) - s(g+1)$.

With a large number of model examples, the effectiveness of the gap statistic proposed by Tibshirani et al. [2001] was the most convincingly demonstrated.