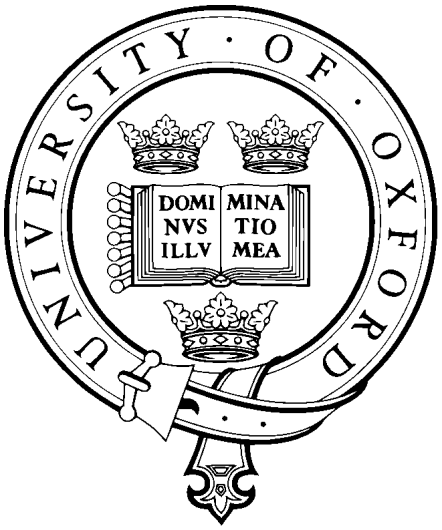
Vital-Sign Data-Fusion Methods to Identify Patient Deterioration in the Emergency Department



Mauro D. Santos

St Hilda's College
University of Oxford

Supervised by
Prof. Lionel Tarassenko
Prof. David Clifton

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*

Trinity Term, 2018

This thesis is dedicated to,

my wife Catarina, our daugther Beatriz,

my father Ilídio, my mother Carolina,

my brother Wilson, and my sister Dejanira.

Thank you for your love and support.

What Is Your Life's Blueprint?

"I want to ask you a question, and that is: What is your life's blueprint?

Whenever a building is constructed, you usually have an architect who draws a blueprint, and that blueprint serves as the pattern, as the guide, and a building is not well erected without a good, solid blueprint. Now, each of you is in the process of building the structure of your lives, and the question is whether you have a proper, a solid and a sound blueprint. I want to suggest some of the things that should begin your life's blueprint.

Number one in your life's blueprint, should be a deep belief in your own dignity, your worth and your own 'somebodiness'. Don't allow anybody to make you feel that you're nobody. Always feel that you count. Always feel that you have worth, and always feel that your life has ultimate significance.

Secondly, in your life's blueprint you must have as the basic principle the determination to achieve excellence in your various fields of endeavour. You're going to be deciding as the days, as the years unfold what you will do in life - what your life's work will be. Set out to do it well. [...] and finally, in your life's blueprint, must be a commitment to the eternal principles of beauty, love and justice."

- Dr. Martin Luther King, Jr., October 26, 1967

Acknowledgements

This thesis would not be possible without the help of my family, friends and colleagues. I would like to thank: my family and friends, whose constant presence helped finish this stage of my life, and continue my efforts to learn every day; my supervisors, Professor Lionel Tarassenko and Professor David Clifton, for guiding me through this DPhil, I was lucky to have such exemplary supervisors; the clinical research team, namely, Dr. Richard Pullinger, Rob Way, Dr. Sarah Wilson, Soubera Rymel, Karen Warnes and Sally Beer, for their work and guidance in the Emergency Department large-scale study that led to this thesis; the colleagues from The Learning Clinic and OBS Medical for their technical support during the data collection process, in which their technology was used; my colleagues from the BSP & mHealth research group, the CHI lab research group and the Centre of Doctoral Training in Healthcare Innovation (CDTHI) graduate program, from the University of Oxford, namely, Marco Pimentel, David Wong, Glen Colopy, Julien Oister, Alistair Jonhson and Tingting Zhu for their collaboration and discussions regarding my thesis work; Carlos Arteta, João Domingos, Elnaz Gederi, Maxim Osipov, Ali Maraci, David Springer, Carmelo Velardo, Dario Salvi, Ahmar Shah and Mayella Zamora, for the great discussions and their collaboration on various projects in the lab; Professor Alison Noble and Professor Gari Clifford, who directed the CDTHI, also providing guidance during my graduate studies; Prof. Maarten De Vos (internal) and Dr. Christian Subbe (external) for their useful contributions and comments in the Viva examination; the IBME (Old Road Campus, Oxford) administrative and IT staff that maintained the office and servers running during this period; and finally, I would like to thank the Research Councils UK Digital Economy Programme, for providing partial funding for my research work in the CDTHI.

Vital-sign Data-Fusion Methods to Identify Patient Deterioration in the Emergency Department

Mauro D. Santos
Thesis submitted for the degree of Doctor of Philosophy
St Hildas's College
Trinity 2018

Abstract

In the United Kingdom Emergency Departments (ED), clinical staff requires to diagnose, give treatment and discharge patients, within 4 hours of their arrival. The patients' vital signs are traditionally managed using paper Track and Trigger (T&T) charts, prone to human error, and bedside monitors, whose alerts are often ignored. Consequently, patient deterioration might be missed at and between nurses' observation sets. This thesis has analysed data from a three stage study in the ED of the John Radcliffe Hospital, Oxford, to investigate the use of an electronic T&T system (VitalPac) and a data-fusion system (Visensia) to help staff identify physiological deterioration in patients attending the majors area.

Data was collected from a total of 10,488 ED attendances receiving standard care in stage 1, followed by two technology interventions in stages 2 and 3, respectively, for a total period of 6 months. It was shown that 9% of the observations sets, conducted on stable ED patients in stage 1, were done on unstable patients when staff was guided by VitalPac in stage 2. One of the causes might have been the increase in the Early Warning Score (EWS) completion from 52% to 100% of the observation sets. In stage 3, 35.7% of the Visensia alerts generated on continuous bedside monitor data, were deemed "technical" alerts due to data artefacts. On the other hand, clinical staff responded within 15 minutes to 85% of the "physiological" alerts.

A two-stage Machine Learning (ML) architecture was proposed to fuse intermittent and continuous vital-sign data and use a sub-population novelty detection model to identify multivariate data deviating from normal physiology. This ML approach out-performed the baseline Visensia model (a population-based model, applied over the continuous data), and the National EWS system (applied over the observation sets data) in detecting patients escalated to the Resuscitation area during their ED stay, on a test set of 1,070 ED attendances (AUROCs and 95% confidence intervals were 0.737 (0.623, 0.830), 0.657 (0.521, 0.755), and 0.643 (0.522, 0.749), respectively).

External contributions of code/software

My colleague, Dr. Marco Pimentel, implemented the original code for the multi-instance Gaussian process (MIGP) algorithm. Dr. David Wong helped with developing a semi-automatic Matlab script to match the continuous vital-sign data to the study participants. Prof. David Clifton provided the Microsoft Access template for the double- and single-entry transcription of the paper-based T&T charts data into electronic means. The various software packages that were used in this thesis, are cited in the correspondent section in each chapter. The remaining materials described in this thesis are a result of my own work.

List of Publications

The following lists all publications produced over the course of this DPhil.

Thesis Publications

Two publications are directly related with the work in this thesis: Pullinger et al. (2015)[2] and Santos et al. (2013)[3], which are extended in chapters 4 and 5, respectively. Chapter 8 is further extended in [1].

[1] Santos, M. D., Pimentel, M.A.F., Clifton, D. A., and Tarassenko, L. *Multi-instance Gaussian Processes for time-series modelling of the vital signs for Emergency Department patients.* ***[In review]***

[2] Pullinger, R., Wilson, S., Way, R., Santos, M., Wong, D., Clifton, D., Birks, J., and Tarassenko, L. *Implementing an electronic observation and early warning score chart in the emergency department: A feasibility study.* European Journal of Emergency Medicine, 2015.

[3] Santos, M. D., Clifton, D. A., and Tarassenko, L. *Performance of early warning scoring systems to detect patient deterioration in the emergency department.* In International Symposium on Foundations of Health Informatics Engineering and Systems, pages 159 - 169. Springer, 2013.

Related Non-Thesis Publications

[4] Pimentel, M.A.F., Santos, M. D., Springer, D. B., Clifford, D. B., *Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices*, Physiological measurement 36 (8), 1717, 2015.

[5] Salciccioli, J. D., Marshall, D. C., Pimentel, M. A. F., Santos, M. D., Pollard, T., Celi, L. A., Shalhoub, J., *The association between the neutrophil-to-lymphocyte ratio and mortality in critical illness: an observational cohort study*, Critical Care,19:13, 2015.

Non-Related Publications

[6] Cairns, A. E., Tucker, K. L., Leeson P., Mackillop, L. H., Santos, M.D., Velardo, C., Salvi D., Mort, S., Mollison, J., Tarassenko L., McManus R. J., *Self-Management of Postnatal Hypertension: The SNAP-HT Trial*, Hypertension, 2018.

[7] Pimentel, M.A.F., Santos M.D., Arteta C., Domingos J., Maraci M.A., Clifford G.D., *Respiratory rate estimation from the oscillometric waveform obtained from a non-invasive cuff-based blood pressure device*, IEEE EMBC, Chicago, USA, 2014.

[8] Arteta, C., Domingos, J., Pimentel, M. A. F., Santos, M. D., Chiffot, C., Springer, D., Raghu, A., and Clifford, G. D., *Low-cost Blood Pressure Monitor Device for Developing Countries*, Mobihealth, Kos Island, Greece, 2011.

Contents

List of Figures

List of Tables

Notation

x (italics lower case) is a scalar; $\mathbf{x}$ (bold lower case) is a column vector; $\mathbf{X}$ (bold upper case) is a matrix; X (italics upper case) is a random variable set; a superscript T denotes the transpose of a matrix or vector, and so $\mathbf{x}^T$ is a row vector; entries in matrices and vectors are denoted by subscripts, so the ijth element of $\mathbf{X}$ is written X_{ij}; to explicitly defined its ith row or column we use $\mathbf{X}_{i,:}$ and $\mathbf{X}_{:,i}$, respectively; and finally, the length of a vector is denoted by $\|\mathbf{x}\|$.

Acronyms

List of common clinical abbreviations	
BP	Blood Pressure
CEWS	Centile-based Early Warning Score
CPR	Cardiopulmunary Resuscitation
DBP	Diastolic Blood Pressure
ED	Emergency Department
EPR	Electronic Patient Record
e-T&T	Electronic Track-and-Trigger
FiO_2	Fraction of inspired Oxygen
GCS	Glasgow Coma Scale
HR	Heart Rate
ICU	Intensive Care Unit
JR	John Radcliffe Hospital, Oxford
LOS	Length-Of-Stay
MEWS	Modified Early Warning Score
NEWS	National Early Warning Score
RR	Respiratory Rate
SBP	Systolic Blood Pressure
SDA	Systolic and Diastolic Average
SpO_2	Peripheral Oxygen Saturation
TEMP	Temperature

List of common technical abbreviations	
ARD	Automatic Relevance Determination
AUROC	Area Under the Receiver-Operating Curve
AUC	Area Under the Curve
BCa	Bias-corrected and accelerated
BO	Bayesian Optimisation
CDF	Cumulative Distribution Function
CI	Confidence Interval
EI	Expected Improvement
EM	Expectation-Maximisation
FPR	False Positive Rate
GP	Gaussian Process
IQR	Inter Quartile Range
KDE	Kernel Density Estimate
LR	Logistic Regression
MAE	Mean Absolute Error
MIGP	Multi-instance Gaussian Process
ML	Machine Learning
NLML	Negative Log Marginal Likelihood
OSVM	One-class Support Vector Machines
PDF	Probability Density Function
PPV	Positive Predictive Value
PSI	Patient Status Index
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
SE	Squared Exponential
SEN	Sensitivity
SPEC	Specificity
SMIGP	Sequential Multi-Instance Gaussian Process
s.t.	subject to
SVM	Support Vector Machines
TPR	True Positive Rate
ZOH	Zero-order hold

Chapter 1

Introduction

Following the declining costs of digital sensing and actuation, large quantities of data became available to create Machine Learning (ML) models able to automate real-time critical clinical tasks such as in-hospital patient condition monitoring. E.g. recent research as analysed the application of ML models to monitor the patient condition in step-down unit wards (Tarassenko et al., 2006), in post-operative cancer patient recovery wards (Pimentel, 2015), and in the intensive care units (Johnson et al., 2014). Computers lack human biases, and their vigilance need not be interrupted by rest breaks or lapses of concentration, which brings great advantages to healthcare services (Frey and Osborne, 2017).

However, building robust clinical decision support tools for patient condition monitoring entails a comprehensive set of scientific and engineering steps, such as: (i) the formulation of the correct clinical decision support problem, i.e. understanding the clinical context, and acquiring the corresponding data examples; (ii) modelling the normal and abnormal patterns of the complex human physiology; (iii) coping with noisy sensor data; (iv) working with ordered data from multiple sources, with different biases, precisions and sampling frequencies; and finally (v) finding the correct metrics to optimise the alerts generated by the decision support tools so that they are useful for the real-time (or offline) tasks in the hospital setting.

1.1 Thesis objectives and overview

In this thesis, we are concerned with the application of Machine Learning to the problem of patient condition monitoring in the Emergency Department (ED). The following structure is followed to discuss our main technical contributions:

- chapter 2 reviews the state of the art of patient condition monitoring in the ED of the John Radcliffe Hospital, Oxford (JR);
- chapter 3 describes the data collection process and the main outcomes of a large-scale ED study, which evaluated the use of an electronic Track and Trigger system (e-T&T) system (VitalPAC) and a data-fusion system (Visensia), prospectively, to help identify patient physiological deterioration in the majors area of the ED of the JR (contribution (i));
- chapter 4 analyses the performance of clinical staff in following the ED frequency of observations protocol, when guided by an e-T&T system (contribution (ii));
- chapter 5 investigates the use of the patient demographics (age) and clinical context information (use of O_2 support) to optimised EWS systems for ED patients. It also investigates the use of a non-quantised EWS system design (contribution (iii));
- chapter 6 reviews recent work on ML methods that use data-fusion for in-hospital patient condition monitoring;
- chapter 7 evaluates the clinical staff response to Visensia, a system able to detect patient physiological deterioration, using continuous vital-sign data from bedside monitors, in the majors areas of the ED of the JR (contribution (iv));
- chapter 8 discusses the use of Gaussian processes to fuse intermittent and continuous ordered vital-sign data (i.e. multi-instance data), and cope with artefacts present in these data (contribution (v));

- chapter 9 evaluates novelty detection (one-class classification) and Logistic Regression (two-class classification) approaches, trained from the multi-instance vital-sign time-series model, and which included patient demographics (age and sex) and clinical context information (e.g. use of O_2 support) to better identify physiological deterioration in the ED setting (contribution (vi));
- finally, chapter 10 provides a general conclusion and directions for future work.

Chapter 2

Vital-sign monitoring in the ED

The ED[1] is often the first contact with hospital care services for patients experiencing an acute health problem. It is a medical facility specialising in the acute care of patients who present without prior appointment, either by their own means or by ambulance (NHS, 2013). It is a busy setting, combining a large number of admissions with a wide variety of high-acuity patient conditions. 16.2 million ED attendances were reported between April, 2010, and March, 2011, from 187 hospitals in England. 26% of all ED admissions arrived by ambulance or helicopter, with these typically being the patients with the most critical conditions. 49% of these latter patients were subsequently admitted to the hospital after being discharged from the ED, while only 12% of patients who arrived by another method of transportation were admitted (Health and Social Care Information Centre, Hospital Episode Statistics, 26 January 2012).

These unscheduled admissions can cause overcrowding (Higginson, 2012) and, although in the UK 98% of the ED patients are required to be discharged within 4 hours (National Institute of Clinical Excellence, 2007), it is reported that the average length-of-stay (LOS) is 4 hours for 95% of the patients, with a median time-to-treatment of 55 minutes (Health and Social Care Information Centre, Hospital Episode Statistics, 26 January 2012).

[1]Also known as Accident and Emergency, the Emergency Room, or the Casualty Department.

Staffing the ED 24 hours each day, for all seven days of the week, is becoming a concern for the UK Department of Health (NHS, 2013), and recent studies report that an increase in crowding can cause delays in treatment, low patient satisfaction, and increased mortality rates (Bernstein et al., 2009; Johnson and Winkelman, 2011).

It is known that patient morbidity is improved if diagnosis leads to early treatment (Becker et al., 1977; Buist et al., 2002), and it has been shown that early treatment in the ED can decrease hospital length-of-stay (LOS) and the level of care required in the Intensive Care Unit (ICU) (Huang, 2004). Furthermore, decreased LOS, with a consequently increased patient throughput, may have a substantial financial benefit for the healthcare system.

Physiological measurements and subsequent interventions have been shown to predict the need for higher level of care among hospital patients (Goldhill and McNarry, 2004), and in this thesis we analyse such measurements and interventions, and how could they be improved to monitor patient condition during the ED stay.

2.1 Patient care in the ED

Figure 2.1 is a diagram of the change in the flow of patient care through the ED of the JR, between 2011 and 2013. Figure 2.2 shows an example of a patient cubicle and the medical equipment used to care for patients in the "majors" area of the ED. When patients arrive in the ED, their condition is first evaluated through a triage process. This typically takes less than 10 minutes, and involves a nurse performing a set of physiological measurements, which are subsequently scored using a triage algorithm. The latter assigns each patient to a risk category. If patients arrive by ambulance, depending on the severity of their condition, they may be immediately admitted to a specialised hospital ward. Otherwise, patients requiring preliminary investigation are admitted to one of the ED areas: the resuscitation ("resus"), "majors", or "minors" areas.

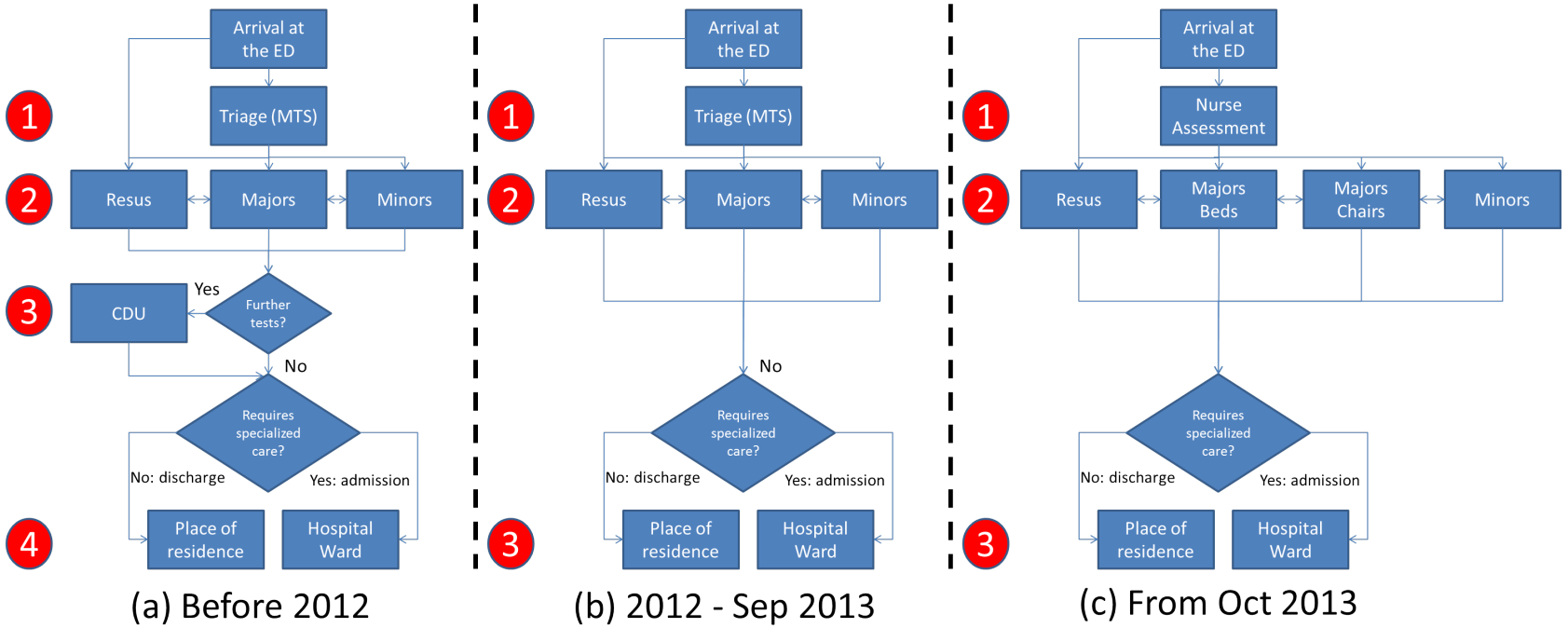


Figure 2.1: *Patient flow in the JR ED. (a) (1) The Manchester Triage System (MTS) algorithm is used to prioritise patients; (2) they are allocated to the "minors" when they are not presenting life-threatening symptoms, or to the "resus" or "majors" areas if immediate treatment is required; (3) for further observation, they are transferred to the Clinical Decision Unit (CDU). (4) patients are either discharged home (stable), or admitted to the next hospital ward (unstable). (b) From 2012 to September 2013. From 2012 the CDU operated independently from the ED. (c) From October 2013 the triage process was based on "Nurse Assessment", and patients that may be at risk of deteriorating, but do not require staying in the majors beds, are also assessed in majors chairs.*

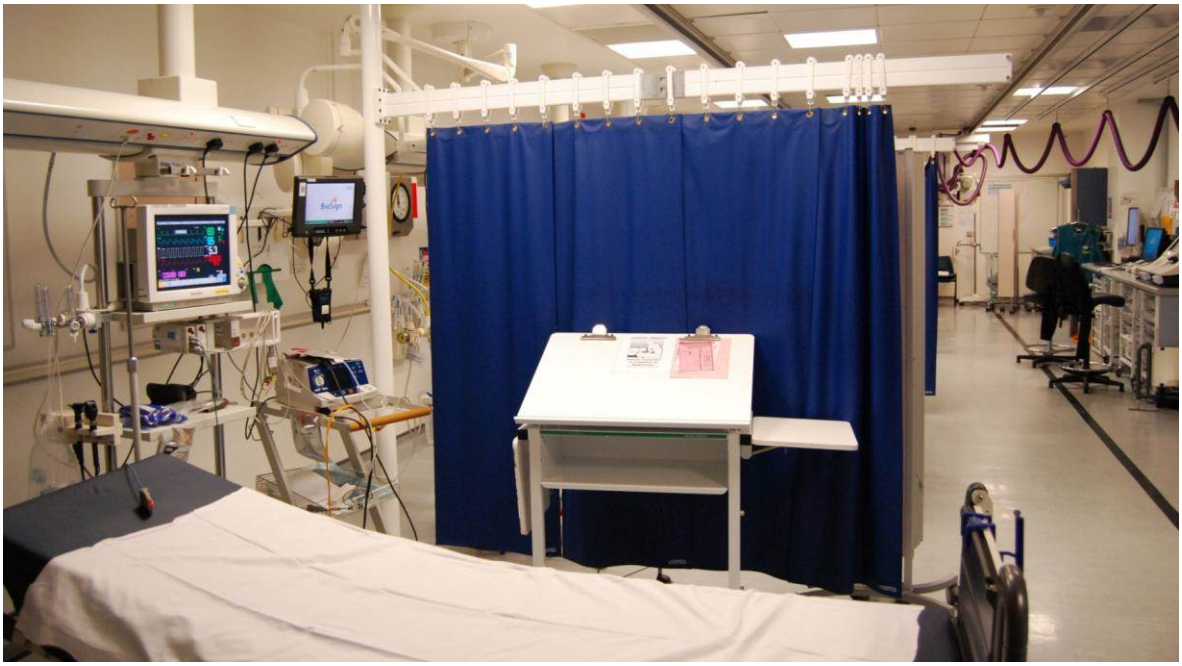


Figure 2.2: *Patient cubicle from the ED majors area of the JR. The Philips bedside monitor and the Visensia module - described in the next chapter - are the two monitors on the left side of this Figure, respectively.*

The resources and time allocated to patients in each of the three areas of the ED are proportional to the acuity of the patient condition, which is evaluated using T&T charts in the UK EDs. The patient vital signs are checked more frequently when the Early Warning Score (EWS), recorded in the T&T chart, is high (T&T and EWS are reviewed in sections 2.3, and 5.1). In the resus room, the nurse:patient ratio is 1:1, and vital signs are measured every 15 to 30 minutes. In the majors area, the ratio changes to 1:4, and vital signs are observed every 1 to 2 hours. In both areas patients are considered to be physiologically unstable and are therefore connected to bedside monitors, that collect their vital signs continuously.

The majors staff:patient ratio also applies to the minors area, with observations every 2 hours. In the minors area in the JR ED, the patients' vital signs are measured and recorded intermittently by nurses using mobile vital-sign monitors on stands (called "spot-check monitors", see Figure 2.4 for an example). In some hospitals, the Clinical Decision Unit (CDU) may be used to investigate patients that require longer than 4-hour stays (National Institute of Clinical Excellence, 2007), before being discharged home or

to the next hospital ward.

An important component of the effective delivery of medical care is matching the severity of illness to the appropriate level of care. Over-triage to critical care units results in unnecessary resource consumption, and under-triage to lower-acuity wards may result in worse patient outcomes. Therefore, it is important to ensure accurate vital-sign scoring during the patient ED stay.

2.2 Vital-sign measurements

Physiological homeostasis

As mentioned, the main purpose of the ED clinical staff is to diagnose and provide initial treatment, in order to restore "normality" to the ill patient. In general terms this "normality" is enabled by physiological homeostasis, which is the tendency of the body to maintain critical physiological parameters of its internal environment (e.g. blood glucose level, blood salinity, blood pressure, core body temperature) within specific ranges. As shown in Figure 2.3a, this homeostatic regulation is an on going process involving (i) the body receptors, which receive information of changes in the environment, ii) the body control centre, which processes the information and iii) the body effectors, which responds to the commands of the control centre by either opposing (negative feedback) or enhancing (positive feedback) the stimulus.

For example, in human thermoregulation, temperature (TEMP) receptors in the skin communicate information to the anterior *hypothalamic nucleus* and the adjacent preoptic area regions of the *hypothalamus*, in the brain, the control centre, and the effector is our blood vessels and sweat glands in our skin. In the baroreflex mechanism (Figure 2.3b), the baroreceptors (located at the aortic arch, carotid sinuses, auricles of the heart and *vena cavae*), respond to pressure induced stretching of the blood vessels, and send information to the nucleus of the *solitary track* in the *brainstream*. The latter controls the *sympathetic* and *parasympathetic* branches of the autonomic nervous system, which

have opposing effects on the blood pressure (BP). The effector is the heart and the blood vessels which increases or decreases its contractibility and their vasoconstriction, respectively, increasing (by activating of the sympathetic nervous system) or decreasing (by activating the parasympathetic nervous system) the blood pressure[2].

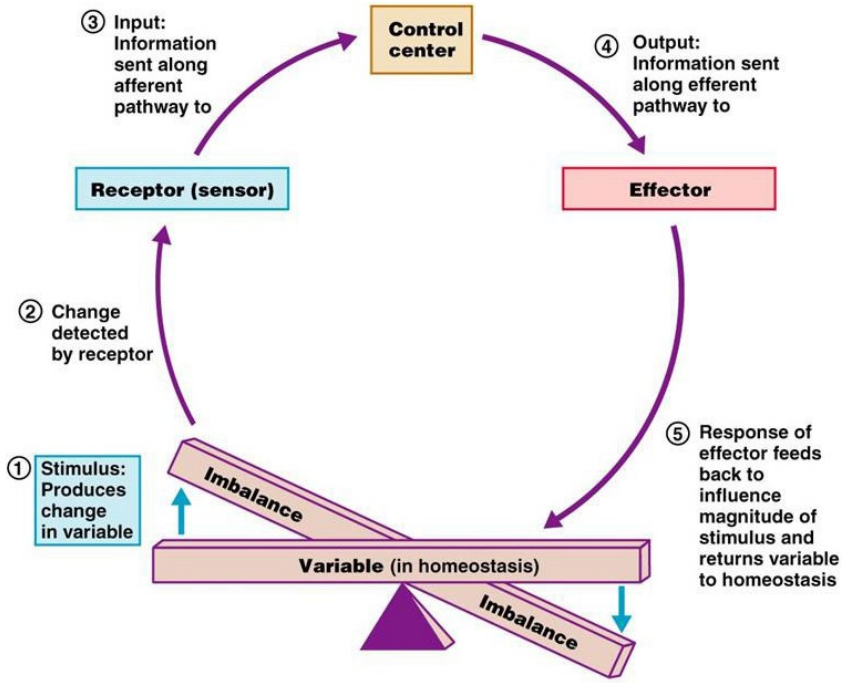
Thermoregulation and baroreflex are examples of a negative feedback control. Blood clotting (coagulation), which is activated to prevent bleeding when a blood vessel is injured, is an example of a positive mechanism.

For the purposes of this thesis it is important to know that in the initial contact with the patient clinical staff uses non-invasive and immediately available vital-sign measurements, such as Heart Rate (HR), Respiratory Rate (RR), Oxygen Saturation (SpO_2), Temperature, Blood Pressure and level of consciousness (usually the Glasgow Comma Scale, GCS), to have minimal information to diagnose conditions that impair the patients' physiological homeostasis (thermoregulation, baroreflex, respiration, etc...). These lead to or are caused by cardiac arrests, myocardial infarctions, mental illness, trauma, asthma and cardiopulmonary obstructive disease, which are the most treated conditions in the ED. Next we focus on the vital-sign measurements that allow the management of the patient condition in this ward.

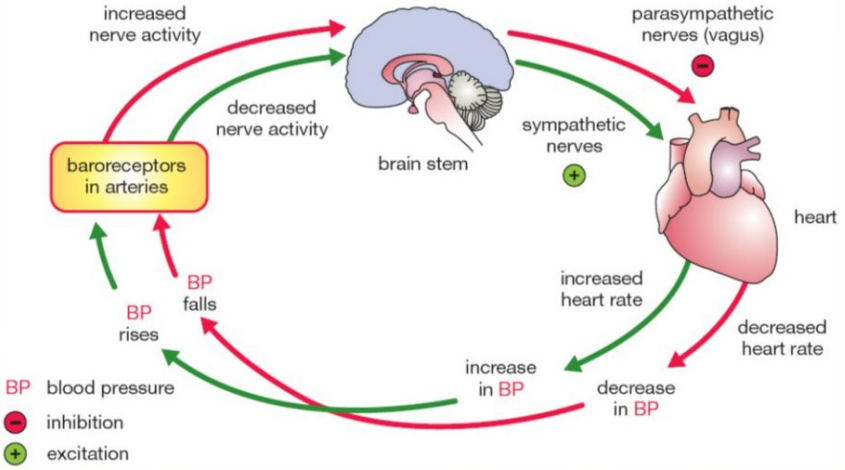
Heart Rate

In the ED, heart rate is recorded using a 3-lead Electrocardiogram (ECG), with 5 electrodes. HR is usually estimated by a bedside monitor using software that determines the number of R-peaks present in short-time windows (typically a few beats) of the ECG waveform, being reported in beats per minute (bpm). The R-peak is the most salient feature in the ECG waveform, consisting of a sharp spike, created by the depolarisation of the ventricles. HR is influenced by many factors including age, existing conditions (e.g. fever), medication (e.g. beta-blockers) and fluid status (Elliott and Coventry, 2012). Low

[2]These are illustrative examples of homeostasis, as the body contains two other, slower acting systems to regulate BP: the heart releases atrial natriuretic peptide when BP is high, and the kidneys sense and correct low BP with the renin-angiotensin system.



(a) *Illustration of the human physiological homeostasis mechanism, from Marieb and Hoehn (2006).*



(b) *Physiological homeostasis example: influence of the baroreflex mechanism to control the blood pressure in the human body. The illustration was accessed from Slideplayer (2018).*

Figure 2.3

and high heart rates are called bradycardia and tachycardia, respectively, and may precede physiological deterioration. Finally, HR data can also be acquired from ambulatory ECG and blood pressure devices present in the triage room (Figure 2.4).

Figure 2.4: *(a) Example bedside monitor screen with HR, RR, SpO_2 and BP values and ECG, respiration, and photoplethysmography waveforms. (b) "Dinamap" spot-check monitor used for intermittent vital-sign measurement in triage room and ED ward. (c) Example tympanic-based temperature probe.*

Respiratory Rate

Respiratory Rate is a measurement of the number of respirations per minute (rpm). In the ED, RR is measured either by direct observation, by counting the number of chest wall movements in 1 minute, or by observing the RR estimates from the bedside monitors. Impedance pneumography is a commonly-used technique to monitor a patients respiration rate, with a bedside monitor. It is implemented by either using two or four electrodes (usually a pair of ECG electrodes present in the monitors), that measure changes in the electrical impedance of the patients thorax caused by respiration. When high-frequency current is injected into the thorax, through the electrodes, the potential difference between them is related with the resistivity of the tissue. During inspiration there is an increase in (i) the gas volume of the chest in relation to the fluid volume and in (ii) the length of conductance paths due to thorax expansion. These cause the conductivity to decrease, and, consequently, the electrical impedance (resistivity of the tissue) to increase (Gupta,

2011).

Low and high respiratory rate are called bradyapnea and tachypnea, respectively. Although RR is one of the most sensitive indicators of critical illness, being an early indicator for cardiac arrest (Cretikos et al., 2007), it is not always recorded accurately by clinical staff, that use short 15 seconds window, instead of the recommended 30 seconds window (Mitchell and Van Leuvan, 2008), specially in busy wards. In the ED ward, staff usually takes RR values from the continuous monitors.

Oxygen Saturation

Oxygen Saturation is the concentration of oxygen in the blood (or oxygenated haemoglobin), expressed as a percentage. It is usually acquired via a pulse oximeter, with a probe attached to the finger. The probe contains two Light Emitting Diodes (LEDs), which emit light at 660 nm and 940 nm (Shelley, 2007). These wavelengths are absorbed differently by both oxygenated and deoxygenated blood. The transmitted light is then detected by a photo-diode. The relative intensities of the transmitted light at the two wavelengths are then used to determine SpO_2. In the ED the SpO_2 is measured from the pulse oximeters connected to bedside monitors or to the ambulatory monitors in the triage room.

Blood Pressure

Blood pressure refers to the pressure exerted by blood against the arterial wall, being influenced by cardiac output, peripheral vascular resistance, blood volume and viscosity, and elasticity of the arterial walls. Changes in blood pressure may reflect underlying pathophysiology or the body's attempt to maintain homeostasis (Elliott and Coventry, 2012). In the work described in this thesis, BP values were taken either from the bedside monitors or from ambulatory BP monitors in the triage room. Both acquire a pressure signal (in mmHg) by using a cuff, usually attached to the patient's arm, which is inflated until the brachial artery is occluded. During the deflation phase the oscillations caused

by the pressure of blood against the arterial wall are captured, and the "oscillometric method" uses an algorithm to determine the Systolic (SBP) and Diastolic (DBP) Blood Pressure (Ramsey III, 1991).

Level of Consciousness

In the JR ED, the Glasgow Coma Scale is the most commonly-used system for evaluating the level of consciousness of a patient. In the GCS, the patients' eye, motor and verbal responses have scores assigned to them. The sum of the three scores is the GCS score, which can range between 3 (indicating deep unconsciousness) and 15 (full alertness).

Temperature

Body temperature (TEMP) is a function of the balance between heat produced and heat lost, a process known as thermoregulation, which tries to stabilise it at about 37^oC in the human body. In the JR ED, temperature is usually measured using an infra-red tympanic thermometer (Figure 2.4 (c)). This measures the infra-red energy emitted from the patient's eardrum in a calibrated length of time. The temperature measurement from the eardrum has been found to be a clinically reliable indicator of body core temperature (Amoateng-Adjepong et al., 1999).

Fraction of inspired oxygen

Patients experiencing difficulty in breathing can be provided with oxygen-enriched air, with a higher-than-atmospheric fraction of inspired oxygen (FiO_2). Natural air is 20.9% oxygen by volume (FiO_2 of 0.21). Oxygen-enriched air is typically maintained below an FiO_2 of 0.5, to avoid oxygen toxicity (Bitterman, 2009). If a patient is wearing a nasal cannula or a face mask, each additional litre of oxygen adds approximately 0.04 to the FiO_2. These values depend on the type of oxygen therapy applied to the patient, and so both the type of device and the FiO_2 value should be recorded on the patient charts. The

use of mask is not a physiological measurement, but it could influence, at least, the SpO_2 that is being measured, and hence it is also an important parameter that is recorded in the patients charts, so clinical staff evaluates their physiology accordingly.

2.3 Track-and-trigger systems

With a view to ensuring in-hospital patient safety, NHS hospitals have developed and used Rapid Response Systems (RRSs) to identify deteriorating patients prospectively. A RRS generally has 3 components: (i) an "afferent limb", which is the mechanism by which the deteriorating patient is identified; (ii) an "efferent limb", which involves teams such as medical emergency teams (MET), rapid-response teams, or critical care outreach (CCO) teams; and (iii) a quality-improvement team, which collects and analyses data and ensures improvement in performance over time (Winters et al., 2013).

Track and Trigger (T&T) systems are an implementation of the "afferent limb" mechanism that triggers a response from the "efferent limb". They traditionally consist of paper-based charts in which the clinical staff periodically record and score physiological measurements, usually known as vital-sign observations, from hospitalised patients. The observations typically recorded are those mentioned in the previous section: HR, RR, SpO_2, Temperature, BP and level of consciousness, either the GCS, or the "Alert-Voice-Pain-Unresponsive" scale, AVPU (National Institute of Clinical Excellence, 2007).

T&T charts use an EWS system to assign scores to vital-sign values, with a higher score being assigned to abnormally low or high values. If the scores exceed a pre-defined threshold, care of the patient is escalated, according to the ward protocol. EWS systems can be categorised as single- or multi-parameter, depending on whether the total score is calculated on a single or the aggregation of vital-sign scores, respectively. The Medical Emergency Team (MET) criteria are an example of a single-parameter EWS system, used in wards to escalate care when there are significant changes in a single vital sign, or a sudden decrease in the level of consciousness. Underlying the use of EWS systems is the

assumption that the physiological processes that produce a catastrophic deterioration, such as cardiopulmonary arrest, are identifiable and treatable 6 to 8 hours before the adverse event occurs (DeVita et al., 2004). The Modified Early Warning Score (MEWS) is an example of a multi-parameter system, in which an alert is triggered if a combination of vital signs deviates from normality (Subbe et al., 2001).

Scoring systems have traditionally been defined heuristically by groups of expert clinicians, and often tuned for use in a specific hospital and ward. Various systematic reviews have shown that this results in low reliability, validity, and usefulness (Gao et al., 2007; Smith et al., 2008a,b). More recently evidence-based EWS systems, in which the score assignment and threshold selection are based on data-driven approaches, either from statistical properties of the data (Tarassenko et al., 2011), or from a score-optimisation process (Prytherch et al., 2010), have been developed and are currently undergoing validation. An example is the National Early Warning Scoring (NEWS) system (Williams et al., 2012), which is being proposed to become a standard for tracking patient vital signs in the UK hospitals.

As observations are taken more frequently in the ED than on general wards, the additional workload and frequent overcrowding can mean that T&T scores are calculated incorrectly. To mitigate this issue, electronic T&T (e-T&T) systems that allow nurses to record vital-sign observations using an hand-held Personal Digital Assistant (PDA), are now being tested in hospital wards. The PDA software has validation steps that check the inputs, automatically calculates the patient score using an EWS and may also notify clinical staff when the time for the next observation is due (Prytherch et al., 2010).

2.4 Evaluation of EWS systems in the ED

Griffiths (2012) reported the results of a survey carried out in 2010, in which 90% of 145 UK EDs supported the use of EWS systems to track patients' physiology. Development of these systems has been an active area of clinical research as EDs experience an

increasing number of patients, and resource allocation is constantly being evaluated to avoid overcrowding while trying to identify patients requiring a higher level of care as efficiently as possible.

The performance of EWS systems in detecting patients at risk of deterioration can be assessed via: (i) retrospective studies, by evaluating the relationship between clinical outcomes and the scores assigned to previously-acquired data; (ii) before-and-after studies, in which an EWS system is introduced as an intervention, and the frequency of clinical markers/outcomes (e.g. length-of-stay, mortality, cardiac arrest or escalation of care) are compared before and after its use; (iii) Randomised Control Trials (RCT), in which patients are allocated randomly to control and intervention groups and clinical outcomes are compared between groups Wong (2011).

Table 2.1 summarises the studies found in the literature evaluating the use of EWS in ED settings (excluding the CDU or other wards that support the ED, but in which the patients can stay up to a day before being discharged). In a total of 9 selected studies two were found to be before-and-after studies, and the remaining were retrospective studies.

One of the before-and-after studies, Olsson (2003), studied the adaptation of the Rapid Acute Physiological System (RAPS, described in Rhee et al. (1987)), using only four non-invasive physiological variables (mean arterial pressure, HR, RR and GCS), readily available in emergency transportation systems and hospital wards, to predict mortality in patients that attended the (non-surgical) ED of a large teaching hospital in Sweden, and were latter admitted to the hospital. Age and oxygen saturation were subsequently added to RAPS to create the Rapid Emergency Medicine System (REMS). RAPS and REMS were applied to 885 consecutive ED admissions.

Table 2.1: *Evaluation of EWS systems in the ED. MTS - Manchester Triage System; REMS - Rapid Emergency Medicine Score; RAPS - Rapid Acute Physiology Score; MEWS - Modified Early Warning Score; MET - Medical Emergency Team criteria; CIC - Clinical Instability Criteria; ASSIST - Assessment Score for Sick patient Identification and Step-up in Treatment. P - Prospective; R - Retrospective; E - Exploratory.*

Country	Design	Author (Year)	Sample	Process	Systems	Main result
SE	P	Olsson (2003)	885	ED stay	REMS	REMS More associated with in-hospital mortality than RAPS.
UK	R	Subbe et al. (2006)	151	Triage	MEWS MET ASSIST	EWS not better than MTS in predicting ICU admission.
CH	R	Etter et al. (2008)	452	ED stay	MET	EWS associated with hospital mortality, need for mechanical ventilation, and hemodynamic instability.
TR	P	Armagan et al. (2008)	309	ED stay	MEWS	EWS associated with ICU admission, ED and in-hospital deaths.
SA	R	Burch et al. (2008)	790	Triage	MEWS	EWS associated with in-hospital admission and death.
USA	R	Heitz et al. (2010)	299	ED stay	MEWS	EWS associated with composite outcome (24-hour death or escalation to higher level of care).
AU	E	Considine et al. (2012)	204	ED stay	ED CIC	Hypotension and tachycardia were the most common reasons for EWS alerts in the ED.
SG	R	Le Onn Ho et al. (2013)	1027	Triage	MEWS	EWS associated with composite outcome (30-day mortality, admission to ICU).
AU	R	Hosking et al. (2014)	200	ED stay	ED CIC	ED CIC better than MET at identifying deteriorating patients.

The authors reported that REMS improved the prediction of in-hospital mortality in ED patients when compared with RAPS, using the Area Under the Receiver-Operator-Characteristic (AUROC) metric, which is used in this context to rank the probability of a scoring system in predicting the correct clinical outcome[3] (AUROC of 0.910 and 0.872, respectively).

Armagan et al. (2008) used the Modified Early Warning Scoring System (MEWS) to prospectively score the vital signs of 309 patients attending the ED of an urban academic care centre in Turkey, and stratify their risk of dying in the ED ward, or having an unscheduled ICU or hospital admission, dying in-hospital after the ED stay. Patients with MEWS ≥ 4 were considered high-risk. Patient care and disposition[4] were based on the judgement of physicians according to the usual standard of practice. The authors showed that patients in the high-risk group had an odds ratio (OR) of 1.95 of being admitted to the ICU, 35.13 of dying in the in ED and 14.81 for in-hospital death compared with low-risk patients (MEWS < 5, reference category). The authors stated that the small sample size (for a 5-month study) did not allow the results to be extended to other ED settings.

Regarding retrospective studies, Subbe et al. (2006), Le Onn Ho et al. (2013), and Burch et al. (2008) evaluated the clinical usefulness of EWS systems at triage time, i.e, when scoring the clinical observations acquired at that stage of the ED admission. In Burch et al. (2008) a version of MEWS reduced to 5 vital signs (SBP, HR, RR, TEMP and level of consciousness) was shown to be associated with risk of in-hospital death, and hospital admission (p-value < 0.001) for 790 patients presenting to the ED of a public hospital in Cape Town, South Africa.

In Le Onn Ho et al. (2013) MEWS showed poor performance (AUROC $= 0.71$, PPV[5] $= 0.17$) in identifying those ED patients whose outcome was death, or admission

[3]The use of AUROC to determined the performance of EWS is described in section 5.1.1, chapter 5

[4]Patient condition at discharge from the ED ward.

[5]Positive Predictive Value, defined in chapter 6

to the ICU or High Dependency Unit (HDU), from a cohort of 1024 attending a large tertiary ED in Singapore. As it was an Asian cohort, the authors comment that the MEWS system used in western hospitals may not be calibrated for the Asian population.

Subbe et al. (2006) compared the use of MEWS, ASSIST (Assessment Score for Sick patient Identification and Step-up in Treatment), and the MET criteria, with the Manchester Triage System (MTS) in identifying patients at risk of ICU admission. The study only made use of the vital signs accessed at triage time, on patients from a UK district general hospital, and found that these systems were not significantly different from MTS in finding patients at risk of ICU admission. The authors indicated that one of the limitations of these studies is the fact that the results might have improved by analysing the EWS of the subsequent observations, and using EWS just at the triage phase may not be the most efficient way of determining the risk of ED patients deteriorating.

Etter et al. (2008) and Heitz et al. (2010), studied the association of the maximum EWS score during ED patients' stay and adverse event outcomes. Etter et al. (2008) analysed 452 consecutive adult patients admitted to intensive care from the ED, and showed that both the maximum MET criteria and the initial MET criteria values were associated with hospital mortality (OR were 3.392, and 3.867, respectively), ICU admission (OR were 4.151, and 4.292, respectively) and hemodynamic instability [6] (OR 1.548, and 1.685, respectively) (all p-value < 0.0001). Heitz et al. (2010) applied MEWS (without SpO_2) to 299 ED patient admissions from one tertiary care academic medical centre in the US. Logistic regression was used to model the log odds of needing higher levels of care (primary outcome composite included all-cause mortality and higher care utilisation within 24 hours) as a function of the $MEWS_{max}$ score. The $MEWS_{max}$ score was shown to be associated with the composite outcome (odds ratio was 1.6).

Considine et al. (2012) created an EWS for the ED similar to the MET criteria, which they called the Clinical Instability Criteria (CIC), for adult and children populations. The CIC has lower trigger thresholds for tachycardia, and tachypnoea, and higher

[6] Defined as the need for vasopressors or inotropes throughout an ICU stay.

thresholds for bradycardia and bradypnoea. The authors showed in a prospective study of 204 ED patients from an Australian urban district hospital, that hypotension and tachycardia were the most common reasons for ED EWS activation. The ED EWS resulted in at least two reports of clinical deterioration in ED patients per day, indicating reasonable uptake by clinicians.

Hosking et al. (2014) compared the MET criteria (designed for general wards) with the ED CIC, and showed that the latter identified a higher number of ED patients at risk of death, cardiac arrest or activated MET calls. However, this result was not statistically significant due to the small sample size (200 patients from a 400-bed regional hospital, Australia). Also the EWS systems were evaluated using just a single set of vital signs per patient.

In summary, most of the described studies (i) use a subset of vital-sign data collected in the ED to evaluate vital-sign scoring systems in this environment; (ii) indicate that the original heuristic EWS systems require re-calibration to be more clinically useful for the ED population; (iii) have used a small population sample which makes it difficult to draw conclusions.

A recent clinical review of 44 studies on the prognostic value of EWS in the ED and in the CDU, from Panday et al. (2017), also notes that this evaluation is difficult as each study uses a slightly different variation of the EWS system. They therefore propose that future studies should use at least NEWS, and the Mortality in Emergency Department Sepsis score (MEDS) system to assess the value of EWS systems in the general ED population, and in ED patients with an infection or sepsis, as the systems showed the best performance in these populations, respectively, in their review. We note also that all of the described studies and all those in the review of Panday et al. (2017), except that of Considine et al. (2012), use death, unscheduled ICU admission or a combined outcome (death and unscheduled ICU admission) to assess the performance of EWS in the ED. Next we introduce a pilot study that assessed the performance of EWS systems when using physiological deterioration events occurring during the ED stay (prior to death or

escalation to the ICU).

2.4.1 Identifying physiological deterioration during the ED stay: Pilot observational study

Wong (2011) analysed the data from the first ED observational study (Jan. 2009 to Jan. 2010) in the JR ED, a single-centre, prospective, study, recruiting ED adult patients ($\geq$ 18 years of age) during the times when there was a member of the research team on duty (typically daytime hours). Patients that did not consent, or with fewer than three observation sets for their ED visit, were excluded.

The primary outcomes were to (i) identify those types of clinical escalation that can be tracked by an EWS system; (ii) estimate the accuracy of paper-based T&T charts for identifying patient deterioration, prior to the escalations; and (iii) study the benefits of using a data-fusion system to detect physiological deterioration from the continuous data collected from bedside monitors.

The dataset for this study included (i) observational data comprising HR, RR, SpO_2, Temperature, SBP, DBP, and GCS, collected from patients' observation charts; (ii) continuous vital-sign data from bedside monitors comprising HR, RR, and SpO_2, sampled each 30 seconds, and SBP and DBP, sampled every 15 or 30 minutes; and (iii) patients' hospital notes, including demographics and clinical outcomes (in-hospital mortality, ICU admission, and other clinical escalation types and timing).

Figure 2.4 (a) shows an example bedside monitor screen with these vital signs.

2.4.1.1 Labelling of clinical escalations

A clinical escalation is an increase in the patients' clinical care during their ED stay, with the aim of avoiding physiological deterioration. The labelling process of the clinical escalations in this study comprised two steps: (i) an ED consultant and an ED nurse consultant reviewed the clinical notes independently, and then labelled the times when

patients were deemed to have been escalated; (ii) the two assessments were reviewed by an ED specialist registrar to determine correct labelling when the first pairs of labels differed.

Escalations were categorised as being either due to **(a) cardiorespiratory** condition, **(b) neurological** condition, or **(c) "non-physiological"** cause (in which no abnormal observations were recorded but the patient was reviewed by clinical staff due to clinical concern). When multiple types of escalations existed concurrently for the same patient, a hierarchy was defined to assign one escalation category to the patient: cardiorespiratory escalations were given priority over neurological escalations. The latter were given priority over non-physiological escalations. These escalation groups were further subdivided into patients **(i) escalated at arrival**, and patients **(ii) escalated after arrival** to the ED.

2.4.1.2 Summary of study results

472 patients fulfilled the study inclusion criteria; 51% were male and the mean age was 61. The median LOS was 4.6 hours. 387 escalations were labelled as occurring for 204 patients, 84% of whom were subsequently admitted to another hospital ward after their ED stay. The remainder were discharged home. 75% of escalations occurred on arrival to the ED, as shown in Table 2.2. The median time to an escalation after arrival was 3.2 hours, as shown in Figure 2.5.

Serious adverse events (in-hospital mortality, unplanned admission to ICU, and escalation to the resus area) occurred with low frequency (2.5%). Only one ICU admission and three escalations to the resus area were considered to have occurred after arrival.

2965 observations sets were recorded, with only 34.5% having a corresponding T&T score, of which 29.6% were incorrect. Most of the errors (93%) were attributed to incorrect assignment of the score to an individual vital sign, and the remainder (7%) were deemed to be due to incorrect addition of individual EWS scores. In some cases it was found that the previous EWS value was copied to the next observation set (Wilson et al., 2012).

Table 2.2: *Location and timing ("At" or "After" the patient arrival to the ED) of clinical escalations that occurred in the pilot study dataset.*

Location	Timing (minutes) (At arrival)	(After arrival)
CDU	2	24
Majors	38	22
Resus	250	51

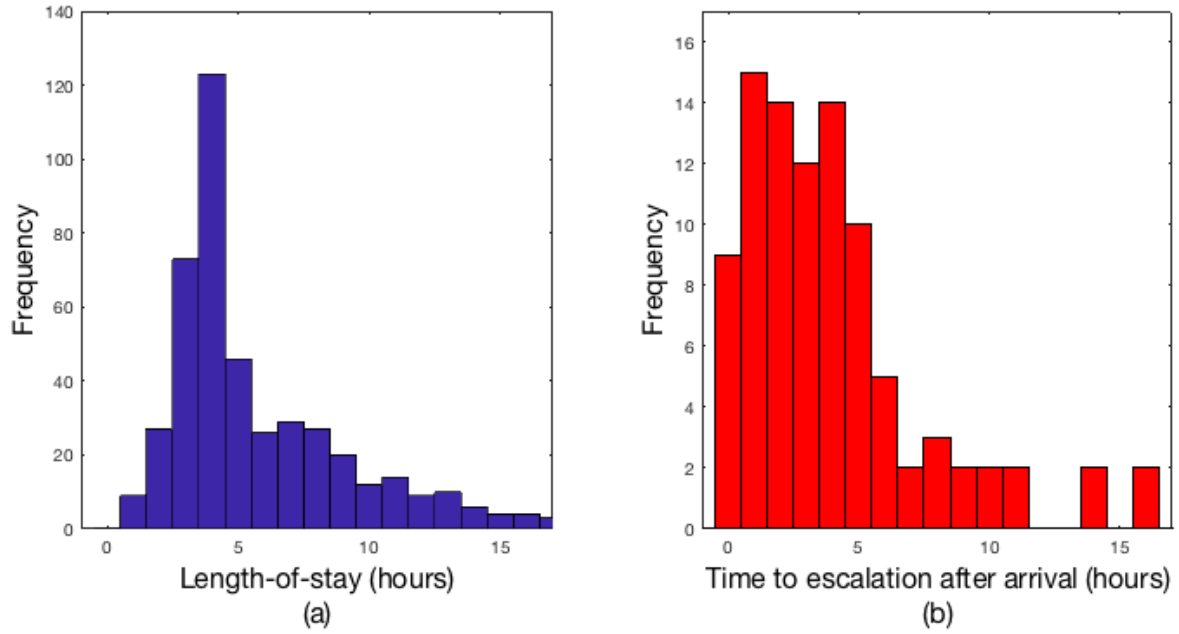


Figure 2.5: *(a) ED patients' length-of-stay (LOS). (b) Time to clinical escalations (including physiological, neurological or non-physiological) occurring after arrival to the ED.*

Using a patient-wise performance analysis (detailed in section 8.3.4), it was shown that the performance of EWS registered in the T&T charts during the study, improved from 0.47 to 0.94 in sensitivity, and decreased from 0.87 to 0.70 in specificity, when re-computed electronically, retrospectively.

Continuous data existed for 85% (2169 hours, for 400 patients) of those patients that were attached to bedside monitors, with mean data loss of 21%. Skin temperature acquired continuously by skin-mounted sensors was shown to be unreliable (as also reported in (Hann, 2008)). Retrospective analysis of the continuous Patient Status Index (PSI, described in section 6.3.1), determined from these data, showed that the labelled cardiorespiratory and neurological escalations occurring after arrival could be identified earlier than when using the EWS, for some patients (sample size was small to provide a significant statistic), with a calculated false-alert rate of 1.13 alerts/bed-day (Wilson et al., 2014).

2.5 Continuous monitoring of the patient condition

Taenzer et al. (2011) reviewed some of the causes behind in-hospital death after adverse events, a phenomenon described as "failure-to-rescue". They identified two main limitations in the afferent limb: (i) intermittent clinical observations may not have enough temporal resolution to identify early deterioration, and (ii) these observations rouse patients resting in hospital beds, temporarily altering their vital signs, making their assessment inaccurate. Due to their intermittent nature, T&T systems may be unable to identify sudden clinical deterioration in patients.

Continuous bedside monitoring systems, which are standard in higher acuity units such as the ICU, may identify deterioration that might otherwise be missed between clinical observations. Figure 2.6 shows a representation of the different components in traditional continuous monitoring systems, in this area in the ED. DeVita et al. (2010) suggested that the HR, RR, SpO_2, BP, and temperature should be included in patient

surveillance systems, and most bedside monitors collect them.

Bedside monitors typically alert when any one of the vital signs is outside the range defined by fixed thresholds. The latter may be chosen by nurses. Such systems typically have a very high sensitivity [7] and very low specificity [8](Tsien and Fackler, 1997). Consequently, the false-alert rate is high, and results in staff ignoring the alarms, as was found in three international studies of critical care monitoring systems (Phillips and Barnsteiner, 2005; Graham and Cvach, 2010; Siebig et al., 2010). Additionally, if alarms are temporarily disabled by a clinician, other staff may not be informed, making the system potentially more hazardous than having no alarms at all.

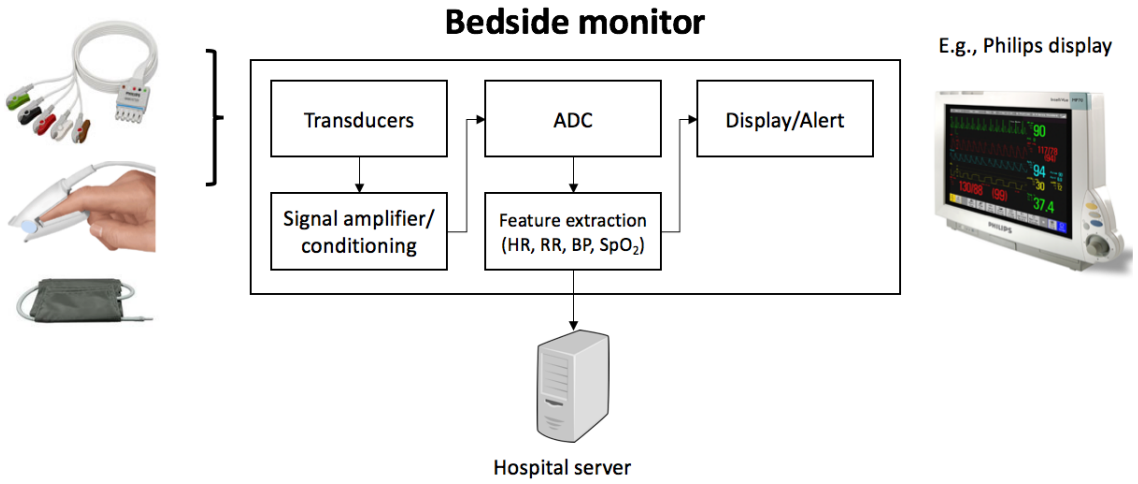


Figure 2.6: *Main components of a bedside monitor used in ED wards. Transducers are used to convert the physiological signals into electrical signals which are then amplified and filtered. An analogue-to-digital converter (ADC) is used to acquire data at a high sampling-frequency, and algorithms are used to extract physiological variables such as HR, RR, SpO_2 and BP. The waveforms, features, and alerts are then displayed on the monitor.*

Whilst their availability in an ED can be viewed as a positive step in meeting the standard for early recording of vital signs (National Institute of Clinical Excellence, 2007), little or no evaluation of using continuous monitoring systems in UK emergency care has been undertaken. Way et al. (2014), which also used data from the pilot study described in the previous section, analysed the frequency, duration and type of audible monitor alarms from 110 patients admitted to the majors area or resuscitation Room of the

[7]System's ability to correctly detect patients with the condition.
[8]System's ability to correctly detect patients without the condition.

JR ED, using the standard manufacturer's classification (Philips Intellivue physiological monitors). Alarm noise was generated for 29% of the observation time. From a total of 93 hours of data, 429 alarms lasting 21.5 hours were judged to be positive and 143 alarms lasting 5.8 hours, negative. 74% of Resuscitation Room and 47% of Majors alarms were silenced or paused by the nursing staff. Alarm limits were altered on a small proportion of patients in both areas of the ED (5% in the Resuscitation Room and 6% in the Majors area). This highlights that nurses in this department do not generally configure the alarm limits in relation to an individual patient's physiology. There was therefore a high probability of near-continuous alarm noise from patient monitoring in a 10-bedded Majors area. The authors concluded that whilst high-level monitoring is desirable from a patient safety perspective, it contributes to a significant ambient noise level, which can be detrimental to patients and staff.

Taenzer et al. (2011) noted that automatic alerts could be managed more effectively by the correct adjustment of (i) alarm thresholds and (ii) notification delays. The authors suggest that thresholds could be set based on the "normal" distribution of typical patient physiology for a ward. Notification delays would help eliminate transient false alerts, by stipulating an amount of time that alerts must persist before sounding.

2.6 Conclusion

The use of T&T is becoming standard in most in-hospital wards, and EWS systems are perceived as useful tools in identifying patient deterioration in the ED. Most studies evaluating heuristic EWS systems in the ED showed that short-term adverse events were regularly missed. However, most used a subset of vital-sign data and were conducted retrospectively. Early corrective treatment depends on many other factors apart from the EWS, such as the protocol being followed by clinical staff, and their ability to respond to alerts in a busy setting.

From the pilot study carried at the John Radcliffe Hospital ED it was possible to

observe that the EWS process is undermined by the (i) known human error in manually recording and scoring vital signs on paper charts, and (ii) the low completion rate of vital-signs sets. It was therefore hypothesised that electronic T&T, based on data entry on hand-held devices, could help decrease errors by automating key steps, such as the calculation of the EWS, the recording of the observation time, and the display of when the next observation set is due.

In addition, the amount of physiological information available for decision support in the ED can be maximised by taking into account the continuous vital-sign data collected from bedside monitors, provided that false alerts are avoided through the use of multi-parameter criteria and of notification delay approaches.

Chapter 3

Large-scale ED study

3.1 Background

The large-scale ED study, which occurred in the JR ED, 2011-2013, was designed to test the feasibility of using two systems to help identify patient deterioration in the ED, prospectively: (i) an e-T&T system, installed on PDAs, and (ii) a data-fusion system using the continuously-acquired data from bedside monitors to generate automatic alerts in periods of physiological deterioration. In this chapter, a description of the overall design and dataset of the ED study is given.

Paper-based T&T system

T&T systems are usually implemented in paper charts, which allow clinical staff to record vital-sign data, their timings and their individual scores, which are added up to determine the final aggregated EWS. Figure 3.1 shows an example of a paper chart used in the ED.

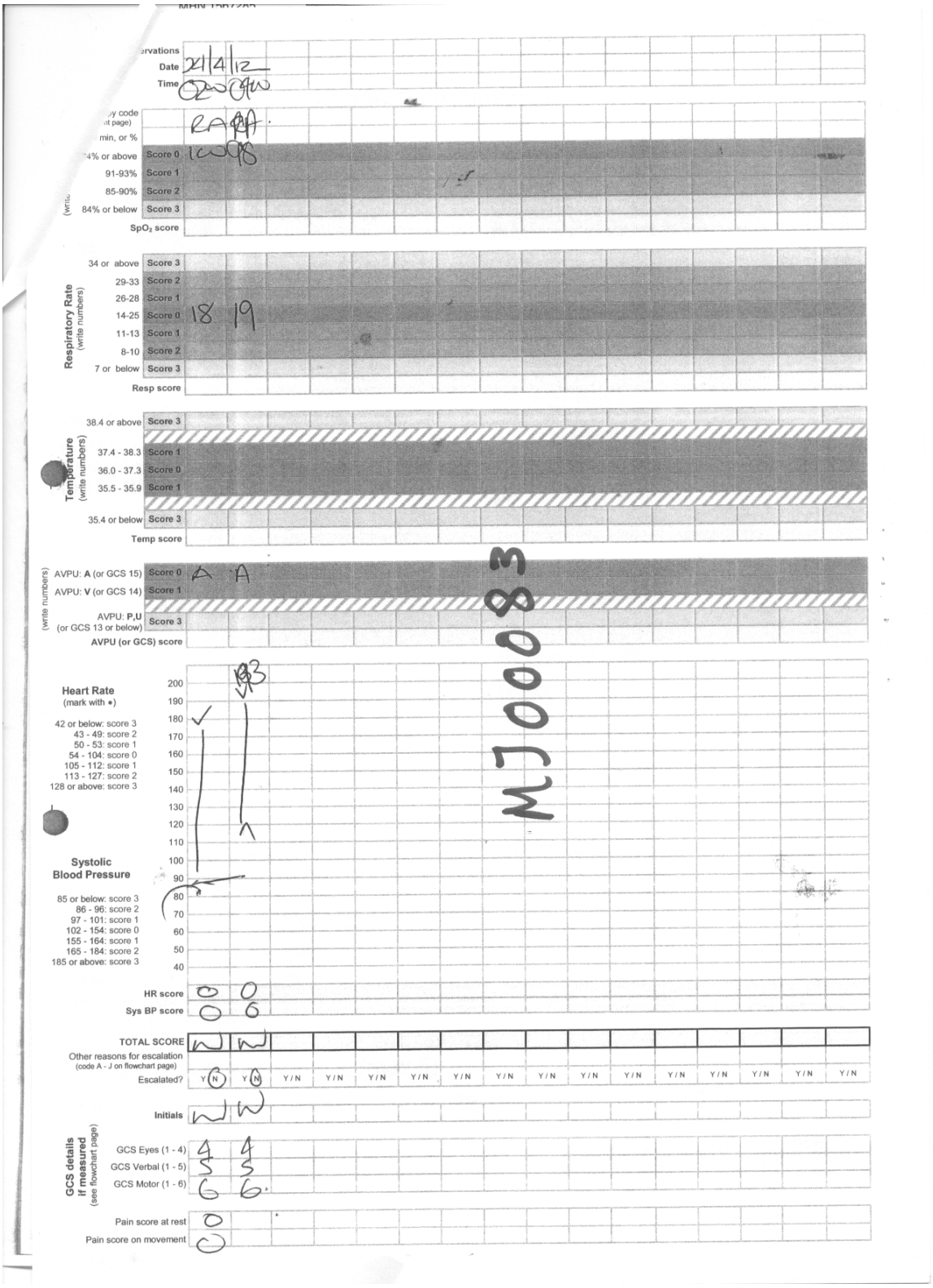


Figure 3.1: *Paper-based T&T charts used in phase 1 of the large scale ED study (explained in the next section). An error can be observed in scoring high SBP in the second observation set (filled in the second column of the blood pressure grid).*

This chart represented the method of recording the T&T data in the JR ED, up to 2016[1]. It can be observed that the hand-writing can be illegible, and the location of the HR and BP marks, can be mis-interpreted, undermining their interpretation and scoring. In this particular chart a score of 0 was assigned to high SBP ($\geq$ 185 mmHg) incorrectly, and the patient was not escalated, and deemed to have "normal physiology", as marked in the chart. In such cases deterioration may be missed. As described in Wilson et al. (2012), this shows that paper-based T&T in the ED is prone to human error.

Additionally, ward-specific clinical data can be recorded on complementary paper charts. In the ED, for instance, a paper-based Triage-chart, is used to record the first set of vital-sign observations, and their timings, in the triage room. EWS are not used with these charts, but sometimes these data are copied to the T&T chart to be scored.

Electronic T&T system

More recently, electronic implementations of T&T systems have been found more suitable to track patients' vital signs on the ward (Prytherch et al., 2010). Figure 3.2a, shows an example e-T&T system, called "VitalPAC", which was customised for the JR ED, and used during the large-scale study.

On arrival at the ED, patients are registered in the hospital system, and become available to the VitalPAC software. Clinical staff can then enrol these patients in the VitalPAC, and assign the specific ED area (minors, majors, resus) in which they are receiving care. Vital-sign observation sets data can then be recorded using the VitalPAC PDA for each patient while they are in the ED. Automatic validation checks notify nurses in case extreme values are recorded, otherwise, the values are scored, according to the EWS thresholds as in paper charts. The physiological measurements data and timing are then saved into a central server managed by the Oxford University Hospitals NHS Foundation Trust.

[1]Since 2016 the SEND system described in Wong et al. (2015) has been used for e-T&T in the JR ED. The paper-charts are used as a backup system.

The observations are only scored if a "full observation set", which includes HR, RR, SpO_2, SBP, TEMP and GCS is made. In case it is not possible to record a vital sign, such as temperature, a reason needs to be given for the parameter not being recorded, and the aggregated EWS is still calculated. In case clinical staff decides not to the give the reason for the missing data, the observation set is not scored, and is called a "special observation set". Scoring the observation set generates the time the next observation is due, guiding clinical staff as to when to make the next set of observations.

The system allows clinical staff to review the last EWS assigned to each patient on PCs located at three nursing central stations in the middle of the ED (an example is shown in Figure 3.2a). Tablet computers, running VitalPAC software, fixed to each bedside monitor, allow the clinical staff to review the time-series of vital signs recorded through the PDAs.

Finally, these systems allow the reconciliation of the observational data with the patients' Electronic Patient Record (EPR) through electronic means.

Data-fusion system

A data-fusion system is a device that scores continuous vital-sign data, collected from bedside monitors, using a multi-parameter model, and generates an alert when the score is above a pre-defined value generated by the model. It is a clinical decision support tools that helps clinical staff to identify patient deterioration from continuous streams of vital-sign data.

(a) VitalPAC views

(b) Visensia views

Figure 3.2: *a) VitalPAC components used in the ED: top - example of a view of electronic T&T displayed on a tablet; bottom left - example list of active patients on a hospital ward and their risk scores (EWS), located at "nursing central stations"; bottom right - views displayed on PDAs carried by the nursing staff in the ED. One view shows the nurse's patient list, and the other shows the time-series of observed vital-sign values for one patient. b) Visensia components used in the ED: left - display of the latest Patient Status Index values and its time-series for a group of six patients attending a hospital ward, displayed on the desktop computed monitor located at a nursing central station; right - PSI score and vital signs contributing to that score, for a single patient, displayed on a tablet located at the patient's bedside.*

Tarassenko et al. (2006) developed and validated a data-fusion system to monitor vital-sign data on general wards. This system used a novelty detection approach to score a set of vital signs, including HR, RR, SpO_2, the Systolic and Diastolic blood pressure Average (SDA), and Temperature, against a joint probabilistic model of vital-sign distributions collected from in-hospital patients whose condition was deemed "normal". The furthest a set of data is from the high probability values of the joint probabilistic models, the more "abnormal" they are deemed to be. The score produced by this system is called the Patient Status Index (PSI). Details of the algorithm are reviewed in section 6.3.1.

Figure 3.2b (right), shows an early commercial version of this system called *Visensia* (previously *Biosign*), installed on tablet-computer devices, fixed to each patient bedside monitor. On the left side of this figure, a Visensia central station is displayed on the screen of a desktop computer (typically located at each nursing station). This allows clinical staff to check the novelty score (usually from 0, normal, to 6, very abnormal) associated with each patient, changing at a rate that varies from 5 to 30 seconds, depending on the frequency at which the bedside monitors make their vital-sign data available to other devices. This system obtained approval from the Food and Drug Administration (FDA) in 2008.

3.2 Study design

All patients over 16 years of age in the majors area of the JR ED, were enrolled sequentially during three study phases. Each phase ran for two months. It was initially estimated that 3,000 patients would attend the majors area in each phase, for an expected total of 9,000 ED visits. An illustration of the systems used to monitor patients' vital signs in each phase is shown in Figure 3.3, and each phase was set-up as follows:

Phase 1 (usual care): Vital signs were recorded using both Triage and T&T paper charts. EWS values were manually calculated and written on the T&T chart for each

Table 3.1: *Centile-based EWS (CEWS) system, described in Tarassenko et al. (2011), and used in the large-scale ED study.*

	Centile-based early warning score						
Variable	3	2	1	0	1	2	3
HR (bpm)	≤ 42	43 - 49	50 - 53	54 - 104	105 - 112	113 - 127	≥ 128
RR (rpm)	≤ 7	8 - 10	11 - 13	14 - 25	26 - 28	29 - 33	≥ 34
SpO_2 (%)	≤ 84	85 - 90	91 - 93	≥ 94			
SBP (mmHg)	≤ 85	86 - 96	97 - 101	102 - 154	155 - 164	165 - 184	≥ 185
TEMP (^{o}C)	≤ 35.4		35.5 - 36.9	36.1 - 37.3	37.4 - 38.3		≥ 38.4
GCS	≤ 13		14	15			

patient. The Centile-based EWS system (CEWS), shown in Table 3.1 (explained in detail in section 5.1.1, chapter 5) was used to score the ED patients vital-signs sets. Clinicians used the protocol described in Table 3.2, to escalate the care of ED patients.

Patients in the majors area in the JR ED, are routinely connected to Philips bedside monitors, which use standard alert systems based on pre-set thresholds for each vital sign. At this stage, the data-fusion system was only operating as a data collection system for the continuous vital-sign data (HR, RR, SpO_2 and BP).

Table 3.2: *Clinical protocol for the desired frequency of taking observations according to EWS. TTNO - time to the next observation set.*

EWS	Risk	TTNO	Message shown to recorder
0	Low	2 hours	-
1-2	Medium	1 or 2 hours	1. Review observations in the context of triage category & consider medical review; 2. Inform senior nurse in area of escalation.
0-2 with concern	High	1 hour	You are concerned about the patient: 1. Review observations in the context of triage category & consider medical review; 2. Inform senior nurse in area of escalation.
3-5	High	1 hour	1. Review observations in the context of triage category & consider medical review; 2. Inform senior nurse in area of escalation.
> 5	Critical	30 minutes	1. Carry out medical reviews; 2. Inform senior nurse in area of escalation; 3. Consider move to resus.

Phase 2 (e-T&T intervention): All vital signs collected during the patient's ED stay

(including triage) were recorded using VitalPAC PDAs (Prytherch et al., 2006).

The VitalPAC devices prompted clinical staff to repeat observations at intervals defined in Table 3.2. For this phase and for the next phase of the study, vital signs from the bedside monitors were also sent to the VitalPAC system whenever an automated BP measurement was made, using the data-fusion software. These observations were displayed as "unvalidated" observations sets on the clinical staff PDAs, because they were sent from the monitors rather than directly observed and recorded by clinical staff ('validated' observations, which include the standard and "special" observations described in section 3.1). With this approach clinical staff could have access to patient's vital-sign estimates between the nurse's clinical observations.

In this phase the data-fusion systems were also used to collect continuous vital-sign data from the bedside monitors.

Phase 3 (e-T&T + Visensia intervention): Observation sets were recorded and scored with the e-T&T system as in phase 2. Visensia (Tarassenko et al., 2005) was used with the continuous vital-sign data from the bedside monitors, and generated alerts using the persistence criteria (section 6.3.1). These alerts comprised visual and audible prompts from the Visensia tablet computers mounted near the bed. Nurses were asked to record an extra set of observations whenever an alert was generated. Clinical staff could silence the alert for a default period of 30 minutes. In this way Visensia would not alert during treatment or diagnosis of a clinical problem.

We note that the ED triage process changed between phase 2 and phase 3 (review Figure 2.1, chapter 2). The Manchester Triage System was replaced by the **Nurse Assessment** procedure. As a result, in phase 3, some patients in the majors area were allocated to majors chairs, and thus not monitored by the data-fusion system. Also the nurse assessment was performed in majors cubicles 1 to 6, in which the alerts of the data-fusion system were not operational. Once the patients had been initially assessed, if they required further care in the ED they were transferred to one of the remaining majors area beds, in which the alerts of the data-fusion system were operational.

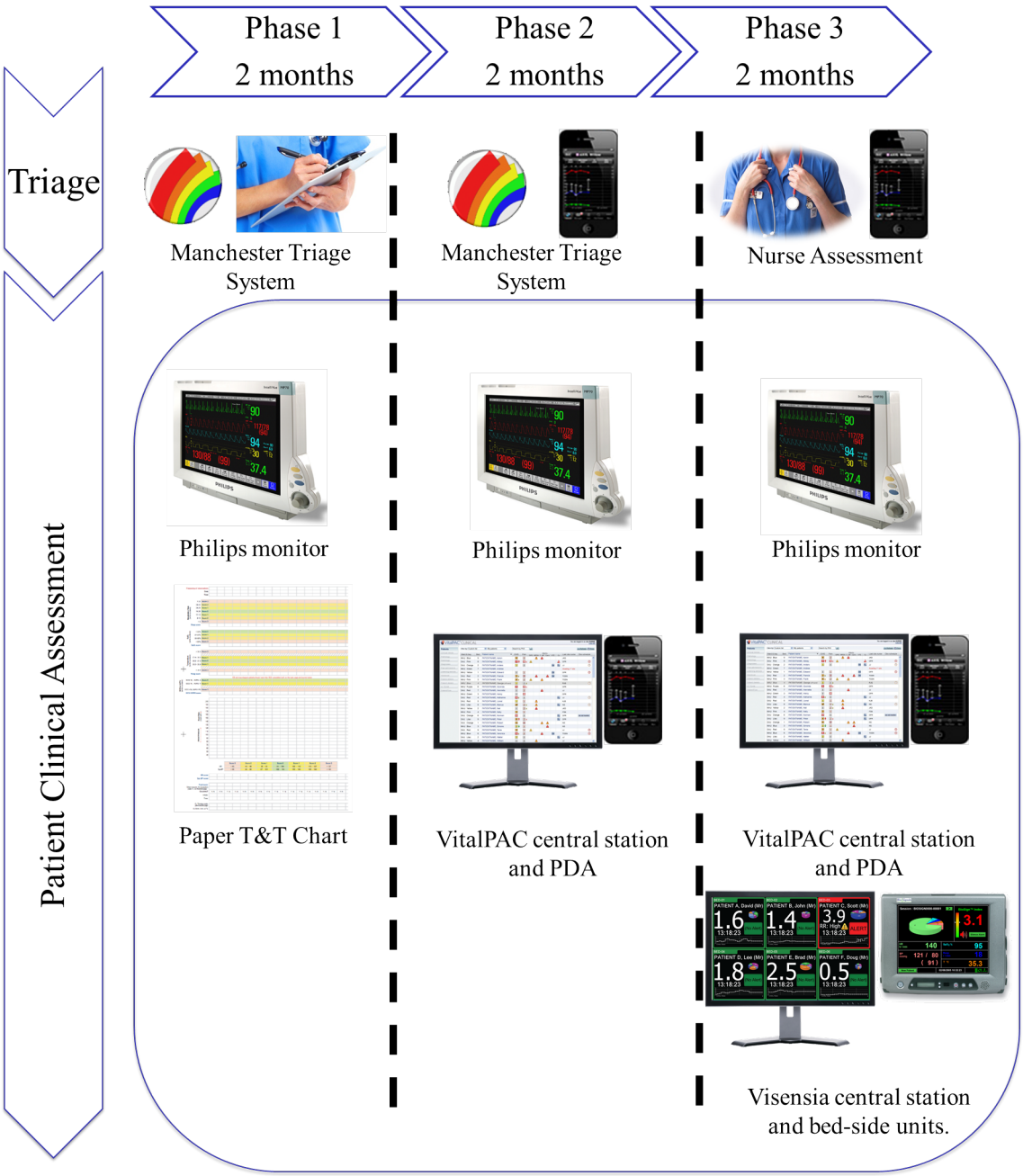


Figure 3.3: *The devices used to monitor patients' vital signs in each phase.*

3.3 Clinical outcomes

Permission for the study was granted by UK National Research Ethics Service, South Central (12/SC0074). With the agreement of the National Information Governance Board, consent was not required prior to patient enrolment. The primary and secondary outcome measures for the study are described in Table 3.3.

Table 3.3: *Primary and secondary outcomes planned for the large-scale observational study.*

#	Primary outcomes
I	% of accurately recorded EWS.
II	% of patients for whom Visensia alerts prompted further clinical observations.
	Secondary outcomes
III	Frequency and duration of periods of physiological abnormality, defined by local EWS.
IV	24-hour, 48-hour, 15-day and 30-day mortality (including out-of-hospital).
V	Hospital length-of-scale (LOS).
VI	% of patients escalated to the resus area.
VII	% of patients requiring hospital admission $\geq$ 5 days after arrival.
VIII	% of 30-day in-hospital Cardiopulmonary Resuscitation.
IX	In-hospital mortality.
X	% escalations to intensive therapy units.
XI	Incidence and causes of downtime of the monitoring systems.

The analysis of the first outcome was published in Pullinger et al. (2015). Chapter 7 analyses the % of patients for whom Visensia alerts prompted further clinical observations (outcome II), and its influence on the duration of physiological instability (outcome III). All remaining outcomes are reported in the current chapter. In addition, the ED length-of-stay, the percentage of patients transferred from resus to the majors area, and the percentage of 2-, 5- and 30-day ED re-attendance rate, are also reported in this chapter.

3.4 Database preprocessing

The data collected in this study were preprocessed as follows:

1. **Observational data (physiological measurements)**:

Phase 1: Data collection for phase 1 was time-consuming because it required transcribing information from paper charts into electronic databases.

Nurses' observations sets were collected from Triage and T&T paper charts from patients meeting the inclusion criteria, and transcribed into a Microsoft Office Access database, located within the hospital and password-protected. Accuracy of transcription of vital-sign values from the observation chart to the research database during stage 1 was assessed using duplicate data entry for an initial sequential sample of 200 ED admissions, 6.2% of a total of 3219 ED admissions. An error was defined if the differences between data entry exceeded the following values: 0.1^oC for temperature, 10 beats per minute for heart rate, 1 breath per minute for respiratory rate, 10 mmHg for systolic and diastolic blood pressure, and 1% for FiO_2. Errors occurred in 35 (1.34%) of 2621 vital-sign values.

Phase 2 and 3: Observational data and the corresponding EWS values, were obtained from the VitalPAC system, along with "unvalidated observations".

2. **Continuous data:**

   The PSI (novelty score) values from the Visensia devices, and the times of alerts and silencing commands, were stored with the continuous vital-sign data. The continuous data were matched to each patient using the following steps:

   (a) The timestamps from BP measurements recorded in the observational data (transcribed in phase 1 and automatically extracted in phases 2 and 3) were compared with the BP measurements in the continuous data, for the beds to which each patient was assigned to.

   (b) Any patients from phase 1 with observations and without a bed assignment were considered separately: the continuous data were matched to the observational data by inspection. With any unmatched patients for phases 2 and 3 the "unvalidated" observations sets captured by the VitalPAC system were used to match with the continuous data.

(c) Patients without observations had continuous data assigned to them by matching their bed numbers and arrival/depart times collected from the EPR.

3. **Demographics data:**

   The age, sex, type of referral to the ED, arrival mode, ED arrival time, Manchester Triage System category, and presenting complaint, were collected from the EPR.

4. **Patient outcomes:**

   The ED discharge time, clinical follow-up decision, discharge location, mortality (in- and out-of-hospital), hospital length-of-stay, and any admission to the ICUs[2] following the ED attendance (including those occurring during the patients' hospital stay) were obtained from the EPR. Cardiopulmonary Resuscitation (CPR) and movement between ED areas were extracted from the Resuscitation Department database and ED room registers. Movements of patients to the resus room were validated using the ED information system.

3.5 Study description

3.5.1 Dataset

Phase 1 occurred between April and June, 2012 (49 days), phase 2 between June and July, 2013 (51 days), and phase 3 between September and November, 2013 (62 days).

Figure 3.4 shows the consort diagram for the data collected in this study. Data from a total of 10,016 patient attendances to the ED majors area were collected, with 3,219, 3,052 and 3,445 attendances in phases 1, 2 and 3, respectively. The diagram shows the number of patients in each of the following categories:

[2]The ICUs included those in the JR Hospital and in the Churchill Hospital, Oxford: Cardiac ICU and adult ICU, and the JR Hospital Cardiac and Thoracic Critical Care Unit and Neuro ICU.

A - all recruited ED attendances, for whom adverse event data (hospital LOS, escalation to resus, escalation to ICU, CPR, and mortality) were collected;

B - all attendances for whom ED documentation (triage/observation charts or e-T&T registration) existed;

C - all attendances for whom complete documentation existed; i.e, all vital-sign data for those patients were transcribed from paper charts or obtained from the e-T&T system for transfer to the study final database;

The sub-indices 1, 2 and 3 are used to group these data by study phase, e.g. group A_1 corresponds to all attendances collected for phase 1. Figure 3.5 shows the daily amount of clinical observations data loss (i.e. patients for whom the paper or e-T&T were not found) for all phases. Data loss occurred randomly, apart from two clear e-T&T system downtimes ($<$ 12 hours) in phases 2 and 3.

Bedside monitor data

Continuous data were matched with 84.6%, 91.31%, and 93.9% of the ED attendances in groups A_1, A_2 and A_3, respectively. The data loss rate in phase 1 is comparable to that of the small-scale study described in chapter 2. Figure 3.6 shows the daily frequency of patients without continuous data in the final database, in each phase. Missing continuous data occurred for three reasons: **(1) Inaccurate or lack of observational/EPR data timestamps**: there was insufficient accurate timestamp information to assign the continuous data to the patients. This was likely to happen in phase 1, during which 33.8% of the patients had incomplete documentation. In all phases, the bed times recorded in the EPR were used when no observation data were available. These timestamps may be imprecise. **(2) Periods of hospital server downtime**. **(3) Visensia devices were deactivated, or Philips bedside monitors were not operational**.

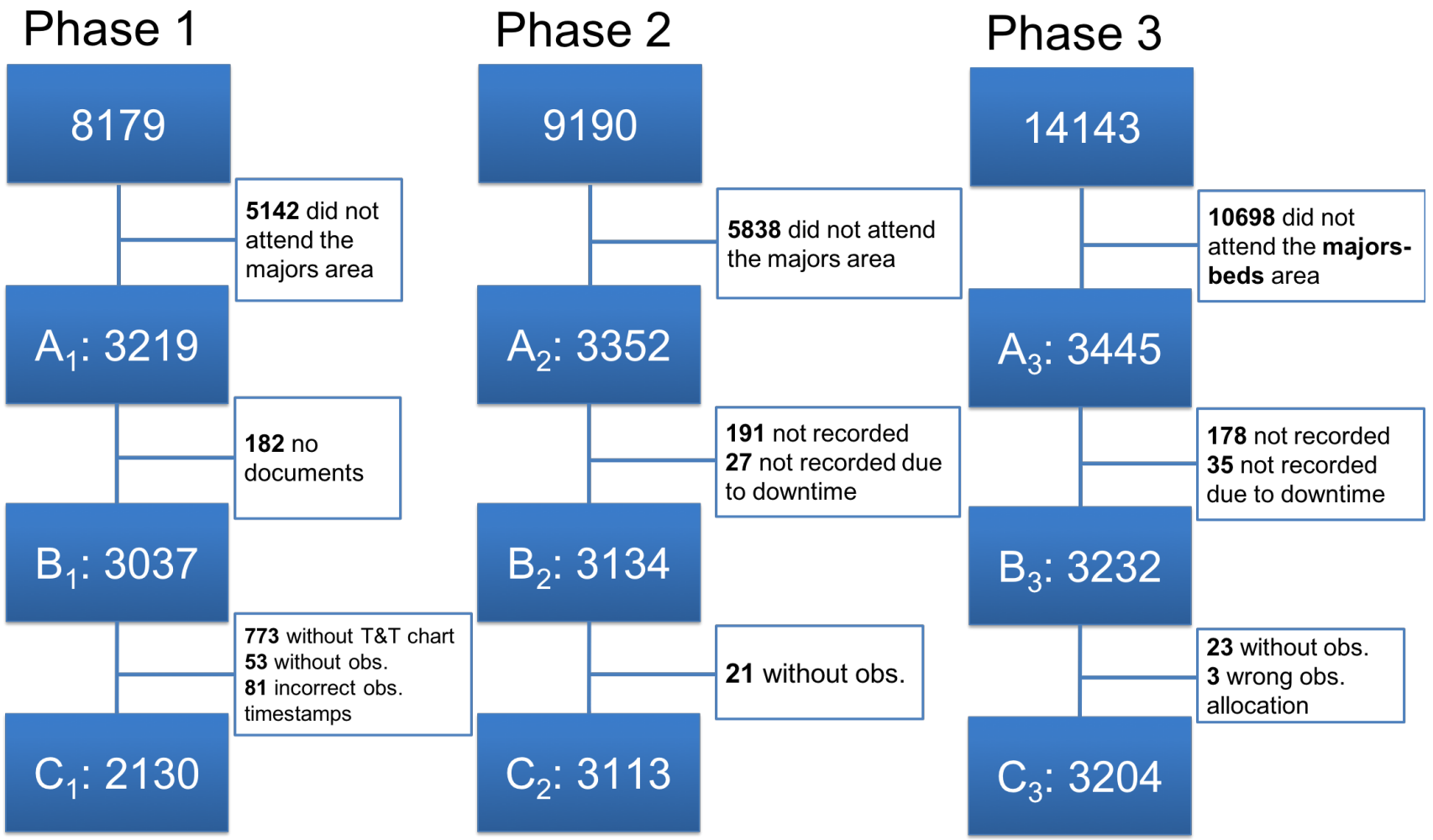


Figure 3.4: *Consort diagram for the number of patients with complete observation set documentation for each phase of the study. The total number of ED attendances for the duration of each phase is shown at the top.*

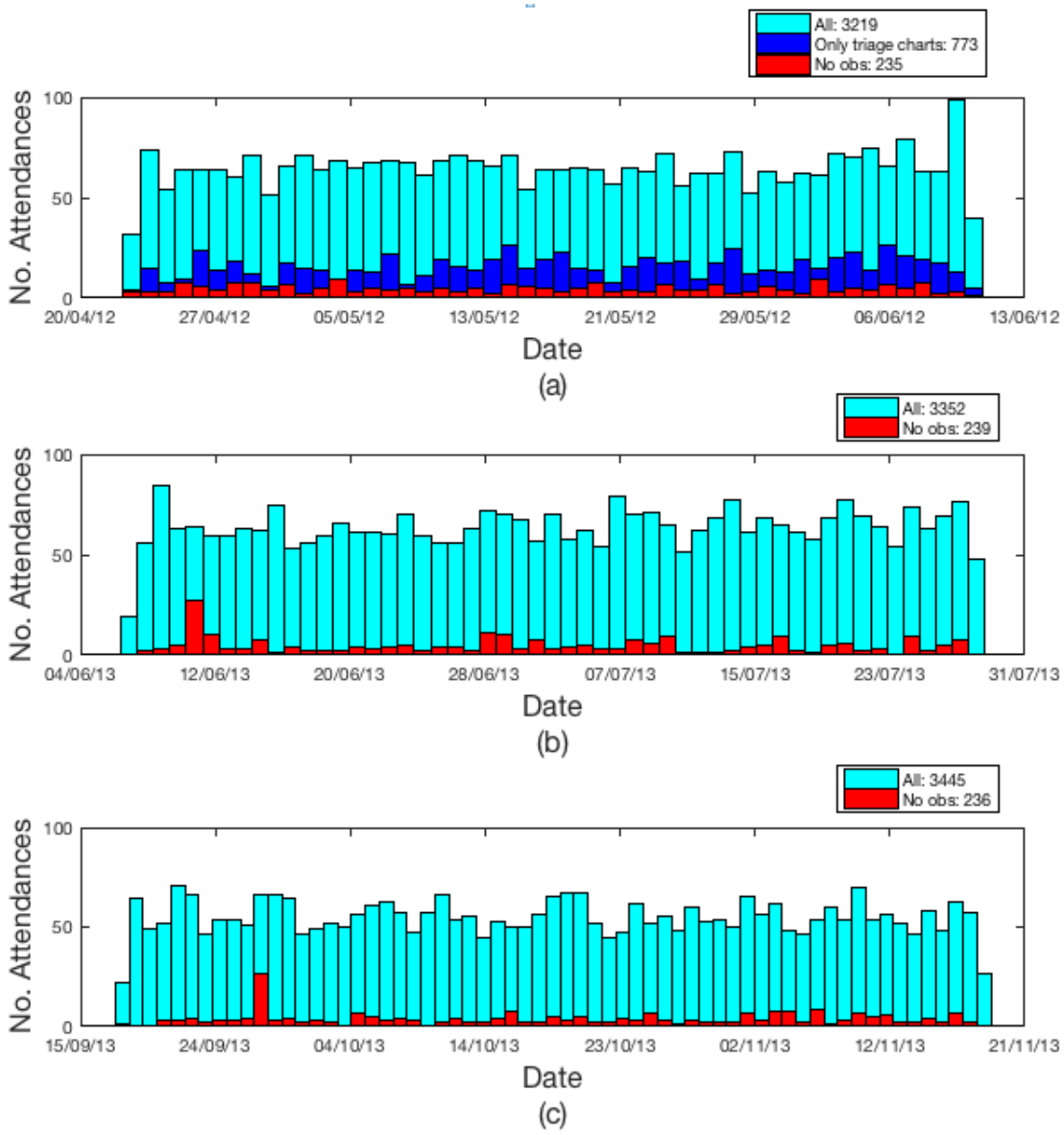


Figure 3.5: *(a), (b) and (c) The number of ED attendances without clinical observations per day (in red), for phases 1, 2 and 3, respectively. In phase 1 the data in dark blue represents the patients who had one observation set extracted from the Triage chart, and for whom T&T charts were not found.*

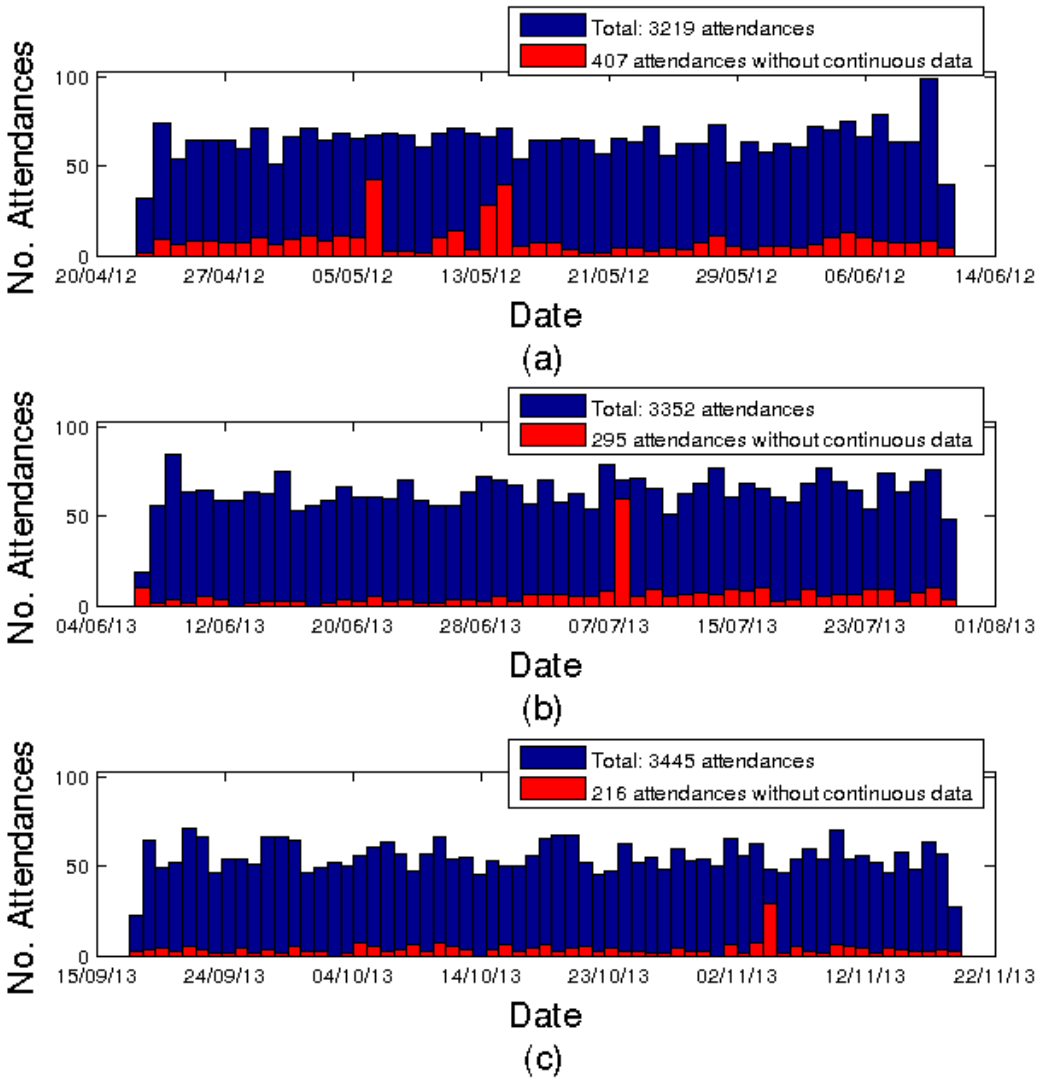


Figure 3.6: *Frequency of ED attendances per day of the study versus frequency of attendances without bedside monitor data per day for each study (a) phase 1, (b) phase 2 and (c) phase 3.*

Continuous data channel availability

93%, 98% and 97% of the total continuous data collected by the data-fusion devices in the majors area were assigned to patients, in each phase, as shown in Table 3.4. The table reports the amount of continuous data collected from each ED area (resus or majors beds) in which recruited patients were monitored. For phase 3, some of the beds were being used for triage, as already mentioned, and the Visensia alerts were not active for those beds, but data was still being collected. The table also presents these data. There

Table 3.4: *Total continuous data collected for majors attendances for all phases, by clinical area. In addition, about 171 hours of continuous data were matched to the majors beds allocated for nurse assessment, in phase 3, and about 277, 301 and 307 hours of continuous data were matched to the resus beds occupied by the cohort of patients that attended the majors area in phases 1, 2 and 3, respectively.*

	Phase 1	Phase 2	Phase 3
Total hours of data collected	5,282.3	6,040.5	5,490.7
Total data matched hours (%)	4,925 (93)	5,892 (98)	5,343 (97)
# Patients with data	2,750	3,057	3,229

is an increase in the quantity of continuous data available from phase 1 to phase 2, which may be because clinicians were making use of the "unvalidated" observations displayed by VitalPAC devices, in phase 2. The quantity of continuous data decreased from phase 2 to phase 3, probably because some patients were first allocated to a majors bed used for triage (where the initial nurse assessment occurred, in phase 3) before being allocated to a non-triage majors bed for further clinical investigation or to receive treatment. This is also reflected in Figure 3.7, in which the quantity of continuous data is shown. We note that phase 2 has at least 11% more data with four vital signs present, than the other phases.

Table 3.5 compares the means and standard deviations (median and IQR for SpO_2) between the distribution of observational and continuous data. Regarding the continuous data, we observed that most distributions are significantly different from the observational data distributions, BP presenting the greatest difference (greater than 2 units, which is not clinically relevant).

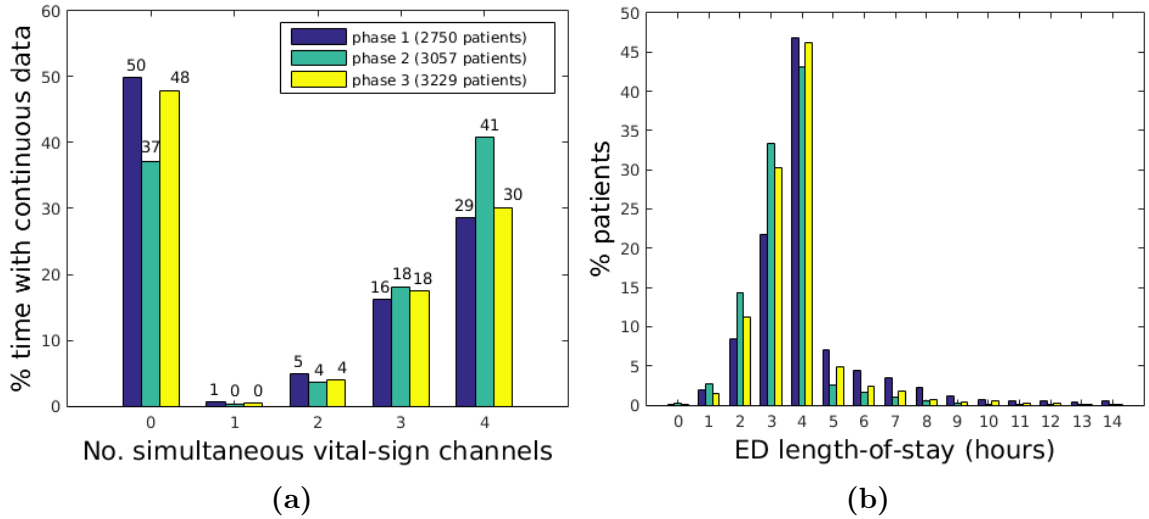


Figure 3.7: *(a) Percentage of data available for total patient ED time, by number of vital signs; (b) Normalized histogram of patient ED LOS.*

Table 3.5: *Means and standard deviations (or medians and IQRs, or proportions) for all vital signs collected from observations sets and bed-side monitors.* [a] *Values presented as median (IQR). The distribution of continuous data is statistically significantly different from that of observational data in most of the cases. The t-test was used for all the continuous-variables, except for SpO_2, for which the Wilcoxon test was used due to its skewed distribution. *p-value < 0.05.*

Vital sign	Source	Phase 1	Phase 2	Phase 3
HR (bpm)	Charts	82.1 ± 18	81.5 ± 18.8	80.8 ± 18.4
	Monitor	80.9 ± 18.8*	81.2 ± 18.8	81.5 ± 19.7*
RR (rpm)	Charts	17.9 ± 3.8	18 ± 4.5	17.9 ± 4.2
	Monitor	18.2 ± 4.8*	18.3 ± 5*	18.3 ± 4.8*
SpO_2 (%)[a]	Charts	98 (96, 99)	98 (96, 99)	98 (96, 99)
	Monitor	97 (95 , 99)*	97 (95, 99)*	98 (95, 99)*
SBP (mmHg)	Charts	134.9 ± 27.4	133.9 ± 27.1	137.9 ± 29.4
	Monitor	132.7 ± 28.8*	131.3 ± 28.2*	133.3 ± 30.1*
DBP (mmHg)	Charts	73.1 ± 17.4	73.3 ± 17.8	74.7 ± 18.4
	Monitor	72.0 ± 19.1*	72.1 ± 19.1*	72.8 ± 19.7*
TEMP (^oC)	Charts	36.3 ± 0.7	36.4 ± 0.7	36.3 ± 0.7
GCS (%)	Charts			
≤10		1.1	1.3	1.1
11		0.2	0.5	0.6
12		0.2	0.8	0.6
13		0.8	1.6	1.9
14		10.7	8.2	9.5
15		86.9	87.5	86.4

3.5.2 Secondary outcomes overview

Hospital Length-of-Stay, Escalation of Care, Mortality and ED Re-admission

Table 3.6 presents the large-scale study secondary outcomes, with the exception of the frequency and duration of physiological abnormality.

While most outcomes are similar between phase 1 and phase 2, that is not the case between phase 2 and 3. We analyse the following outcome categories:

Hospital length-of-stay: The hospital LOS was not significantly different between phases. Regarding the ED LOS, there a was significant decrease of 18-min between phases 1 and 2, and a significant increase of 12-min between phases 2 and 3. Although they were significant, they may not be clinically relevant.

Escalation of care: significantly more patients started in the resus area (a difference of +1.27%, p-value <0.001) in phase 3, than in phase 2. We note that in our population all patients starting in the resus area were transferred first to the majors area, after their initial treatment, and then moved to the next hospital area (or home, if their were deemed "stable"). Between phases 1 and 2, +5.2% patients where admitted to the hospital after their ED stay, but the difference was not significant between phases 2 and 3. However, the proportion of patients with hospital stays longer than 5 days was higher for phase 3.

Mortality: a higher proportion of patients died within 30 days in phase 3 (difference was +1.01%, p-value was 0.013), but 24 and 48-hour mortality were not significantly different.

Re-admission: No differences were found in the re-admission rate of ED patients between phases 1 and 2. The apparent lower re-admission rate (p-value was 0.053) for phase 3, was also not significantly different from that of phase 2.

Table 3.6: *Secondary outcomes for the large-scale ED study for cohort groups: A_1, A_2, A_3. CPR - Cardiopulmonary resuscitation; LOS - Length-of-Stay; ICU - Intensive Care Unit. [a] Continuous values presented as median (IQR). The Mann-Whitney U test was used to compute the p-values. [b] The p-value was determined using the χ-test of independence for multinomial variables. *p-value < 0.05. [c] Mortality occurring after discharge from the ED-related hospital stay.*

Outcome	A_1	A_2	A_3	p-value[b] A_1 v.s. A_2	p-value[b] A_2 v.s. A_3
Mortality (%)					
24-hour	0.25	0.24	0.26	0.935	0.852
48-hour	0.28	0.30	0.55	0.888	0.109
15-day, post-ED	1.34	1.04	1.51	0.275	0.088
30-day, post-ED	1.49	1.43	2.03	0.842	0.058
15-day[c]	1.80	1.46	1.97	0.276	0.105
30-day[c]	2.36	2.36	3.37	0.991	0.013*
LOS					
ED (hours)[a]	3.80 (3.20, 4.00)	3.50 (2.80, 3.90)	3.70 (3.00, 3.90)	0.001*[a]	0.001*[a]
Hospital (days)[a]	1.00 (0.00, 4.00)	1.00 (0.00, 4.00)	1.00 (0.00, 4.00)	0.665[a]	0.083[a]
Escalation of care (%)					
Started in Resus	3.42	2.63	4.30	0.06	0.001*
Moved to Resus	2.30	2.57	2.29	0.483	0.466
Hosp. admission	61.60	66.80	68.40	0.001*	0.139
$\geq$ 5-day admission	14.07	14.98	16.87	0.299	0.033*
CPR	0.16	0.24	0.32	0.447	0.679
ICU	0.78	0.72	0.49	0.775	0.236
Re-admissions (%)					
2-day	1.20	1.00	1.00	0.52	0.721
5-day	1.50	1.10	0.60	0.242	0.013*
30-day	3.80	4.20	3.20	0.386	0.026*
Total	7.80	8.10	6.80	0.734	0.053
Operational downtime					
VitalPac (%)	-	1.14	1.27	-	-
Visensia (%)	3.60	5.30	2.10	-	-
Total Attendances	3,219	3,352	3,445	-	-

3.6 Discussion

e-T&T intervention

There was a gap of one year between phase 1 and phase 2, and although there were no reported logistical changes within the ED, statistical analysis revealed differences in patient age, triage, and referral (see Table A.1, in the appendices).

The e-T&T intervention did not have any effect on secondary outcomes except for a 5% increase in hospital admission, which could be again due to factors external to the ED logistics.

The benefits of e-T&T in this study when compared with paper charts were reported in Pullinger et al. (2015). Paper charts were found to have an error rate of about 20% in the calculation of the EWS, confirming the result found in Wilson et al. (2012). Errors in e-T&T occurred at a much lower frequency (<2%), caused by user mistake. Examples of user mistake were: (i) using codes to justify missing data, that would influence the EWS score, when other more adequate codes were available - this was noticed as this error was only related with missing temperature data, which is sometimes hard to measure as the temperature probes may not be readily available at the patient's bed -, an error that may be corrected with user training; and (ii) some observations were recorded on patients incorrectly due to errors in registering the patient in the system, as this was a manual process, in which nurses had to match the patient to a list of names coming from the system that registered all patients at ED admission. The latter may be corrected by changing the patient registration process, for example, by scanning the patient wristband first, in case those are available (Bonnici et al., 2016). The data loss was much greater when using paper charts[3], as T&T charts were not available for 33.8% of the collected data, while the data loss in e-T&T was 7.13% and 6.7% for phases 2 and 3.

[3]This is only retrospective data-loss; clinical charts are available for 100% of patients during their ED stay.

Data fusion system intervention

We hypothesise that there was a change in the recruitment of patients from the majors areas in phase 3 as, previously, all patients would be allocated to cubicles/beds, and in phase 3 only those deemed more unstable by senior nurse assessment (rather than MTS) received care in beds. Only the latter benefited from Visensia, and therefore only these data were collected for review. Also the senior nurse assessment increased the ED LOS between phase 2 and 3 by 12 min (p-value < 0.001), as this may be less time-efficient than MTS. However, this difference may not be clinically relevant.

Phase 3 patients were older than those from phase 2, by 5 years (see Table A.1, in Appendices). A higher proportion attended the ED resus area, and died within 30 days of their admission (p-values < 0.001 and 0.013, respectively). Our hypothesis is that this result is influenced by the change in patient flow in the majors area. Patients from the resus area, were moved to the majors area and then moved to the next ward, while patients that were deemed more stable occupied the majors chairs. Therefore, there were more data from patients at risk of deteriorating in the phase 3 data, which may explain the higher 30-day mortality.

3.7 Conclusion

The large-scale observational study analysed the feasibility of using e-T&T and a data-fusion system to alert staff when clinical observation measurements or continuous data were deemed "abnormal". e-T&T was shown to improve EWS calculation, data availability and decrease the risk of retrospective data loss. Also, e-charts can be augmented with data acquired automatically from bedside monitors, showing vital-sign information between patient observations. Although the clinical observations measurement standards were improved, e-T&T had no significant impact on patient clinical outcomes, which remained similar between phase 1 and phase 2.

It was observed that changes in the ED triage process and patient flow between

phases 2 and 3, may prevent a direct comparison between these phases' clinical outcomes. Clinical staff were asked to act according to the ED ward protocol in the case of an e-T&T alert, and to take an extra set of observations in the case of a data-fusion alert.

Chapter 4

Electronic Track-and-Trigger intervention

4.1 Background: Frequency and completeness of observation sets in the ED

Matching the correct frequency of clinical observation to severity of illness is of vital importance to optimise ED resource allocation and early recognition of patient deterioration. Although the assessment of vital signs is a common task for an ED nurse, there is limited information regarding the optimal frequency with which vital signs should be monitored. In principle the frequency of observations should be proportional to the patient acuity, but a few studies have shown that this is not the case in busy EDs.

Alcock et al. (2002) analysed the frequency and completeness of vital-sign measurements on all patients admitted to an UK hospital via the ED in a four-week period. From a total of 739 admitted patients, observation charts were only available for 728. 88% of the patients had an observation set within 15 minutes of arrival and 36% a repeat observations while waiting for admission. Only 52% had temperature, HR, RR and BP recorded. The authors concluded that a standard protocol was needed in the ED to safeguard against the inconsistent observations pattern, so that treatment call be priori-

tised according to the patients' physiological status, with repeat observations required to monitor the response to the intervention.

Armstrong et al. (2008) showed that only 58% of 400 consecutive ED admissions to the majors and resus areas of a district general hospital in the UK had a vital sign recorded within 15 minutes of arrival and only 7% had all vital signs repeated at 60 min. The study found that the poor recording of vital signs was unrelated with staffing levels or number of patients attending the ED, and a significant relationship was found between the failure of vital-sign recording and lower triage categories, the authors stating that this failure was undermining the strategy to detect and manage ill patients.

Johnson et al. (2012) analysed the association between clinical (e.g. the triage category, assigned using the Emergency Severity Index, ESI[1], in this study), social (e.g. ethnicity), and environmental (e.g. ED crowding) factors, and the time between clinical observations in 202 patients. Low patient acuity was found to be the strongest predictor of increased time between clinical observations. Triage category had the greatest impact on the time between successive nurse observations. Although the majority of patients in this study were assigned a triage category 2 or 3 (high and medium risk, respectively), the results were consistent with previous studies where a more acute patient required more resources as evidenced by more frequent vital signs. The authors concluded that the triage category may be a good instrument to guide emergency nurses in determining the frequency of vital sign monitoring required for patients. The most critically ill patients did not meet the inclusion criteria because of their short length of stay and hence the results from this study cannot be generalized to ED patients with high acuity. A similar significant relationship was observed by Miltner et al. (2014), using a dataset of 43,232 patient visits to 94 US EDs (which included patients from high-acuity areas).

Adding to the inconsistent observation pattern and low vital-sign completeness, in the analysis of the pilot study, Wilson et al. (2012) reported the completeness of

[1]ESI is a triage algorithm that ranks ED admissions from 1 to 5, from requiring life-saving intervention to requiring no clinical resources when the patient is stable.

observations in 472 patients with at least 3 observation sets admitted to the high-acuity areas (majors/resus) of the JR ED [2] and showed that although 85.8% of patients had at least one full set of observations, only 34.5% of the observation sets (2,965 observations in total) had a corresponding T&T score (21% of paper-based scores were incorrect).

These results show that current protocols and tools to record and score vital-sign measurements undermine the management of acutely-ill patients in the ED, current practices being based almost entirely on expert opinion and tradition. This motivated the implementation of e-T&T to track patients' vital signs in the ED.

In the UK, standards were published in 2007 (National Institute of Clinical Excellence, 2007) for the frequency and completeness of vital-sign observations in the ED, but a recent study showed that changing hospital protocol is challenging. Hands et al. (2013) analysed electronic charting patterns from all adult inpatient areas (except high-care areas, such as critical care units). Their study found that the hospital protocol was not followed, as the frequency of observations was related with clinical staff habits rather than the EWS of clinical observations sets. Early recognition and timely activation of Medical Emergency Teams were consequently impaired.

One of the benefits of the use of an electronic device to track the patients' vital signs is the fact that it can notify clinical staff of the time to the next observation, based on the previous calculated EWS, helping clinical staff to follow the protocol for the frequency of observations. This chapter analyses the frequency and completeness of clinical observations sets, with the following contributions:

- comparison between paper and electronic T&T charting patterns (frequency and completeness of observation sets) between phase 1 and phase 2 of the large-scale ED study;
- comparison of the compliance of clinical staff with the ED frequency of observations protocol between phases 1 and 2 of the large-scale trial.

[2]Details of this pilot study are given in section 2.4.1.

4.2 Methods

4.2.1 Cohort selection

We aim to compare charting patterns between phases, and so the complete documentation of the clinical observations was required. Therefore this analysis was conducted using subsets C_1 and C_2 of phases 1 and 2, respectively (see Figure 3.4, chapter 3).

4.2.2 Data preprocessing

Physiologically-implausible values were removed from the observational data, using limits from previous literature (Wong, 2011; Hugueny, 2013), displayed in Table 4.1. Inter-arm or postural BP measurements were merged into one observation set (i.e. the same row now has the right and left BP measurement in the case of an inter-arm BP assessment). Observation sets done within less than 5-minutes of a previous set of vital signs were aggregated, i.e. the most updated vital-sign value was kept to compute the EWS. The latter comprised extra measurements that were made to complement a previous one, such as a temperature measurement. Table 4.2 summarises the post-processing merging and aggregation operations, the total vital-sign measurement sets (initial database rows), and the final observation sets (final merged rows). Data completeness was determined from the resulting database.

Table 4.1: *Low and high threshold values to filter implausible vital-sign values.*

Variable (units):	Low threshold	High threshold
Heart Rate (bpm)	30	300
Respiratory Rate (rpm)	3	60
SpO_2 (%)	60	100
Temperature (^oC)	32.0	42.0
Systolic BP (mmHg)	40	300
Diastolic BP (mmHg)	20	200

Table 4.2: *The table shows counts for different groups of observations sets analysed in this chapter: (i) total unprocessed vital-sign measurements recorded in the final database; (ii) final count for processed observation sets, after merging inter-arm and postural BP assessments (the last BP value being considered in the observation set), and aggregating observation sets within 5-minutes of each other (the most updated values being considered for the observation set). * These observation sets were merged by a research nurse, consisting of those recorded in the triage charts, and then copied, and sometimes merged with complementary data in the T&T charts. **Cohorts C_1 and C_2 correspond to patients with complete documentation from phases 1 and 2, respectively. For phase 1, 773 and 81 ED attendances were removed because of incomplete documentation and observation sets with incorrect timestamps, respectively.*

Data category	# rows (# patients)	
	Phase 1	Phase 2
Manually merged rows*	1168(1168)	-
Unprocessed rows	7292 (2984)	7154 (3113)
Merged rows	76 (64)	4 (4)
Aggregated rows	121 (114)	156 (136)
Final processed rows	7,094 (2,984)	6,998 (3,113)
Cohort C**	6,094 (2,130)	6,998 (3,113)

4.2.3 Patient acuity

Patient acuity was defined according to the patients' highest EWS assigned to their observation sets, represented by $EWS_{\max}$. Histograms of the number of patients per $EWS_{\max}$ were analysed for each phase, and stratified in four groups: $EWS_{\max} = 0$, $EWS_{max} = 1$, $EWS_{max} = 2$ and $EWS_{\max} \geq 3$. These correspond to cases in which patients are (i) stable, at (ii) medium risk ($EWS_{\max} \in \{1,2\}$), requiring a senior nurse assessment, and at (iii) high or critical risk, in which case medical review should be considered. These patient severity levels correspond to the ED protocol, defined in Table 3.2, in chapter 3.

4.2.4 Evaluation of the frequency of observation

The following metrics were compared between phases 1 and 2:

1. The frequency of observation for each patient acuity group, defined by the patients' $EWS_{\max}$.

2. The percentage of attendances with at least two observation sets in at least 1 hour

of their ED stay (hourly charting patterns);

3. Compliance of clinical staff with the ED frequency of observation protocol: The ED frequency of observation protocol that was in place during the study is specified in Table 3.2, in chapter 3. This protocol suggests the time to the next observation (TTNO) is due given an EWS calculated from the previous observation. Three EWS groups with different TTNO in the protocol were derived: EWS $= 0$, EWS $\in \{1, 2, 3, 4, 5\}$ and EWS ≥ 6, which have a TTNO of 2 hours, 1 hour and 30-min, respectively. In addition, ED clinical staff were required to repeat an observation within the first hour after the first observation set. It was not possible to determine the cases in which the TTNO changed because of clinical staff "concern" parameter, specified in the protocol as these data were not available. The compliance of with the TTNO protocol was determined using the following rules:

   (a) Compliant observation set (True Positive) - A pair of observations with a TTNO within that defined in the protocol. In addition, if the total EWS of the first observation set was below 6, a second observation set was required to be made within one hour, and within 30-min otherwise.

   (b) Non-compliant observation set (False Negative) - A pair of observations with a TTNO greater than that defined in the ED protocol.

The clinical staff compliance rate (a measure of sensitivity) in following the ED frequency of observation protocol was determined using these rules on all observation sets. The optimal observation interval is illustrated in Figure 4.1. A metric for the false positive rate was not defined, as it was assumed that patients in the ED setting always require intermittent observations, otherwise they are ready to be discharged. The time between the last observation set and the patient discharge was not taken into account for this analysis. Different timings would need to be applied if the patient was discharged home or to another hospital ward.

It is difficult to assess which tolerance period should be allowed in order to evaluate

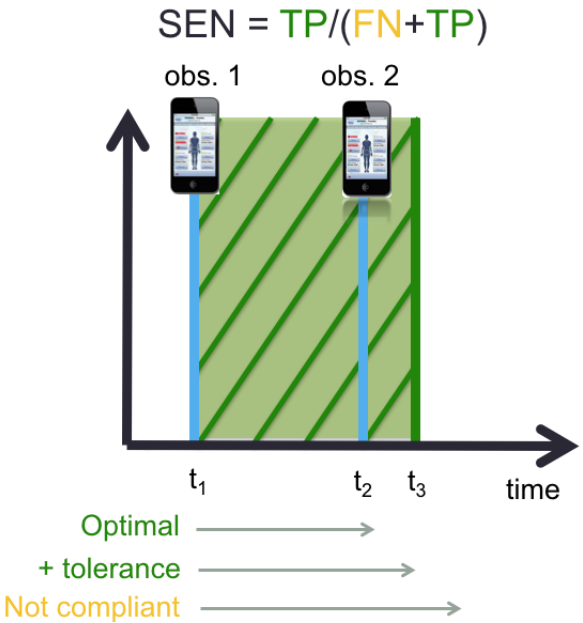


Figure 4.1: *Illustration of the framework to evaluate the compliance of clinical staff with the frequency of observation protocol.* t_1 *- time of first observation (obs. 1);* t_2 *- optimal time for the next observation (obs. 2);* t_3 *- time of the next observation with a tolerance interval, allowing for the fact that some timestamps may be imprecise. SEN - sensitivity, which in this context is a measure of compliance rate.*

the compliance rate; for example, it is possible that if the required TTNO is set to 30 minutes, the nurse will complete the observations set at 30-min + 5-min, or + 10-min, or + 15-min. Hence the total compliance rate is determined here for varying tolerance periods. The curves are presented for the three defined EWS groups with different TTNO in the protocol. As before, patients were allocated to these groups according to their highest EWS, represented by EWS'_{max}, in this case.

4.2.5 Statistical tests

The χ^2-test of independence was used to determine significant differences between groups. Significance corresponded to a p-value of less than 0.05.

4.3 Results

4.3.1 Cohort characteristics & data completeness

In phases 1 and 2, 2,130 and 3,113 patients had complete observation set documentation available, with a total of 6,094, and 6,998 observations sets, respectively. The statistics of patient demographics and outcome data for subsets C_1 and C_2 are equivalent to those of A_1 and A_2, respectively.

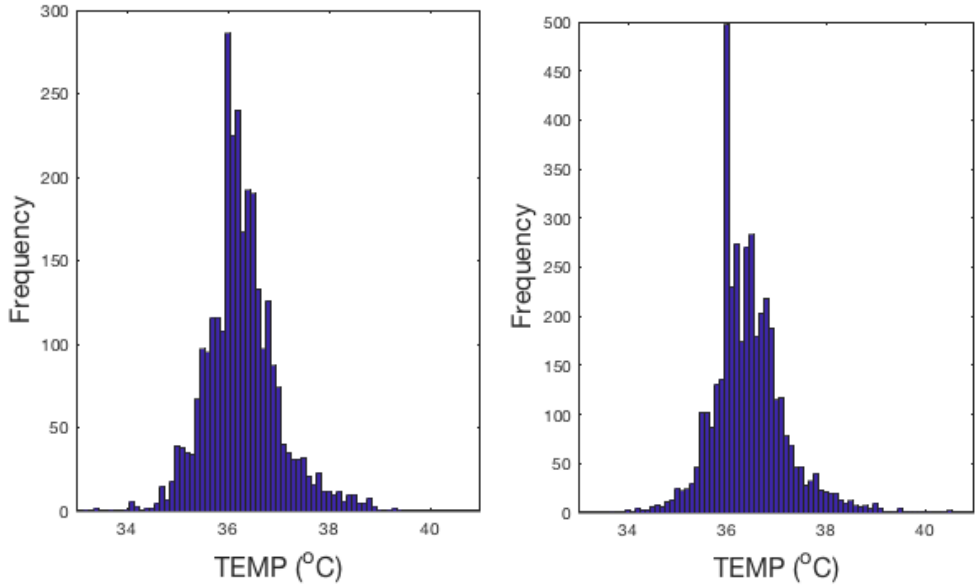
Table 4.3: *Completeness of vital signs, EWS and FiO_2 in T&T charts, for all documented ED admissions from phases 1 and 2 (groups C_1 and C_2, respectively).*

	Observations (patients)	
	C_1 (%)	C_2 (%)
HR, RR, SpO_2, BP	92.24 (98.64)	97.99 (99.71)
HR, RR, SpO_2, BP, GCS	60.93 (83.00)	97.99 (99.71)
HR, RR, SpO_2, BP, GCS, TEMP	34.13 (69.62)	61.53 (86.83)
Recorded EWS	52.13 (76.71)	99.74 (100.00)
FiO_2	76.63 (92.68)	9.42 (10.41)
Total # (%)	6,094 (2,130)	6,998 (3,113)

Table 4.3 shows that the completeness of the observation sets for patients with complete documentation. The EWS completeness for at least one of the patients observation sets was 100% in phase 2 (76.71% in phase 1), and 86.83% of the patients have at least one complete observation set in phase 2 (it was 69.62% in phase 1). A greater proportion of completed GCS and especially temperature can be observed in phase 2. Temperature histograms for each phase are shown in Figure 4.2, and there are clear modes at 36.0 ^oC, in higher number in phase 2. Two hypotheses may contribute to this apparent bias: (i) clinical staff round temperature values to 36.0^oC; (ii) in phase 2, clinical staff, requiring to complete the clinical observation when using the VitalPAC software, recorded a normal temperature value which does not influence the EWS, with a preference for this number.

The completeness of the FiO_2 decreased considerably when electronic charts were introduced. With the paper charts the nurses would record if the patient was room air conditions by writing 21%. With the electronic charts, our data showed that the nurses

only recorded it if a mask with active oxygen therapy was applied to the patient. For our analysis, we assumed no O_2 support was given otherwise, for the data from phase 2.



(a) *Phase 1* **(b)** *Phase 2*

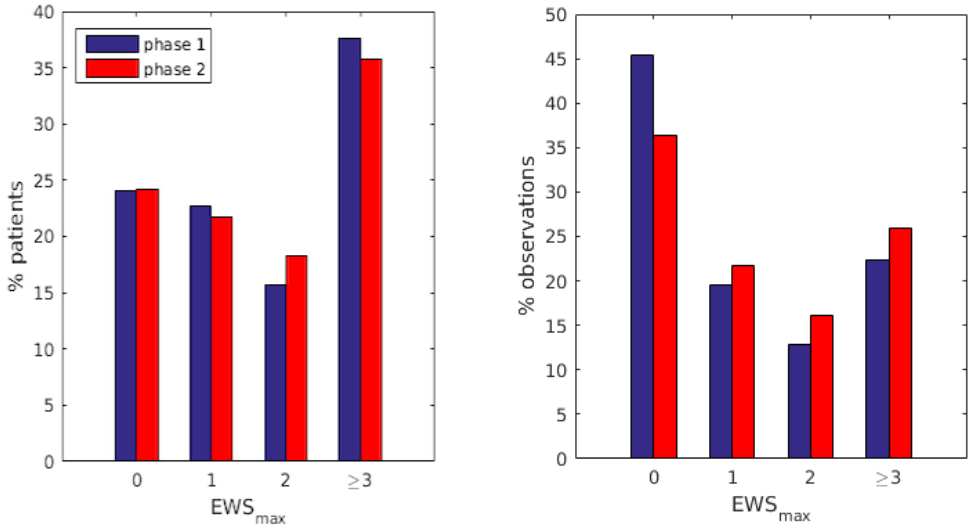
Figure 4.2: *The image shows histograms of temperature data from phase 1 and 2 of the large-scale study. 36 °C is much more frequent in phase 2, as a consequence of "enforced" vital-sign data completeness by the e-T&T system.*

4.3.2 Patient acuity

Figure 4.3 (a) shows the number of ED patients categorised by their EWS_{max} group, phases 1 and 2. The patient acuity (or criticality) was the same between phases (p-value = 0.095), with 24% of the attendances presenting stable physiology ($\text{EWS}_{max} = 0$), for both phases, and 35% and 37%, respectively, requiring medical review ($EWS_{\max} \geq 3$) during their ED stay.

4.3.3 Frequency of observations versus patient acuity

Having demonstrated that patient acuity was the same between phases 1 and 2, we now stratify the number of observations sets by $EWS_{\max}$. Figure 4.3 (b) shows that 9.1% fewer observations were done in stable patients ($EWS_{\max} = 0$) in phase 2, but 3.6% more observations were done in patients with $EWS_{\max} \geq 3$ (p-value < 0.001). This suggests that more observations were made in the more acutely-ill patients when using the e-T&T system, with respect to phase 1.



(a) *Phase 1* **(b)** *Phase 2*

Figure 4.3: *(a) Percentage of patients in each EWS_{max} group, for each phase (p-value = 0.095, for phase 1 v.s. phase 2); (b) Percentage of clinical observations in each EWS_{max} group, for each phase. EWS_{max} - Maximum Early Warning Score during the patient's ED visit.*

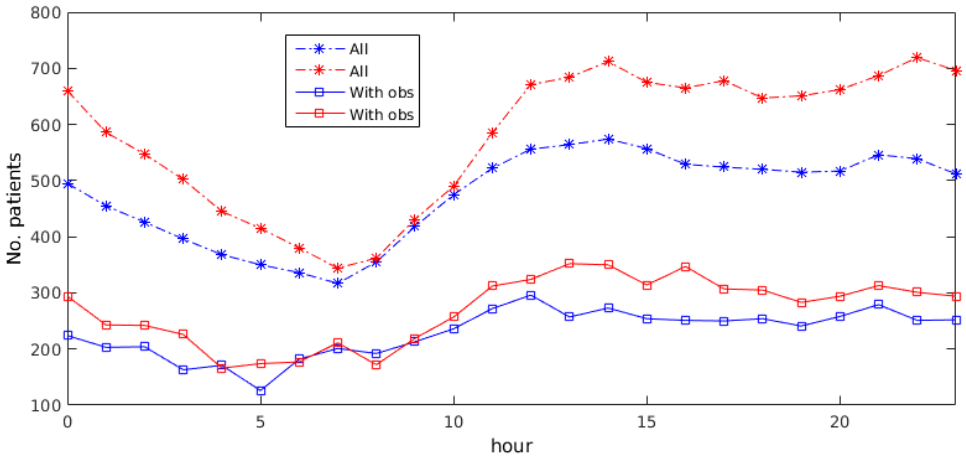
4.3.4 Hourly frequency of observation

Figure 4.4 shows the percentage of ED attendances with observations per hour of the 24-hour cycle, during both phases. About 50% of the patients have at least one observation set per hour. This percentage is higher when there are fewer patients on average, at 7 am, coinciding with the morning clinical staff shift change. There were no clinically relevant differences when these data were analysed per week day.

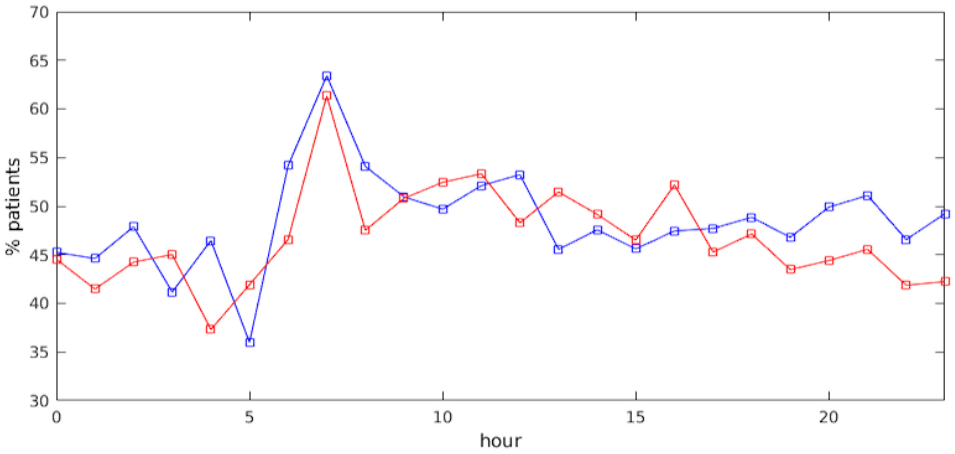
Figure 4.5 shows the percentage of patients with more than one observation set per hour, stratified by their $EWS_{\max}$. These were the patients requiring more attention from the clinical staff. In phase 2 only 10.7% of these were patients with stable physiology, versus 17.8% when using paper charts in phase 1. 55.6% of these patients, in phase 2, had at least one observation set that would consider a medical review ($EWS_{\max} \geq 3$) while that occurred for 49.6% of the patients in phase 1. This was a significant change (p-value = 0.027), and confirms that clinical staff spent more time with patients presenting higher EWS, when guided by e-T&T.

4.3.5 Compliance with frequency of observation protocol

Figures 4.6a and 4.6b show the percentage of patients versus observation pairs, and the compliance rate (%) of clinical staff in following the frequency of observation protocol, for the different $EWS'_{\max}$ groups, in each phase, respectively. Figure 4.6b shows that in general nurses are more compliant in tracking the more critical patients, which can be observed for both phases 1 and 2. Assuming a 15-min tolerance interval, the compliance rate in phases 1 and 2 for critical risk patients ($EWS'_{\max} \geq 6$) was 72% and 70% while it was lower for the low risk patients ($EWS'_{\max} = 0$), about 66% and 63%, and much lower for medium/high risk patients ($EWS'_{\max} \in \{1,2,3,4,5\}$), about 60% for both cases, respectively. As expected for a busy ward, staff are mostly occupied with the critical patients, at the cost of a lower compliance rate for non-critical patients.



(a)



(b)

Figure 4.4: *Blue and red lines correspond to phase 1 and 2 data, respectively. (a) Frequency of majors patients (asterisk markers), and majors patients with observations (squared markers) attending the ED per hour, for the 24-hour cycle; (b) Percentage of patients with at least 1 observation per hour, for the 24-hour cycle.*

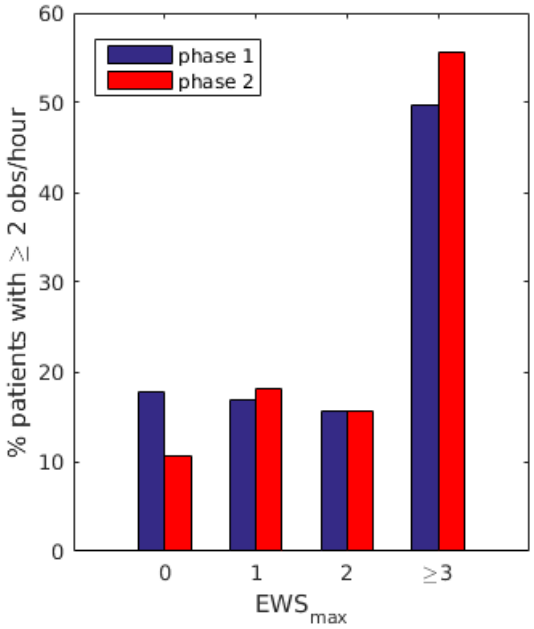


Figure 4.5: *Percentage of patients with at least two clinical observations per hour, stratified by their EWS_{max} group, p-value = 0.027 for phase 1 v.s. phase 2.*

4.4 Discussion

This chapter has compared vital-sign charting patterns, characterised by the frequency and completeness of clinical observations, between paper and electronic T&T charting in the ED.

The data completeness improved with e-T&T, with respect to paper charts, as 100% of the patients, with complete documentation, had at least one EWS recorded (improved from 76.71% in phase 1), and 86.83% had at least one observation with a complete set of vital signs in phase 2 (improved from 69.62% in phase 1). e-T&T doubled the observation completion rate when compared with paper charts (34.12% versus 61.53%, respectively), potentially at the cost of temperature data being imputed artificially, either by holding the initial value, or by making sure that it would not influence the EWS score. This may require more analysis in future work.

It was demonstrated that between phases 1 and 2 (e-T&T intervention) a higher proportion of observations (+ 9.1%) were conducted in the unwell or at risk group of patients ($EWS_{\max} \geq 0$), when clinical staff was guided by the electronic devices, as patient

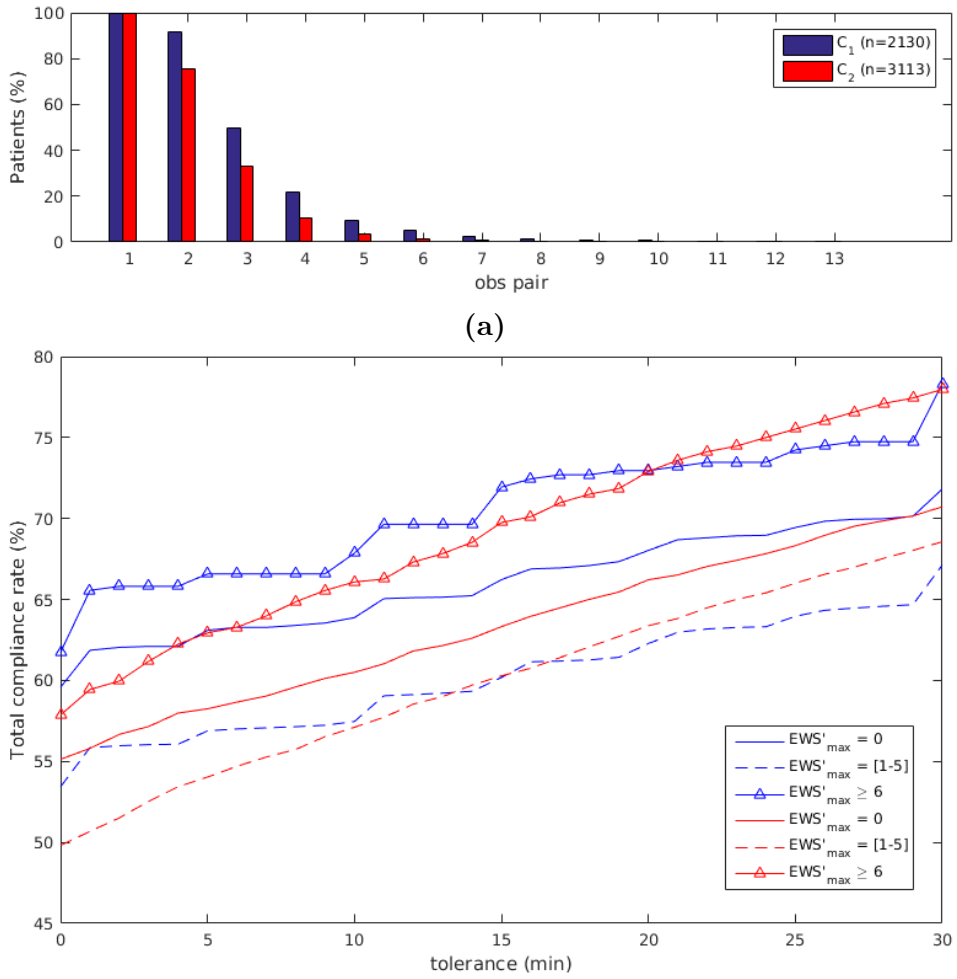


Figure 4.6: *(a) Percentage of patients by observation number, for each phase; (b) Compliance rate (%) of clinical staff following the ED frequency of observations and escalation protocol for each tolerance interval. Blue and red correspond to phases 1 and 2, respectively. The solid line, dashed line and the triangular marker line correspond to each* EWS'_{max} *group,* $EWS'_{max} = 0$*,* $EWS'_{max} \in \{1,2,3,4,5\}$*, and* $EWS'_{max} \geq 6$*, respectively.*

criticality was the same. The electronic devices accurately determine the EWS and the corresponding time to the next observation, and help avoid most of the data recording and scoring mistakes made by clinical staff when using paper-charts (as discussed in chapter 3).

The "clinical concern" criteria that set the frequency of observations to 1-hour when the EWS is 0, 1 or 2 were not modelled, these are not recorded on our study database. These criteria include nurse judgement - which may be subjective -, or specific events such as "Blood loss/melena", "New onset of vomiting/diarrhoea" and "new onset of pain", to name three examples. Although more observations were made in patients requiring a medical review ($EWS_{\max} \geq 3$), nurse compliance with the observation interval protocol (TTNO) did not change for critical patients when using e-T&T. Nurses were less compliant with this protocol in observing the non-critical patients in phase 2.

4.5 Conclusion

In this chapter it was possible to show that when nurses were guided by electronic T&T charts, the data completeness was much higher, and more observations were done on patients with higher acuity, the nurses spending more time with these patients. As the ED is a busy ward, the consequence was a decreased compliance in the frequency of observation protocol for non-critical patients.

Chapter 5

Optimising the Early Warning Scores for the ED

In this chapter we make use of the ED data collected from our studies to:

- investigate the design and evaluation of different EWS approaches, with the objective of optimising the score for use in the ED;
- investigate the addition of clinical parameters such as FiO_2 and patient demographic information (e.g. age) to improve EWS performance in the ED.

5.1 Design of existing EWS systems

An example of a typical multi-parameter EWS system is shown in Table 5.1. Scores are allocated to vital-sign ranges in a weighted manner, with the most abnormal vital signs being scored higher. Seven scoring bands are used: 3, 2, 1, 0, 1, 2 and 3, reflecting abnormally low values (EWS = 3) to normal values (EWS = 0) to abnormally high values (EWS = 3). The sum of the scores allocated to a given set of vital-sign measurements results in the overall score. Overall scores above a pre-determined threshold are used to escalate care, e.g. increase the frequency of observations or call a rapid response team, or more experienced staff. Protocols also usually require an alert to be generated when

a single parameter reaches its maximum score (EWS = 3).

Table 5.1: *The National EWS system, NEWS, as reported in Williams et al. (2012).*

	The National Early Warning Score						
Variable	3	2	1	0	1	2	3
HR	≤ 40		41 - 50	51 - 90	91 - 110	111 - 130	≥ 131
RR	≤ 8		9 - 11	12 - 20		21 - 24	≥ 25
SpO_2	≤ 91	85 - 90	91 - 93	≥ 94			
SBP	≤ 90	91 - 100	101 - 110	111 - 219			≥ 220
TEMP	≤ 35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	$\geq$39.1	
AVPU				A			V, P, U

Table 5.2, adapted from Pimentel (2015), shows the parameters present in 25 multi-parameter EWS used clinically and reported in the literature (Gao et al., 2007; Smith et al., 2008a,b; Prytherch et al., 2010; Tarassenko et al., 2011). The majority comprises the vital signs usually observed in the ED ward (see chapter 2). A few use a score to reflect the use of oxygen therapy and a couple of systems use age as a parameter. In this thesis we focus on EWS systems designed from statistical analysis of the vital-sign data, rather than those for which the thresholds for the individual scores have been chosen heuristically by clinical experts. Four data-driven EWS systems examples can be summarised as follows:

ViEWS was developed by analysing a vital-sign data set with patient outcomes and determining thresholds for individual scores that yielded the highest AUROC value for the detection of in-hospital mortality within 24 hours of an observation set. This system was then compared with another 33 EWS systems using a large data set (198,755 observation sets from 35,585 patients), showing that ViEWS had the highest AUROC (Prytherch et al., 2010). In this system, aside from the six traditional vital signs, a score of 3 is added when patients are receiving oxygen therapy (through an oxygen mask).

NEWS (Smith et al., 2013), displayed in Table 5.1, is a modification of ViEWS, with the presence of oxygen therapy being given a score of 2 and very high values of systolic BP being given a score of 3 (note there was no alert threshold for high SBP in ViEWS). These changes led to NEWS having the highest AUROC (0.873), when compared with the 33 EWS systems mentioned above, which had AUROC values in the

Table 5.2: *Twenty-five aggregate-weighted track-and-trigger systems identified and their physiological components (marked with ✓), adapted from Pimentel (2015).*

Year	Study	Heart Rate	Resp. Rate	O_2 Saturation	Temperature	Systolic BP	FiO_2 Support	Age	Consciousness
2000	Wright et al. (2000)	✓	✓		✓	✓			✓
2001	Subbe et al. (2001)(1)	✓	✓		✓	✓		✓	✓
2001	Subbe et al. (2001)(2)	✓	✓		✓	✓			✓
2001	Riley and Faleiro (2001)	✓	✓		✓	✓			✓
2001	Cooper (2001)	✓	✓		✓	✓			✓
2003	Subbe et al. (2003)	✓	✓		✓	✓			✓
2004	Rees and Mann (2004)	✓	✓			✓			✓
2004	Allen (2004)	✓	✓		✓	✓			✓
2005	Goldhill et al. (2005)	✓	✓	✓	✓	✓			✓
2005	Chatterjee et al. (2005)	✓	✓	✓	✓	✓			✓
2005	Andrews and Waterman (2005)	✓	✓		✓	✓			✓
2006	Paterson et al. (2006)	✓	✓	✓	✓	✓			✓
2006	Smith et al. (2006)	✓	✓		✓	✓			✓
2006	Lam et al. (2006)	✓	✓		✓	✓			✓
2006	Gardner-Thorpe et al. (2006)	✓	✓		✓	✓			✓
2007	Subbe et al. (2007)	✓	✓			✓		✓	✓
2007	Hancock and Durham (2007)	✓	✓		✓	✓			✓
2007	Duckitt et al. (2007)	✓	✓	✓	✓	✓			✓
2007	Lilienfeld-Toal et al. (2007)(1)	✓	✓	✓	✓	✓			✓
2007	Lilienfeld-Toal et al. (2007)(2)	✓	✓	✓		✓	✓		✓
2008	MEWS(Oxford-based)	✓	✓	✓	✓	✓			✓
2010	Prytherch et al. (2010) (ViEWS)	✓	✓	✓	✓	✓	✓		✓
2011	Tarassenko et al. (2011) (CEWS)	✓	✓	✓	✓	✓			✓
2012	Williams et al. (2012) (NEWS)	✓	✓	✓	✓	✓	✓		✓
2014	Badriyah et al. (2014) (DT-EWS)	✓	✓	✓	✓	✓	✓		✓

range 0.736 to 0.834.

DT-EWS (Badriyah et al., 2014), was generated using a decision tree algorithm, in which each EWS score was assigned according to how the outcome risk in each terminal leaf would compare to the mean outcome risk. For example, if the risk of the outcome is between 1 and 2 times the mean risk, a score of 1 is assigned to the vital-sign range in that terminal leaf; a score of 2 is assigned if the risk is between 2 and 3 times the mean risk and a score of 3 for higher outcome risks. This design showed similar performance to NEWS (which, by contrast, was developed by trial and error), in predicting adverse events (mortality, cardiac arrest and ICU admission) within 24 hours of an observation set. One criticism of the methodology used to develop DT-EWS and ViEWS is the fact that although they benefit from a large sample size, no held-out independent data was kept back to validate the system.

Finally, **CEWS** (Tarassenko et al., 2011) proposes the use of a centile-based EWS, in which the scores are assigned according to the distribution of the vital signs. Figure 5.1 shows an example, in which scores 3, 2 and 1 are assigned to the 1%, 5% and 10% tails, respectively, of the Cumulative Distribution Function (CDF) of HR. The CDFs were determined from 64,622 h of vital-sign data, acquired from 863 acutely-ill in-hospital patients, both from surgical and medical wards, using bedside monitors. As the system was not "validated" in wards at the time, Subbe (2011) commented that although the use of centiles in this way could have merit, the difficulty in calibrating EWS systems is that they require validation in an intervention study. CEWS is currently being validated in hospital wards of the John Radcliffe Hospital, Oxford (its use in the acute wards of the ED is described in chapters 3 and 4).

In this chapter the probabilistic approach used for the CEWS is followed to model the vital-sign scores, and the addition of age is modelled. Prytherch et al. (2010) noted that its addition adds further complexity, can convey little benefit for certain hospital wards, and has the potential to raise ethical issues. On the other hand, in specific studies such as in Subbe et al. (2001), it was shown that a version of the MEWS system improved

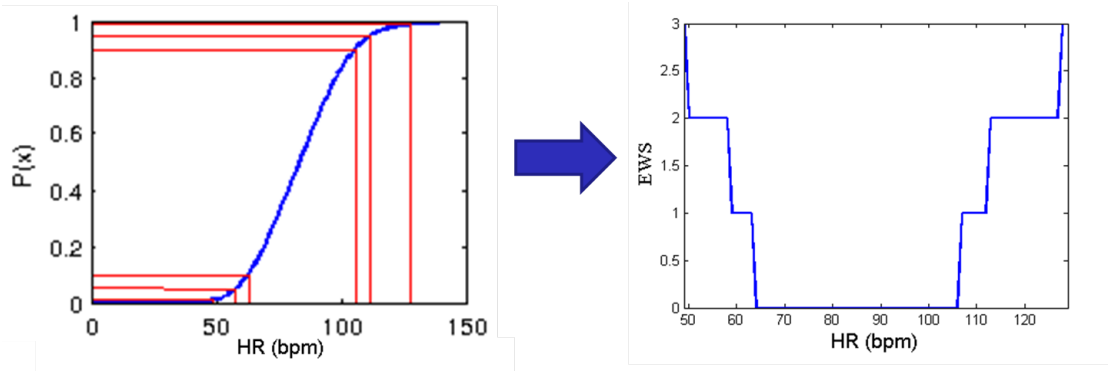


Figure 5.1: *EWS system developed using a data-driven approach (Tarassenko et al., 2011). (left) the cumulative distribution function (CDF) of vital-sign data representative of the patient population being monitored is shown in blue, with scoring centiles shown in red; (right) scores over the HR range.*

its AUROC from 0.67 to 0.72 in predicting mortality, and ICU/HDU unscheduled admission in patients from the Medical Assessment Unit ward, when adding a score of 2 for patients who were 50 years old and above, and a score of 3 for those patients who were 70 years old or above. However, following this rationale, we hypothesise that an elderly patient on the ED ward would often alert, and so we consider an alternative approach, in which the vital-signs CDFs, used to model physiological abnormality, are conditioned on the patient's age.

5.1.1 Observation-wise evaluation of EWS

The goal of any EWS is to alert clinical staff to the need for clinical escalation of a patient in physiological distress (positive class) and not alert when the patient physiology is normal (negative class). If this decision is made for each clinical observation set (i.e. observation-wise), then an alerting or non-alerting observation set occurring within a pre-specified time t prior to a clinical escalation (positive outcome) is classified as being a True Positive (TP) or a False Negative (FN), respectively. Conversely, an alerting or non-alerting clinical observation set without a subsequent clinical escalation within time t is classified as being a False Positive (FP) or a True Negative (TN), respectively.

At every possible EWS threshold value, observations can be allocated, correctly

or incorrectly, to each class as specified before. A confusion matrix is computed from these results and additional discriminative metrics can be evaluated: the Sensitivity, also called True Positive Rate (TPR) or Recall, and the Specificity (SPEC), also called True Negative Rate (TNR), which are defined as the proportion of correctly identified cases within all actual positive and negative cases, respectively; the Positive Predictive Value (also called Precision), and the Negative Predictive Value (NPV), which correspond to the proportions of correctly identified cases within all classified positive and negative cases, respectively; and finally the Accuracy (ACC), which is defined as the proportion of cases that are correctly classified within all cases. These discriminative metrics are summarised in Table 5.3.

Table 5.3: *Confusion matrix. TPR - True Positive Rate; SEN - Sensitivity; FPR - False Positive Rate; PPV - Positive Predictive Value; NPV - Negative Predictive Value; SPEC - Specificity; ACC - Accuracy.*

		Predicted			
		Positive	**Negative**		
Actual	**Positive**	TP	FN	**TPR/SEN** $\frac{TP}{TP+FN}$	
	Negative	FP	TN	**FPR** $\frac{FP}{FP+TN}$	**SPEC** $\frac{TN}{TN+FP}$
		PPV/Precision $\frac{TP}{TP+FP}$	**NPV** $\frac{TN}{TN+FN}$	**ACC** $\frac{TP+TN}{TP+TN+FP+FN}$	

The AUROC metric is often used to identify the model with best discriminating power for binary clinical outcomes. The receiving-operator-characteristic (ROC) curve is plotted as SEN versus [1 - SPEC], or alternatively TPR versus FPR , over a range of values for the alerting threshold. The AUROC is the area under this curve. The optimal operating point (threshold) may be found so that the classifier gives the best trade-off between the cost of failing to detect positives against the cost of raising false alarms. This "optimal" threshold value corresponds to the point on the ROC curve that is closest to the point (0, 1).

5.2 Methods

5.2.1 Cohort selection

Annotation of the clinical outcomes by experts is one of the most important steps in patient data analysis projects, allowing the reliable identification of the patterns we wish to model. However, this process would be time consuming for a database as large as the one described in chapter 3. Hence, in a first approach, we made use of the outcome data collected during the trial by two experienced research nurses, and identified those adverse events that matched the definition of escalation in the ED pilot-study (see section 2.4.1.2, chapter 2).

One type of adverse event was identified as a consequence of physiological deterioration occurring during the patient's ED stay: escalation from the majors to the resuscitation area after ED arrival. We defined "event" versus "no-event" patients as those with and without an escalation to the resus area during their stay in the ED, respectively. Other types of escalations (Table 3.6, chapter 3) were not considered because they were not related with physiological deterioration during the ED stay (e.g. ED re-admissions or death after the ED stay).

Patients without observational data before the escalation event, and with less than 15 minutes of continuous data in at least 3 vital-sign channels, were excluded. In this way we focus on those patients that required more attention, and the use of continuous vital-sign data to diagnose their condition.

The consort diagrams in Figures 5.2a and 5.2b show the final number of patients with and without an adverse event selected in training and in the test data. We use the nomenclature $E_{\{\text{phases}\},\{outcomes\}}$, to represent different sets of patients from the total cohort. Patients from phases 1 and 2 of the large-scale study are chosen as training data, with the corresponding patient indices coded as $E_{\{1,2\}}$. Patients from phase 3 of the study, are chosen as test data with patient indices coded as E_3. The patient outcome group is represented by the second sub-index. Event patients from phase 3 are code with

sub-index $E_{3,e}$. The no-event patients are further divided into those discharged home, and those admitted to another hospital ward. These groups are coded with sub-indices d and a, e.g. $E_{\{1,2\},\{d,a\}}$[1] corresponds to the indices of no-event patients in phases 1 and 2 discharged home and admitted to a hospital ward after their ED stay.

Figures 5.2c and 5.2d compare (i) the normalised histograms of the length-of-stay for no-event (in blue) and event patients (in red), and the normalised histogram of the time from arrival to the escalation event (in magenta), for the training and the test datasets, respectively. The median time from patient arrival to the escalation event was 1.7 [1.0, 2.7] hours, and 2.1 [1.2, 3.2] hours, for the training and test datasets, respectively (the median LOS was between 4 and 5 hours, for the no-event and event patients in both the training and test datasets).

5.2.2 Data pre-processing

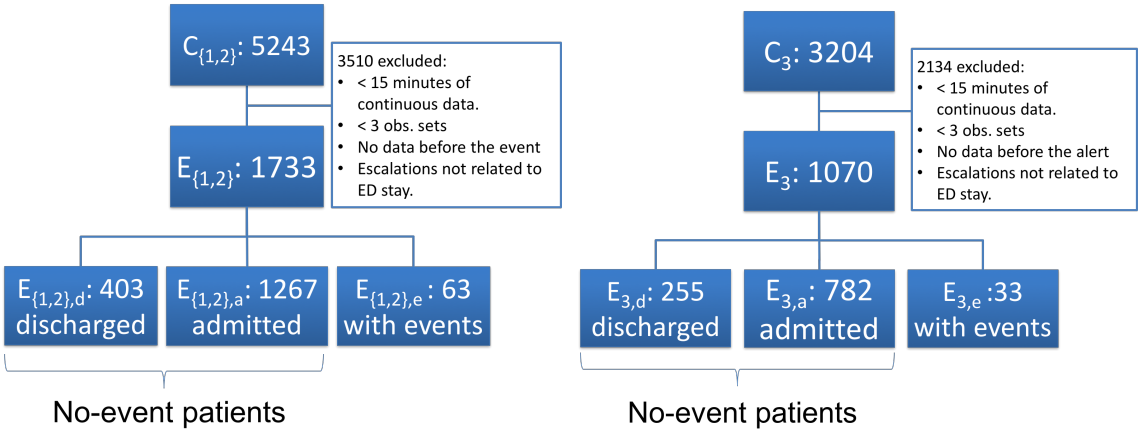
The observational data was pre-processed as described in section 4.2.2. We followed the design of recent EWS systems (Prytherch et al., 2010) and modelled the presence or absence of an oxygen mask, by coding all values reflecting active oxygen therapy as 1, and 0 otherwise.

5.2.3 Univariate analysis

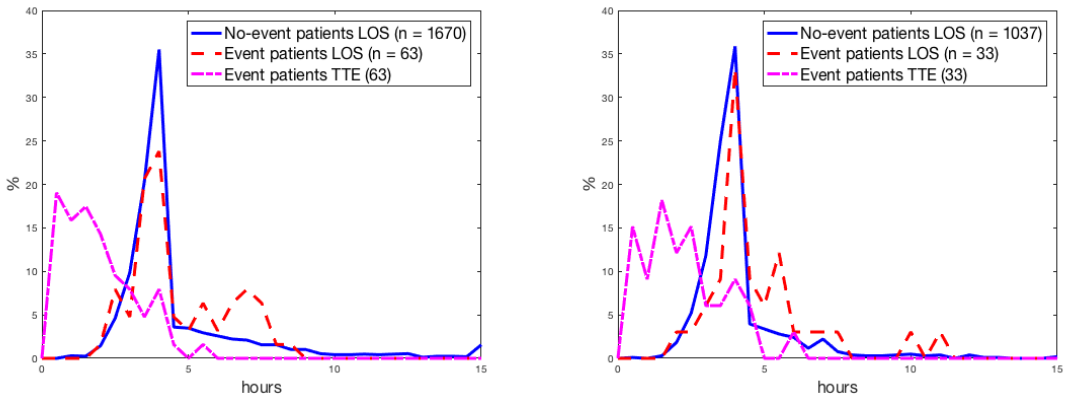
Table 5.4 shows the difference between the vital-sign values of intermittent observations sets recorded by ED nursing staff, and the demographic data (age and sex) between patients with and without events. No difference was found in gender between patient groups, but there is a difference in the age, which increases with greater need for care (median age was 51, 66 and 70 years, for discharged, admitted, and "event" patients, respectively).

Figure 5.3 shows the cumulative density functions for the three patient groups.

[1]To summarise the notation, all outcomes are considered when the second sub-index is omitted, e.g. $E_{\{1,2,3\}}$ is the same as $E_{\{1,2,3\},\{d,a,e\}}$, and E_3 is the same as $E_{3,\{d,a,e\}}$.



(a) *Training dataset.* **(b)** *Test dataset.*



(c) *Training dataset LOS and TTE.* **(d)** *Test dataset LOS and TTE.*

Figure 5.2: *Consort diagram for patients with and without clinical events. (a) Training dataset, comprising data from phases 1 and 2, coded as* $E_{\{1,2\}}$ *(*E_1 *and* E_2 *refer to the patient indices from phases 1 and 2, respectively); (b) test dataset, comprising data from phase 3, coded as* E_3*. Sub-indices d, a and e, represent no-event patients discharged home, no-event patients admitted to a hospital ward, and event patients, respectively. (c) and (d) compare the normalised histogram of the length of ED stay (LOS) for no-event patients and for event-patients, and the normalised histogram of the time to the first event (coded as time-to-event, or TTE), between the training and the test data sets, respectively.*

Table 5.4: *Mean (standard deviation) of the observational data set included in the analysis from 2,803 episodes (all patients included in our analysis, see consort in Figure 5.2): "No-event" data include data from patients without events, discharged from, or admitted to the hospital ($E_{\{1,2,3\},\{d,a\}}$); and "Events" data include data from patients with adverse events occurring after ED arrival ($E_{\{1,2,3\},e}$). [a] Values presented as median [25^{th}, 75^{th}] quantiles. [b] Values presented as %, p-value computed using χ^2-test of independence. *p-value $\leq$ 0.05 (computed using the Kolmogorov-Smirnov test between "event" and "no-event" patients). **Values for all patients without events, i.e. comprising both admitted and discharged ED patients. We summarise the notation so that $E_{\{:\},d}$ represents $E_{\{1,2,3\},d}$.*

	No events			Events	p-value
	Discharged	Admitted	All**		No-events v.s. Events
	$E_{\{:\},d}$	$E_{\{:\},a}$	$E_{\{:\},\{d,a\}}$**	$E_{\{:\},e}$	$E_{\{:\},\{d,a\}}$ v.s. $E_{\{:\},e}$
HR	78 (17)	81 (18)	81 (18)	90 (23)	0.001*
RR	17 (4)	18 (4)	18 (4)	20 (6)	$<$ 0.001*
SpO_2^a	98 (96-100)	98 (96-99)	98 (96-99)	98 (95-100)	$<$ 0.001*
SBP	134 (26)	137 (29)	136 (28)	128 (33)	0.003*
TEMP	36.3 (0.7)	36.4 (0.8)	36.4 (0.8)	36.4 (0.9)	$<$ 0.001*
FiO_2 [b]	2%	9%	7%	44%	$<$ 0.001*
GCS$\leq$11	1%	1%	1%	7%	
GCS 12	1%	1%	1%	1%	
GCS 13	2%	1%	1%	3%	$<$ 0.001*
GCS 14	7%	11%	10%	16%	
GCS 15	90%	86%	87%	74%	
Age[a]	51 (32-70)	66 (44-82)	62 (40-80)	70 (42-84)	0.017*
Male[b]	51%	48%	49%	50%	0.790

The vital-sign distributions for the patients with and without events are significantly different. We observe that patients with events have higher mean heart rates (87 v.s. 81 bpm), higher mean respiratory rates (20 v.s. 18 rpm), lower mean SBP (129 v.s. 136 mmHg). 44(%) of event patients are given oxygen therapy, in comparison to only 6% of the patients without events. Patients with events with low SpO_2 will often be put on oxygen therapy and hence have high SpO_2 values as a result of this intervention, which may explain the similar SpO_2 median values for these and "no-event" patients. Therefore we need to consider the use of oxygen mask to model abnormal physiology, as in many cases this makes the SpO_2 return to normal values.

The similar mean and median temperatures are explained by the high use of the value 36^oC to complete data in the electronic charts (as discussed in section 4.2.2, chapter 4). However, the Kolmogorov-Smirnov test reports a significant difference[2], the two groups having different temperature distributions (see Figure 5.3, TEMP).

Given the characteristics described above we used all the vital-sign data from the patients in the training set without events ($\#E_{\{1,2\},d} + \#E_{\{1,2\},a} = 403 + 1267$ ED admissions) to "build" EWS models of normal physiology.

[2]As it is based on the maximum difference between the CDF's.

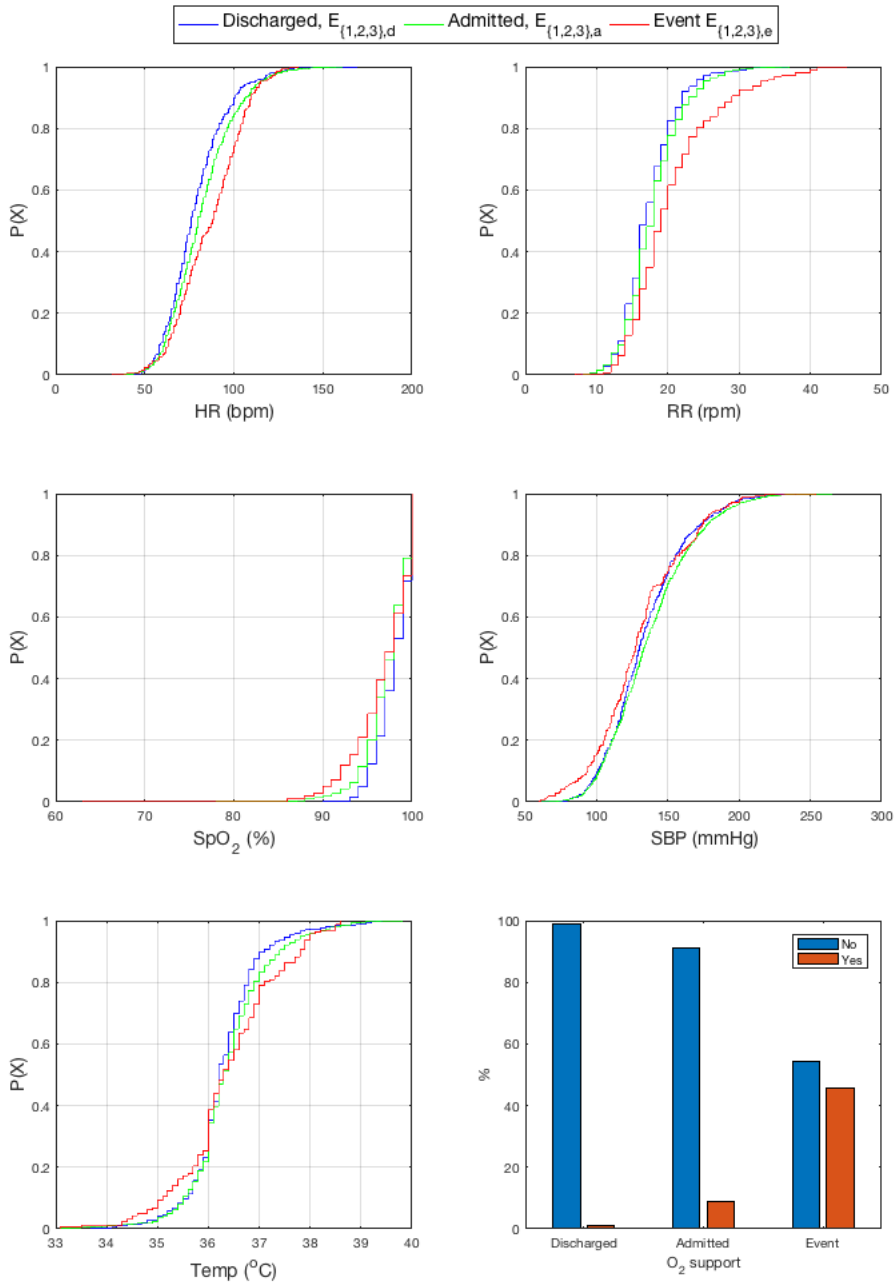


Figure 5.3: *Vital-sign CDFs for "no-events" (discharged or admitted) and "event" patients. There are significant differences between these two patient groups.* $E_{\{1,2,3\},d}$ *- no-event patients discharged after ED stay;* $E_{\{1,2,3\},a}$ *- no-even patients admitted after ED stay;* $E_{\{1,2,3\},e}$ *- events patients.*

5.2.4 Modelling physiological ageing

Figure 5.4 shows the pairwise correlations between some of the vital signs (HR, RR, SpO_2 and SBP), and between the latter and age, of the observation sets of the no-event patients in the training set. The Spearman's correlation coefficient (Altman and Bland, 1983), given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{5.1}$$

where $d_i = r_i - s_i$ is the difference between the ranked variables r and s, and n is the number of samples, is used to evaluate their pairwise correlation (measuring the statistical dependence between two variables)[3]. Values of 1 or -1 correspond to variables fully correlated with each other, and 0 implies that there is no correlation between the variables. Coefficients highlighted in red indicate which pairs of variables have correlations significantly different from zero[4]. The results indicate that age has a significant monotonic correlation with RR, SpO_2 and SBP.

We further analyse the influence of gender on the vital-sign values. Figure 5.5 shows the mean HR and SBP (observational data) conditioned on age estimated by kernel regression (Nadaraya, 1964):

$$\hat{\mathbf{y}} = \frac{\sum_{i=1}^{N} y_i K_\sigma(\mathbf{x} - \mathbf{x}_i)}{\sum_{j=1}^{N} K_\sigma(\mathbf{x} - \mathbf{x}_j)} \tag{5.2}$$

using a Gaussian kernel[5], also known as a Radial Basis Function (RBF), defined as:

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

[3]As the age distribution is not Gaussian, we opted to use a non-parametric correlation measure.

[4]p-values were calculated by transforming the correlation to create a t-statistic with n - 2 degrees of freedom.

[5]A kernel is a weighting function; we define it more formally in section 6.3.3, in chapter 6.

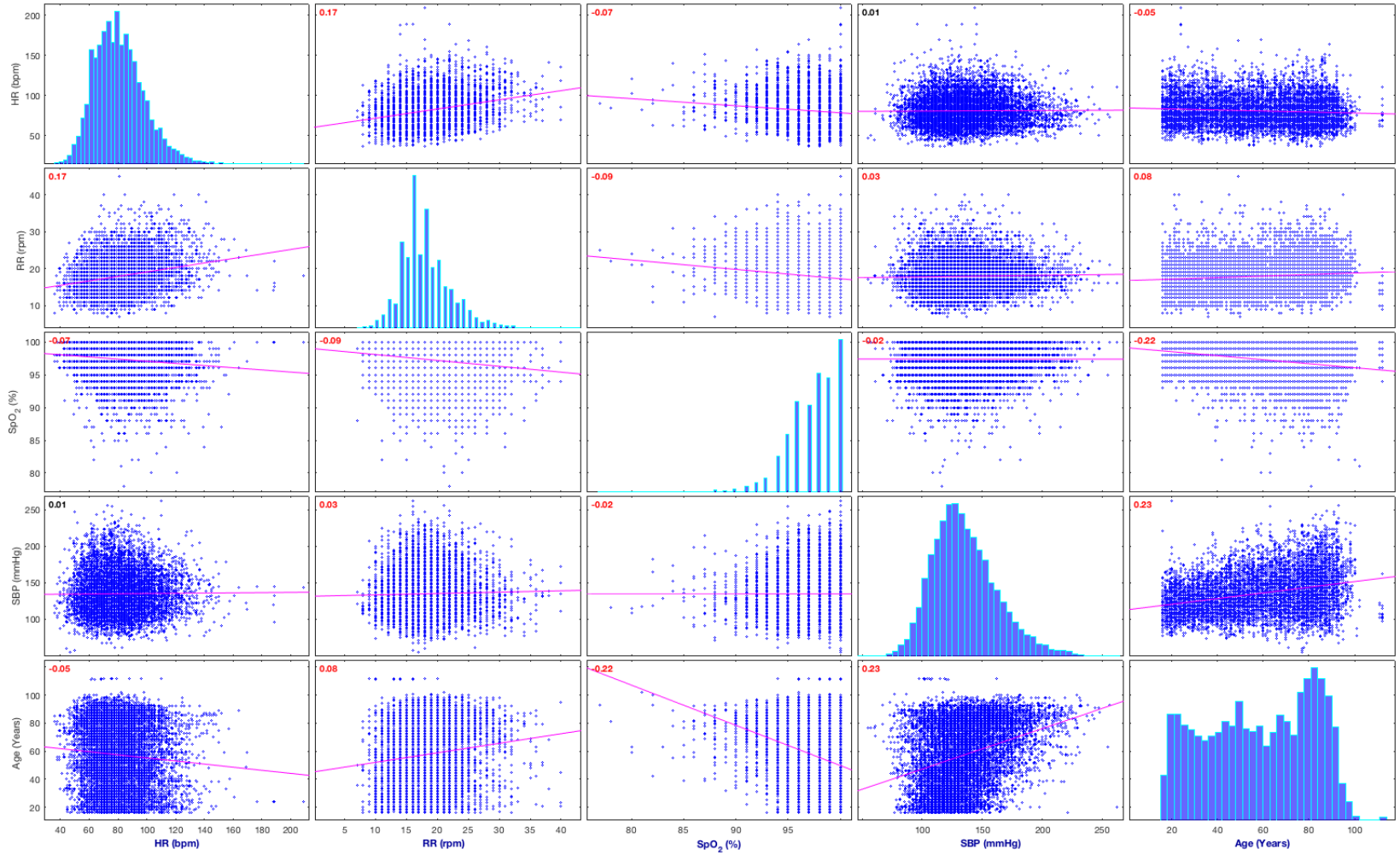


Figure 5.4: *Correlation matrix between vital signs and the age of no-event patients, in the training dataset ($E_{\{1,2\},\{d,a\}}$). The pairwise Spearman's correlations are shown. Significant correlations ($\rho \neq 0$, p-value < 0.05) are highlighted in red.*

where the dimension $D = 1$, in this case. The kernel bandwidth σ was estimated using the rule of thumb given in Murphy (2012), in which, if $\mathbf{x}$ and $\mathbf{y}$ are age and one of the vital signs, respectively, then $\sigma_x = (4/3n)^{1/5}\hat{\sigma}_x$, and $\hat{\sigma}_x = 1/0.6745 * median(|\mathbf{x} - median(\mathbf{x})|)$; σ_y is estimated using the same approach and then the bandwidth for the isotropic Gaussian kernel is given by $\sigma = \sqrt{\sigma_x \sigma_y}$.

It can be observed that women had a consistently higher mean HR than men (+ 5 bpm, may not have clinical significance). The increase in their SBP is non-linear, and emphasised at around age 40 (Maas and Franke, 2009)[6], while in males, in contrast, the increase is more linear. This difference in BP was observed also in vital-sign data from patients monitored in step-down wards in the US (Hann, 2008). As the distributions of vital signs given age and gender are similar, in this chapter we will focus on the introduction of age in EWS systems for the ED, but our methods could easily be extended using the gender covariate for specific EWS parameters.

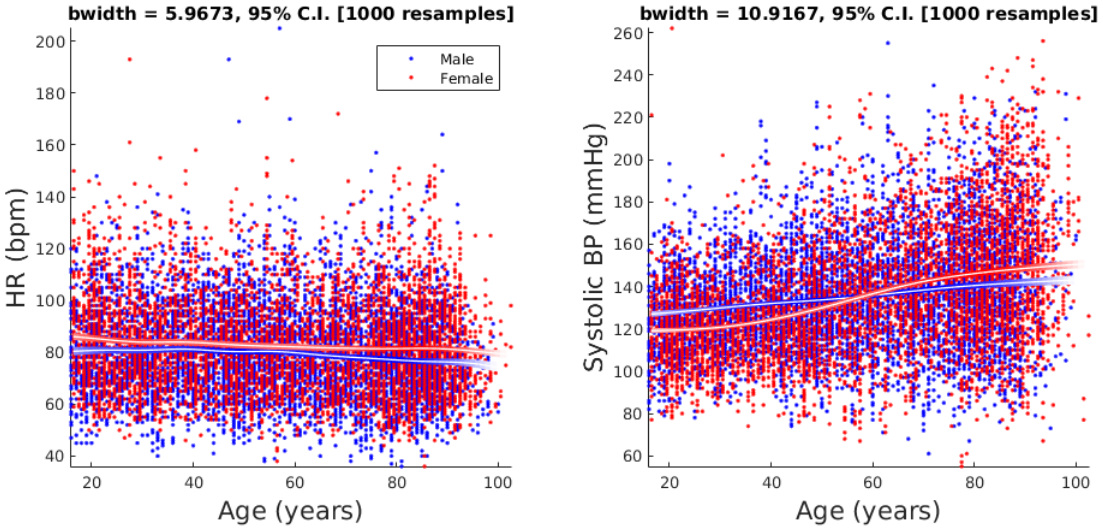


Figure 5.5: *HR and SBP versus age, grouped by the patients sex. The estimated values result from kernel regression (Nadaraya, 1964), using a Gaussian kernel, with bandwidths indicated at the top of each figure. Confidence intervals were computed by bootstrapping (n = 1,000). The mean and the 95% CI are indicated by colour weights using the implementation from Hsiang (2013).*

[6]Its rise with ageing is mainly caused by an increase in vascular stiffness of the great arteries in combination with atherosclerotic changes in the vessel wall. The rise in women older than 40 may be related to the hormonal changes during the menopause.

5.2.5 CEWS models

The following CEWS system modifications were analysed:

(A) **CEWS$_A$**: CEWS is the benchmark EWS system original model described in Tarassenko et al. (2011). CEWS$_A$ represents a version of CEWS in which FiO_2 support is taken into account, adding 2 to the EWS, when present.

(B) **ED-CEWS$_B$**: using the CEWS approach, the EWS cut-off values are obtained from the vital-sign CDFs computed from a sample of the ED training data, which comprises data from patients without events, whose vital signs we consider to be closer to the "normal" physiology of an ED patient (group $E_{\{1,2\},\{d,a\}}$). The lower EWS cut-off values, $\{1, 2, 3\}$, are set to the vital-sign integer values at, or below the $\{10^{th}, 5^{th}, 1^{th}\}$ percentiles, for two-tail distributions, and at, or below the $\{20^{th}, 10^{th}, 2^{th}\}$ percentiles, percentiles in the case of SpO_2, which presents a one-tail distribution, respectively. The upper cut-off values, $\{1, 2, 3\}$, are set to the vital-sign integer values at or above the $\{90^{th}, 95^{th}, 99^{th}\}$ percentiles, respectively.

Figures 5.6a and 5.6b show the difference in the EWS of SBP (solid blue lines) between CEWS and ED-CEWS, respectively. ED-CEWS has significantly higher thresholds for high SBP. Figure 5.7 shows the quantised ED-CEWS for the remaining vital signs (HR, RR, SpO_2 and TEMP, in solid blue lines). The ED-CEWS is derived from the methodology in section 5.2.6.

(C) **ED-CEWS$_C$ (inclusion of age)**: We first use the median age of the training data to separate the vital-sign data into two age groups, from either side of the median, and recompute the CDFs and assign the scores, as before, for each age group.

Figure 5.8a shows an example of the EWS for SBP for both age groups (the median was derived from the methodology in section 5.2.6). In this example, the elderly will alert later as their SBP increases.

We also modelled the EWS cut-off values resulting from vital-sign CDFs conditioned

on each age bin. In this approach, the data are binned around each (integer) age value ± 15 years, in the range $[16, 99]$, for each vital sign. Figure 5.8d shows an example of the EWS for SBP varying with each age bin ± 15 years.

(D) **ED-CEWS$_D$ (inclusion of FiO$_2$)**: as in the NEWS system, for the original and modified versions of CEWS we investigated the addition of a score of 2 when oxygen therapy was given.

(E) **ED-CEWS$_E$ ("Non-quantised" EWS system)**: Hann (2008) proposed a fine quantisation for EWS. We analysed the impact of a non-quantised scoring system, with respect to the standard integer quantisation (scores 0, 1, 2, and 3).

We used cubic interpolation to map the quantised EWS values, {0, 1, 2, 3}, to a "continuous" EWS scale, $[0, 3]$, using the number of vital-sign values between each centile values used as cut-off for abnormal EWS. The cubic interpolation matched better the non-linear curve created by the CDF from which the abnormal EWS cut-off are originally derived.

This process can be done separately for the low (i.e. below the 50^{th} percentile of the CDF), and high (at or above the 50^{th} percentile of the CDF) values for each vital-sign distribution, mapping each of these regions to the interval [0,3]. As with the quantised ED-CEWS, Figures 5.6a and 5.6b show that the non-quantised (in green and red) versions have higher SBP cut-off values, the lower cut-offs remaining the same.

With this approach, every vital-sign value is assigned to a non-quantised EWS value. We also analysed a model for which the EWS interpolation occurs in the interval $[1, 3]$, for low (i.e. at or below the 10^{th} percentile of the CDF) and high (at or above the 90^{th} percentile of the CDF) vital-sign values, with a score of zero being assigned for all the normal range (interval between the 10^{th} and the 90^{th} percentiles, for HR, RR, TEMP and SBP, and above the 20^{th} percentile for SpO_2) . The non-quantised versions, ED-EWS $\in [0, 3]$, and ED-EWS $\in \{0\} \cup [1, 3]$, coded

as ED-CEWS$_{E,1}$, and ED-CEWS$_{E,2}$, respectively, are shown in Figure 5.6b for SBP (in green and red solid lines, respectively). We note that the quantised version of this model is represented by ED-CEWS$_B$. This non-quantised modelling approach can be applied to the CEWS system as well.

(F) **ED-CEWS$_F$ ("Non-quantised" EWS with inclusion of age and FiO$_2$)**: we combine all of the previous changes, from points (B), (C), (D) and (E) into one model, e.g. ED-CEWS$_{F,1}$ comprises the EWS system derived from the ED training dataset, conditioned on the patient's age, with non-quantised EWS $\in [0, 3]$, and with FiO$_2$ as an additional parameter, adding 2 to the EWS score, when present. In ED-CEWS$_{F,2}$ the non-quantised EWS $\in \{0\} \cup [1, 3]$ is used. To synthesise our notation, we make an exception and denote ED-CEWS$_{F,0}$, the quantised version of this model (EWS $\in \{0, 1, 2, 3\}$), representing the inclusion of age and FiO$_2$.

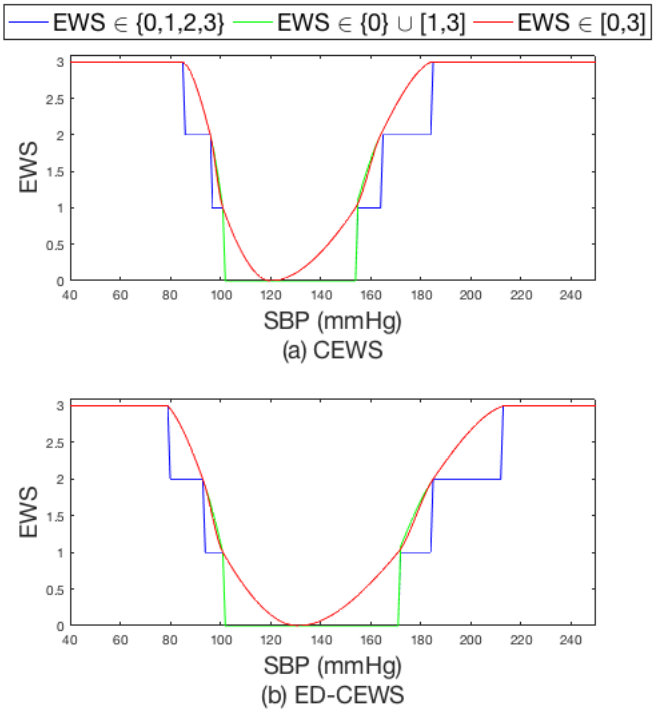


Figure 5.6: *EWS for SBP for the quantised (blue) and non-quantised (green and red) versions. (a) and (b) correspond to the CEWS and ED-CEWS models. A cubic-interpolation is used between the selected centiles to model the non-quantised EWS values.*

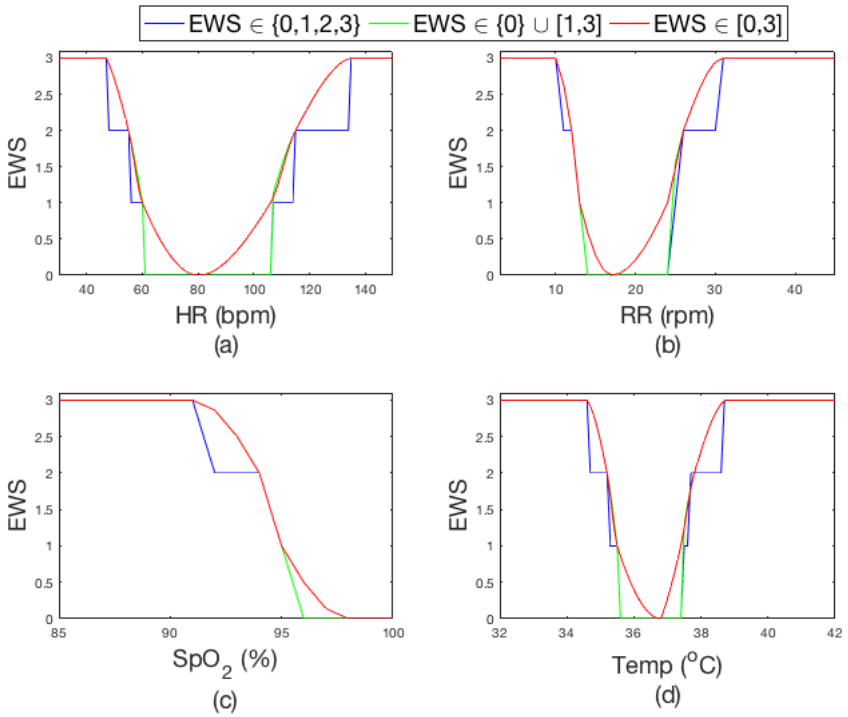


Figure 5.7: *ED-CEWS for HR, RR, $Sp0_2$, TEMP for the quantised (blue) and non-quantised (green and red) versions.*

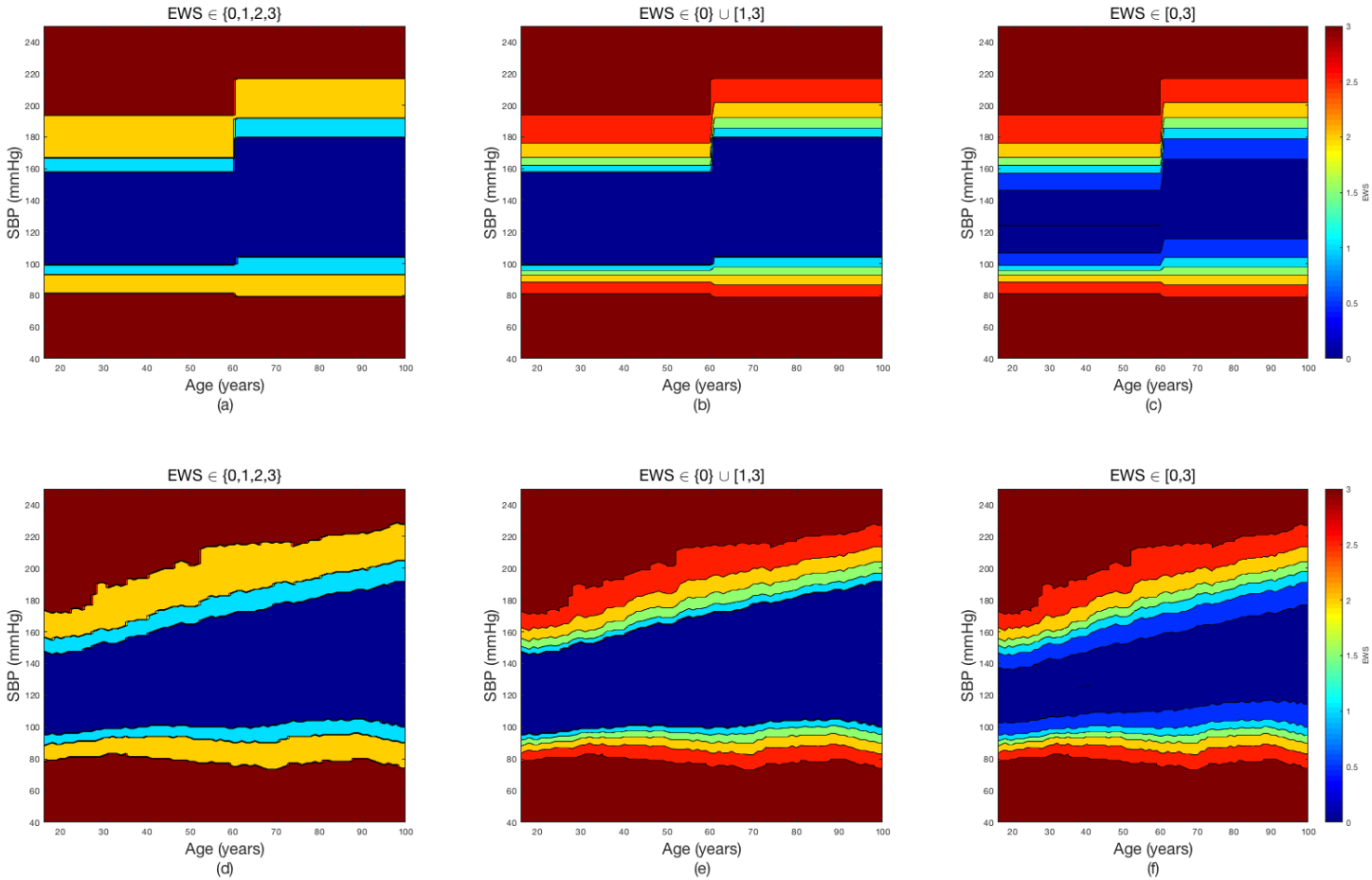


Figure 5.8: *EWS of SBP conditioned on the training set age. Top) CDFs computed for two groups separated according to the training set median age (median age was 60 years, IQR = [59, 61], derived from the 5-fold cross-validation procedure described in the next section); Bottom) Age regression model, based on computing CDFs for each (integer) age bin ± 15 years, in the range* $[16, 99] \pm 15$ *years. From left to right: (a) quantised EWS* $\in \{0, 1, 2, 3\}$*; (b) non-quantised EWS* $\in \{0\} \cup [1, 3]$ *; (c) non-quantised EWS* $\in [0, 3]$.

5.2.6 Model assessment

We followed the approach described in Prytherch et al. (2010) to evaluate the performance of EWS systems, and extended it with the model generalisation method discussed in Pimentel (2015). Figure 5.9 (experiment (A)) shows a block diagram of the strategy used to select the models' single-parameter thresholds. The five-fold cross validation procedure was used on the training dataset ($E_{\{1,2\},\{d,a\}}$) for deriving new cut-off values. Data was partitioned on a per-patient basis, i.e. all observations from a patient were either included in the training or validation sets.

Data from $K-1$ subsets of no-event (or "normal") patients (80%, for $K=5$) were used to compute the CDFs and the new cut-off values for the lower and upper bands of the vital-sign values of the proposed scoring systems, and the remaining subset of no-event patients (20%) and the set of event patients, $E_{\{1,2\},e}$, were used for validation.

In the validation phase, data from each fold were used to evaluate the performance of the EWS systems to identify abnormal observations (positive class) before the escalation time[7], and then used to select the threshold of the aggregated EWS that gave the best performance for the model from each fold. The best threshold was selected using ROC analysis on the validation set. The GCS was not optimised, and as in the original CEWS, a score of 3 is given to a total GCS of 13 or below, 1 if the total GCS is 14, and 0 for a total GCS of 15.

The performance of the selected "modified CEWS" systems were then assessed using an observation-wise AUROC analysis on an independent test dataset, E_3, to identify escalations occurring after ED arrival.

Efficiency Curves

Another way of optimising the EWS system configuration (with respect to its threshold), is analysing the trade-off between clinical staff workload and the prediction of the escala-

[7]This is well within the 24-hour period, used in both Prytherch et al. (2010) and Pimentel (2015), as ED patients stay on average up to 4 hours in the ED. The assumption is that any observation set within this period (from patient arrival) will be related to the escalation event.

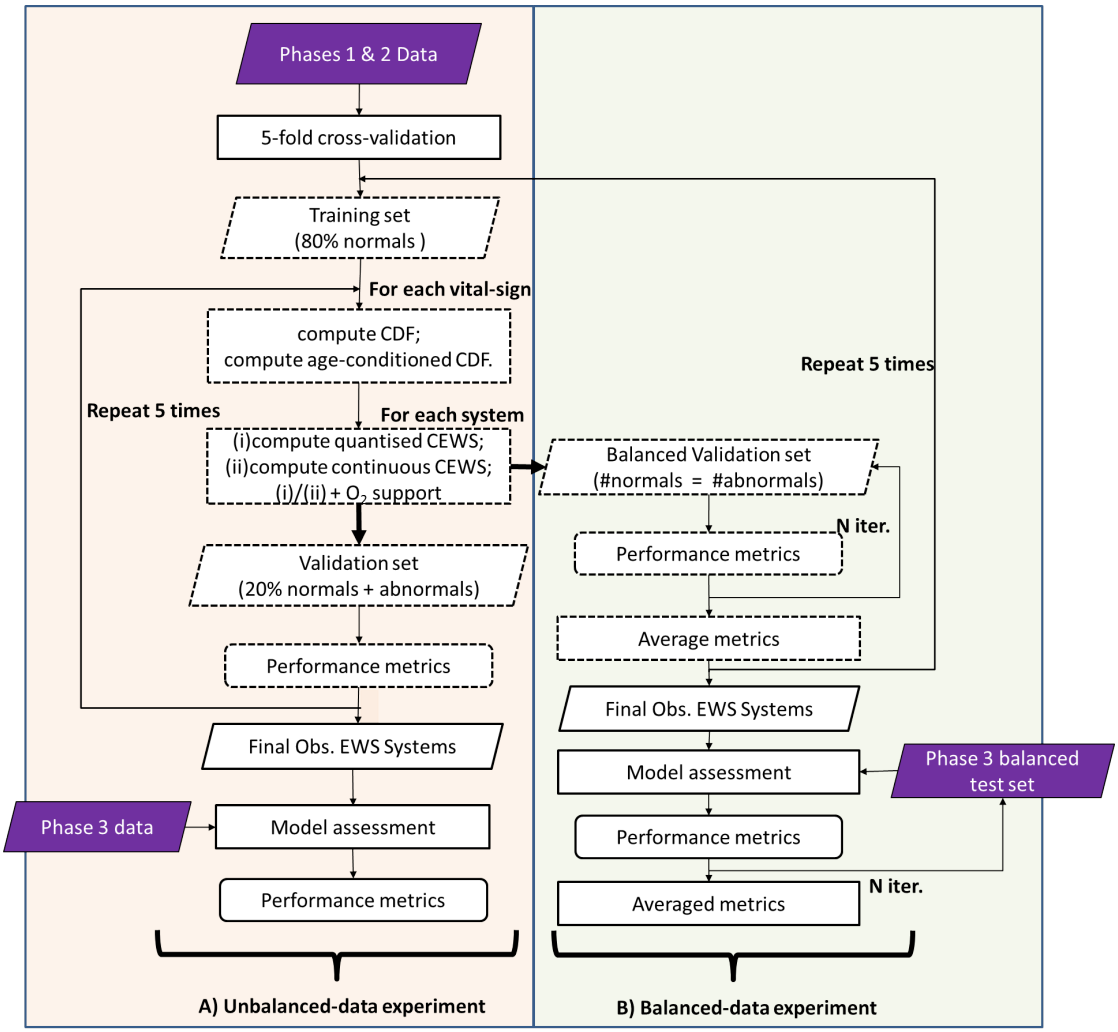


Figure 5.9: *Block diagram for modified CEWS systems model assessment, adapted from Pimentel (2015). A) refers to the evaluation of the EWS on the entire test dataset (unbalanced data); B) refers to the evaluation of the EWS systems using a balanced dataset strategy.*

tion events. We thus use the concept of "Efficiency Curves", as in Prytherch et al. (2010) to analyse this trade-off, by computing the number of observations with an EWS above the alert threshold versus the number of these alerts that are followed by an escalation event within the patient ED stay (i.e. we determine the Positive Predictive Value, PPV). We then compare between EWS systems configurations to determine which provides the most efficient workload for a specific PPV we wish to use for our ED dataset.

Balanced dataset analysis

Due to class imbalance, the evaluation of the EWS systems may be biased towards gains in identifying the class with the highest number of patients. In this case the system providing higher number of True Negatives (lower false positive rate) may rank higher. Solutions for training and testing machine classifiers in an unbalanced data set include (i) sampling methods, (ii) cost-sensitive methods, and (iii) Kernel and Active Learning methods (He and Garcia, 2009). In this case cost-sensitive methods would require learning from clinical experts the cost of misclassifying either class, which can be subjective, as it may depend on the clinical setting or clinical case, and using (iii) is outside of the scope of this chapter.

To recap, there are $n - \frac{n}{5}$ patients in the training set and $\frac{n}{5}$ in the validation set. Thus 63 no-event patients are sampled randomly from the $\frac{n}{5}$ patients in the validation set to match the 63 event patients, and this process is repeated 50 times. The performance metrics are averaged and the system ranked highest is selected.

In the test phase, we also match the number of no-event patients to event patients, by sampling the former from the existing pool of no-event patients in the held-out test set. We repeat this process 50 times and average the performance metrics to derive the final results.

5.3 Results

5.3.1 Performance on the unbalanced dataset

In our test dataset, E_3, of 1,070 patients with at least three observation sets (giving a total of 3,800 observation sets), only 33 patients had an event during their ED stay, with a total of 74 observation sets before the escalation events (note there can be less than three observations sets before the event for some of the event-patients). Table 5.5 shows the AUROC of the models on this held-out test set. The re-computed models (A to F) are displayed in red, those from the literature are displayed in black, and those from the literature for which the presence of oxygen mask is added (Subbe2001(1)′), or removed (DT-EWS′), are shown in blue. The baseline CEWS is shown in green (we keep this colour-code throughout the tables in this section). We summarise the results as follows:

- CEWS$_A$: showed an improvement in the AUROC from 0.626 (0.555, 0.702)[8] to 0.661 (0.584, 0.735), with respect to the baseline CEWS.

- ED-CEWS$_B$: had an AUROC of 0.627 (0.550, 0.696), just above the benchmark system (along with its non-quantised versions).

- ED-CEWS$_C$ (age included): Conditioning the vital signs on two age groups, divided by the median age, had higher AUROC than the age year $\pm$ 15 years binning approach. The latter approach was also tested for the $\pm$ 5 years, and $\pm$ 10 years intervals, but we observed that the $\pm$ 15 year interval smoothed the EWS transition between age bins better, and marginally improved its observation-wise AUROC, over the alternative versions. Hence the table shows the result of the former (i.e. that using the median age). This may be an indication that a better smoothing procedure for the age-bin-wise approach is required. Using the median age gave a higher AUROC, 0.631 (0.557, 0.695), than the benchmark CEWS.

[8]The respective 95% Confidence Interval of the Boostrap distribution, with 2000 resampled, is shown in the brackets.

- ED-CEWS$_D$ (FiO_2 included): The addition of an EWS of 2 for the presence of FiO_2 had the most significant impact in the ranking of ED-CEWS, increasing the performance to 0.655 (0.579, 0.731), over all the other modifications, as it was also observed for CEWS$_A$, mentioned in the first point.

- ED-CEWS$_E$: This approach ranked higher (not-significantly) than the respective quantised approaches, indicating that non-quantised scores may improve, the sensitivity of alerts for the ED dataset.

- ED-CEWS$_F$ (non-quantised EWS system, and age and FiO_2 included): The non-quantised version ED-CEWS$_{F,1}$, i.e. EWS $\in$ [0,3], addition of 2 to the EWS for the presence of FiO_2 support and vital-signs CDFs conditioned on two age groups, divided by the median age, had the best AUROC, 0.664 (0.593, 0.734), amongst the systems re-trained using the ED training dataset, and the CEWS system (including CEWS$_A$).

ED-CEWS$_{F,1}$ performed better than the CEWS-based systems but it was outperformed by 9 EWS systems given our current observation-wise AUROC analysis metric. Although conditioning the vital signs by age groups did not have a significant effect on EWS performance, we note that the system ranking third, Subbe2001(1), with an AUROC of 0.698 (0.630, 0.756), adds a score of 2 and 3 for people above 50 and 70 years old, respectively. However, the systems using the FiO_2 parameter ranked highest in general, DT-EWS first, with an AUROC of 0.711(0.648, 0.769), amongst the systems found in the literature. The impact of the FiO_2 parameter was further investigated by adding it to Subbe2001(1), which showed the highest performance on our dataset, AUROC = 0.720 (0.644, 0.781). In this case, the age parameter together with FiO_2 led to a substantial improvement in performance given that Subbe2001(2), which differs from the original Subbe2001(1) in not scoring the age parameter, showed a much lower rank (AUROC = 0.663), than both Subbe2001(1) and its modified version Subbe2001(1)$'$. Moreover, if FiO_2 is removed from DT-EWS (which scores 3 for this parameter), its AUROC decreases

Table 5.5: *Baseline performance evaluation of EWS (unbalanced dataset, observation-wise analysis case). The table shows EWS performance metrics for existing multi-parameter EWS systems and 6 proposed CEWS systems optimised on our dataset (the benchmark CEWS shown in green). The number of "abnormal observation sets" (positive class) before the event is 74 while the number of "normal observations sets" before the critical interval or from non-event patients (negative class) is 3,800. T is the optimum threshold.* [a] *A version of Subbe2001(1) with an additional individual EWS of 2 for the presence of* FiO_2. [b] *DT-EWS without* FiO_2. *The AUROC 95% Confidence Interval (CI) is shown (determined by the bias-corrected and accelerated (BCa) bootstrap approach, with 2000 resamples).*

System	AUROC (95% CI)	T
Subbe2001(1)′[a]	0.720 (0.644, 0.781)	13
DT-EWS	0.711 (0.648, 0.769)	16
Subbe2001(1)	0.698 (0.630, 0.756)	11
ViEWS	0.685 (0.615, 0.754)	14
DT-EWS′[b]	0.682 (0.622, 0.740)	13
NEWS	0.681 (0.611, 0.746)	13
Andrews2005	0.667 (0.602, 0.724)	9
Smith2006	0.665 (0.602, 0.726)	9
Lam2006	0.665 (0.604, 0.724)	9
Gardner-Thorpe2006	0.665 (0.596, 0.726)	9
Wright2000	0.664 (0.601, 0.725)	9
ED-CEWS$_{F,1}$	0.664 (0.593, 0.734)	14.5
Subbe2001(2)	0.663 (0.595, 0.720)	9
Hancock2007	0.662 (0.593, 0.723)	9
CEWS$_A$	0.661 (0.584, 0.735)	13
Von-Lilienfeld-Toal2007(2)	0.661 (0.588, 0.729)	12
Subbe2003	0.660 (0.598, 0.721)	9
Riley2001	0.660 (0.595, 0.721)	10
Rees2004	0.659 (0.600, 0.715)	9
ED-CEWS$_D$	0.655 (0.579, 0.731)	13
Paterson2006	0.655 (0.589, 0.729)	9
Duckitt2007	0.652 (0.580, 0.719)	10
MEWS	0.650 (0.566, 0.717)	13
Cooper2001	0.649 (0.582, 0.710)	9
Chatterjee2005	0.649 (0.576, 0.718)	10
Goldhill2005a	0.647 (0.567, 0.711)	9
Von-Lilienfeld-Toal2007(1)	0.638 (0.567, 0.709)	11
Subbe2007	0.636 (0.568, 0.703)	9
ED-CEWS$_{E,2}$	0.632 (0.559, 0.701)	12.4
ED-CEWS$_C$	0.631 (0.557, 0.695)	11
ED-CEWS$_B$	0.627 (0.550, 0.696)	11
CEWS	0.626 (0.555, 0.702)	12
Allen2004	0.602 (0.528, 0.672)	8

below that of ViEWS (but is still higher than any of the proposed ED-CEWS systems). Finally, we note that due to the class imbalance (1067 no-event versus 33 event-patients) the difference between the AUROCs is not significant.

5.3.2 Use of efficiency curves

Figure 5.10 presents the percentage of observations that are above a given EWS alert threshold versus the percentage of those that alert and are followed by an escalation event. Efficiency curves allow us to analyse the trade-off between clinical staff workload (number of observation sets that generate alerts - both true and false positives) and the PPV of the EWS systems. This analysis is different from the ROC analysis in that there is no optimal operating point, and we must select, instead, a performance target to make the EWS system clinically useful. For this analysis we require that the EWS systems shall identify at least 2/3 of the observations (49 observation sets in the test set, E_3) occurring before the escalations (true positives), with the lowest workload possible. We choose the closest EWS integer value to the right of PPV = 66.7%, (x-axis in Figure 5.10) for the quantised systems and the closest EWS decimal value for the non-quantised system.

The original threshold value for the aggregated EWS for three of the systems with highest rank in the previous section, and the baseline CEWS, and two versions of the ED-CEWS system with highest rank: 6 for DT-EWS, 4 for Subbe2001(1), 5 for ViEWS, 5 for NEWS, and 3 for CEWS, ED-CEWS$_{F,1}$ and ED-CEWS$_{F,2}$, are labelled in the Figure 5.10. Note that the original thresholds fall to the left of the PPV = 66.7% target.

Table 5.6 summarises those results falling at or to the right of the PPV target (66.7%). We observe that our proposed non-quantised ED-CEWS$_{F,1}$ system would find at least 2/3 of the positive cases, and have the lowest amount of false alerts per TP clinical observation alert, i.e. 26 false alerts for each TP alert. The false alerts were calculated by dividing the number of false alerts by the number of TP alerts at the EWS threshold value closest to 66.7% PPV, e.g. at a threshold of 2.0, ED-CEWS$_{F,1}$ had

$50.0\% \times 3,800/74 = 26$ false alert observation sets per TP alert; and at a threshold of 2, NEWS had $55.0\% \times 3,800/74 = 28$ false alerts per TP alert.

Figure 5.10 also shows that other system configurations could provide more efficient clinical workloads at the cost of a lower PPV, amongst the selected. For example, Subbe2001(1) at a EWS threshold of 5 would provide a 58.1% PPV (# TP = 43 observation sets) with a cost of 14 false alerts for each TP alert. In the presence of an unbalanced dataset, such as our ED dataset, the efficiency curve allows for a more reasonable choice of the EWS threshold than the ROC analysis.

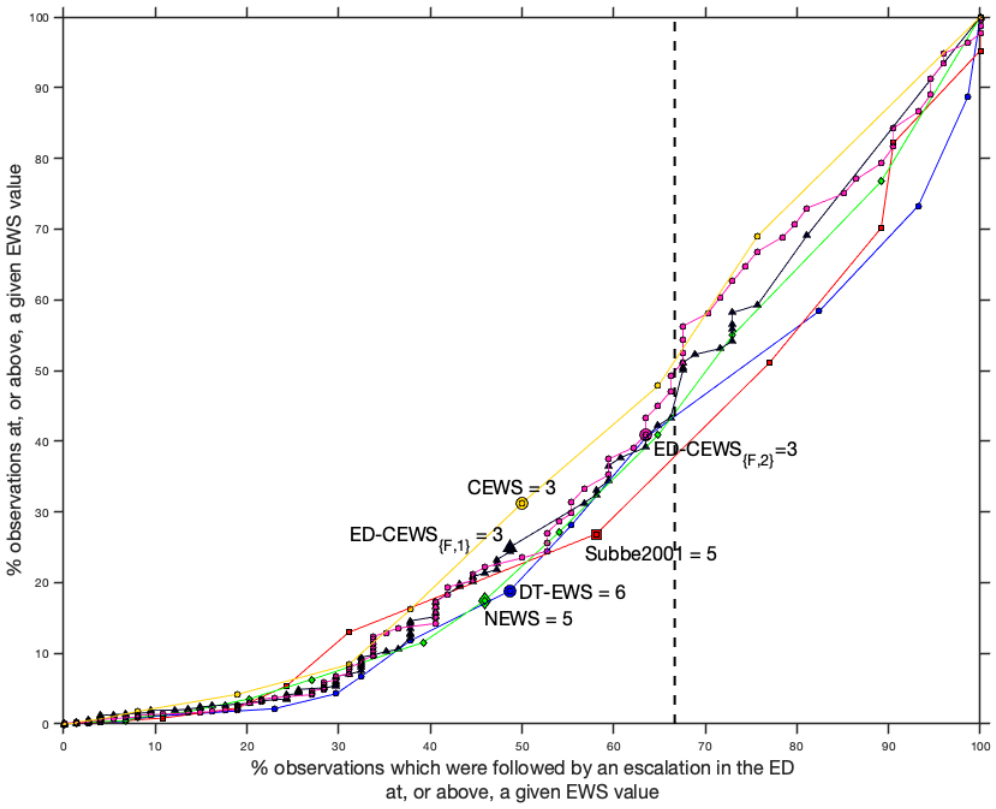


Figure 5.10: *The plot shows on the y-axis the percentage of observations that are above a given EWS alert threshold versus, on the x-axis, those that alert and are followed by an escalation event during the patient ED stay (PPV), for the three of the systems with the highest AUROC in the unbalanced data experiment (metrics shown in Table 5.5) namely, DT-EWS, Subbe2001(1) and NEWS, and the baseline CEWS, ED-CEWS$_{F,1}$ and ED-CEWS$_{F,2}$ (the non-quantised ED-CEWS versions which use the FiO_2 parameter and with vital-sign CDFs conditioned on two age groups, with highest AUROC, amongst the ED-CEWS systems).*

Table 5.6: *We show the EWS system configurations that would allow identifying at least 2/3 of the observation sets before the ED escalations using the training dataset (E_3). For each configuration we also show the number of false alerts that would be generated per each TP alert observed by staff. T - threshold for the aggregated EWS. [# false alerts per TP] = $\frac{\#obs.\ above\ T}{TP}$. PPV - Positive Predictive Value.*

System	T	PPV $\geq$ 66.7%	% obs. above T	# TP	# false alerts per TP
DT-EWS	3.0	82.4	58.4	61	30
Subbe2001(1)	4.0	77.0	51.1	57	26
NEWS	2.0	73.0	55.0	54	28
ED-CEWS$_{F,1}$	2.0	67.6	50	50	26
ED-CEWS$_{F,2}$	2.6	67.6	51.2	50	26
CEWS	1.0	75.7	69.0	56	35

5.3.3 Use of efficiency curves on the entire ED dataset

As the efficiency curves do not require any parameters to be optimised on a training set, we also analysed them using the entire dataset of patients $E_{\{1,2,3\}}$, with a total of 9,959 observation sets, 205 of which before the escalation events. The results are shown in Figure 5.11. The EWS system configurations that would again allow a PPV of at least 2/3 of the positive class are shown in Table 5.7. We observe that the non-quantised ED-CEWS$_{F,2}$ would present the lowest number of false alerts per TP clinical observation alert (17), again, with a reasonable EWS threshold of 3.3 (in contrast with CEWS which presents a higher false alert rate at a threshold of 2 to achieve the same PPV).

We note that ED-CEWS$_{F,2}$ differs from ED-CEWS$_{F,1}$ in depressing the EWS to zero, rather than giving a non-quantised score between 0 and 1, for the range of stable physiology, considered between the 10^{th} and the 90^{th} percentiles of the vital-sign CDFs (except for SpO_2, in which it corresponds to values above the 20^{th} percentile). The effect is lowering the number of false alerts while allowing the system to meet the proposed target PPV $\geq$ 66.7%.

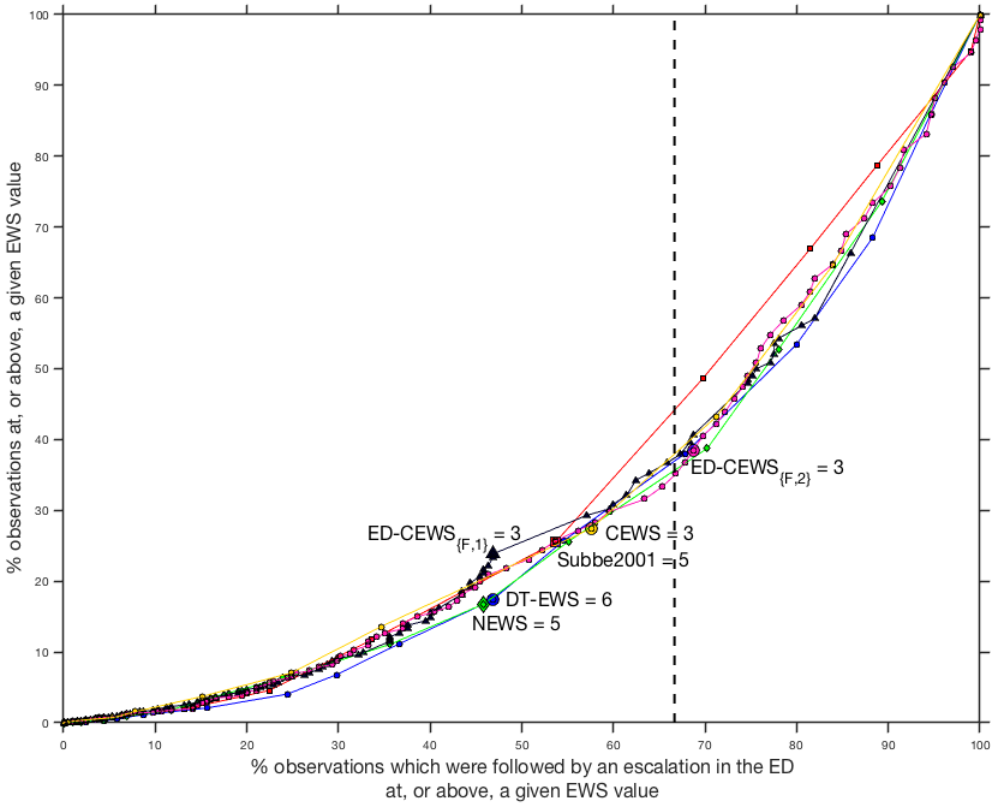


Figure 5.11: *The plot shows on the y-axis the percentage of observations that are above a given EWS alert threshold versus, on the x-axis, those that alert and are followed by an escalation event, for three of the systems with the highest AUROC in the unbalanced data experiment (metrics shown in Table 5.5) namely, DT-EWS, Subbe2001(1), and NEWS, and the baseline CEWS, the ED-CEWS$_{F,1}$ and ED-CEWS$_{F,2}$ systems (non-quantised EWS, which use FiO_2, and 2 age groups to condition the vital-signs CDFs, with highest AUROC amongst the ED-CEWS systems). All data is used in this case, i.e. 96 event patients in a total of 2,803 patients. For completion, their AUROCs for these dataset where: 0.712, 0.674, 0.706, 0.690, 0.695, and 0.698, respectively.*

Table 5.7: *We show the EWS system configurations that would allow identifying at least 2/3 of the observation sets before the ED escalations using the training and test datasets ($E_{1,2,3}$). For each configuration we also show the number of false alerts that would be generated for each TP alert observed by staff. T - threshold for the aggregated EWS. [# false alerts per TP = $\frac{\#obs.\ above\ T}{TP}$. PPV - Positive Predictive Value.*

System	T	PPV $\geq$ 66.7%	% obs. above T	# TP	# false alerts per TP
DT-EWS	4.0	67.8	38.0	139	18
Subbe2001(1)	4.0	69.8	48.6	143	24
NEWS	3.0	70.2	38.8	144	19
ED-CEWS$_{F,1}$	2.3	67.3	37.9	138	18
ED-CEWS$_{F,2}$	3.3	66.8	35.2	137	17
CEWS	2.0	71.2	43.2	146	21

5.3.4 Performance on the balanced dataset

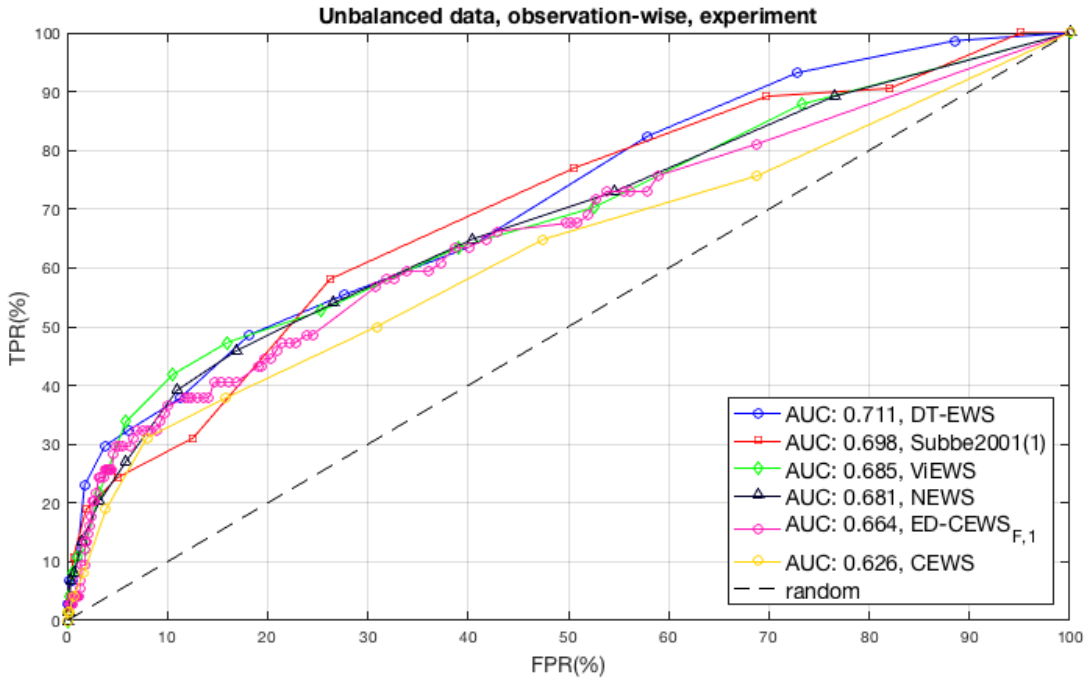
The dominant characteristic in our dataset is the class imbalance. Selecting the majority class, i.e. classifying all patients as normal, would result in near 100% accuracy. Consequently, the optimal operating point would be set at a high threshold when using the ROC analysis, in order to classify most observation sets as TN cases. We thus repeat the ROC analysis using a balanced dataset experiment.

The balanced dataset experiment was carried out by computing the performance metrics using the same number of normal (33 patients, selected randomly from the no-event patients pool, for each fold) and abnormal patients, 50 times, and averaging the results. In each test fold the number of normal and abnormal observations is more balanced than the previous experiment, with 119 ± 7 normal observation sets on average versus 74 abnormal observations sets before the escalation.

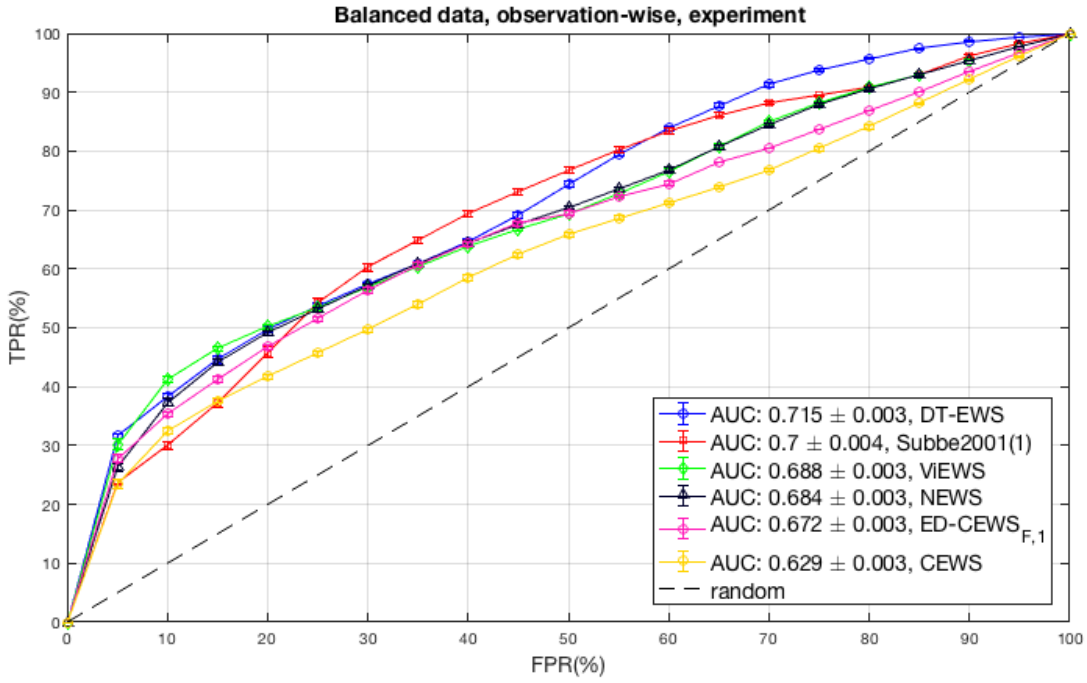
The AUROC results for the observation-wise performance analysis are shown in Table 5.8 and are similar to those of the unbalanced dataset case. The ROC curves in Figures 5.12*a* and 5.12*b*, representing the unbalanced and balanced experiments, respectively, show similar patterns for both experiments. However, as expected, the EWS thresholds for the optimal operating point of each system are much lower than for the unbalanced dataset, for example, the (rounded) optimal aggregated EWS threshold for CEWS_A is now 6 ± 1, rather than 13, that selected for the unbalanced experiment (Table 5.5). There are small differences between the quantised and non-quantised versions of the proposed systems (the latter showing a lower score in this experiment), but the ranking of the proposed systems configurations is similar to that for of the unbalanced dataset.

Table 5.8: *Balanced-dataset, observation-wise, performance analysis. The table shows performance metrics for multi-parameter EWS systems and 6 proposed CEWS systems, for a balanced dataset experiment. T is the optimised threshold. Values are presented as the mean (standard deviation), resulting from 50 balanced data experiments. The number of "abnormal patients" (positive class) is 33.* [a] *A version of Subbe2001(1) with an additional individual EWS of 2 for the presence of* FiO_2. [b] *DT-EWS without* FiO_2 *parameter.*

System	AUROC	T
Subbe2001(1)′[a]	0.721±0.03	6.96±0.67
DT-EWS	0.715±0.03	8.10±1.33
Subbe2001(1)	0.700±0.04	6.36±1.38
ViEWS	0.688±0.03	6.18±0.69
DT-EWS′[b]	0.687±0.03	6.72±1.03
NEWS	0.684±0.03	5.90±0.93
ED-CEWS$_{F,0}$	0.673±0.03	5.88±1.14
Andrews2005	0.670±0.03	4.68±0.59
Lam2006	0.668±0.03	4.76±0.89
Gardner-Thorpe2006	0.668±0.03	4.74±0.85
Smith2006	0.668±0.03	4.28±0.70
Wright2000	0.666±0.03	4.44±0.73
Subbe2001(2)	0.665±0.03	4.56±0.58
Hancock2007	0.664±0.03	4.78±0.76
Von-Lilienfeld-Toal2007(2)	0.664±0.03	4.34±0.63
CEWS$_A$	0.663±0.03	5.54±0.93
Subbe2003	0.662±0.03	4.40±0.76
Riley2001	0.661±0.03	5.26±0.75
Rees2004	0.661±0.03	4.16±0.91
ED-CEWS$_D$	0.659±0.03	5.88±0.77
Paterson2006	0.657±0.03	3.82±0.52
Duckitt2007	0.656±0.03	4.98±1.60
MEWS	0.653±0.03	4.74±1.05
Goldhill2005a	0.652±0.03	3.92±0.27
Cooper2001	0.651±0.03	4.38±0.97
Chatterjee2005	0.648±0.03	4.22±0.58
ED-CEWS$_C$	0.647±0.03	5.04±0.35
Von-Lilienfeld-Toal2007(1)	0.644±0.03	4.24±0.85
Subbe2007	0.641±0.04	3.82±0.72
ED-CEWS$_{E,2}$	0.638±0.03	5.49±0.53
ED-CEWS$_B$	0.632±0.03	5.36±0.80
CEWS	0.629±0.03	4.96±0.49
Allen2004	0.600±0.03	3.74±0.60



(a)



(b)

Figure 5.12: *(a) and (b) show ROC curves for the unbalanced and balanced dataset analysis, respectively, for the 6 selected EWS systems.*

5.4 Discussion and Conclusion

In this chapter the performance of the proposed data-driven EWS systems to detect physiological deterioration in the ED was analysed against other aggregate weighted EWS systems currently used in clinical practice (Prytherch et al., 2010), and our baseline CEWS system, being used in the ED of the John Radcliffe Hospital, Oxford, at the time of the large-scale ED study (chapter 3).

The optimal system would be one that could alert on abnormal observation sets before the escalation event, while not alerting on observation sets from patients without escalation events. Escalations to resus area, was the type of event identified in the dataset, for our analysis. In order to make a robust assessment of the proposed systems, we used cross-validation to tune the thresholds and out-of-sample data to test the EWS model performance.

As a consequence of the highly unbalanced dataset (96 event patients of a total of 2,803 patients, i.e. 3.5%), AUROC-based analysis favours systems with high specificity and low sensitivity. Thus, we designed a balanced dataset experiment, by "down-sampling" the negative class to the same number of examples in the positive class (66 and 33 patients for the validation and test set, respectively). To avoid losing important data structure, we ran the balanced experiment 50 times, sampling randomly (and without repetition in each run) from the negative class at each fold.

Using the ROC analysis we observed that:

1. Systems incorporating the FiO_2 parameter had the highest AUROC, as this parameter was present in 35% of the 74 observations before the escalations in event patients. We showed that by adding this parameter to the Subbe2001(1) EWS system enables this system to achieve the highest score in both the unbalanced and the balanced analysis (AUROC = 0.720, and AUROC = 0.721 $\pm$ 0.03, respectively).

2. The original DT-EWS system, which was trained using a decision-tree algorithm on a large quantity of clinical observations (198,755 vital signs, from 35,585 patients)

and has a score of 3 for the presence of FiO_2, consistently scored highly on our dataset. We also note that the DT-EWS dataset was acquired from patients from the Medical Decision Unit, which is a ward that complements the ED in caring for unscheduled patient admissions, whose case requires more than 4 hours to reach an accurate clinical assessment.

3. The ED-CEWS version with the best performance was that combining the addition of age and FiO_2. The performance gain is larger for the addition of the FiO_2 parameter. Conditioning on age worked better when using two age groups, rather than binning the vital signs per age year. We note, however, that the system ranked just behind NEWS. The addition of age allowed the Subbe2001(1) system to achieve a better performance than Subbe2001(2), which does not use the age parameter, in our dataset. This approach to model the age parameter should also be tested for the CEWS based systems in future work.

4. The non-quantised scores did not have a significant impact in the performance of ED-CEWS, as far as AUROC analysis is concerned.

The use of a balanced dataset also presents disadvantages, as it creates an artificial experiment to optimise EWS system configuration. We thus used the concept of efficiency curves, as in Prytherch et al. (2010), to specify a desired target performance for our systems. We can then analyse which system configurations present the best trade-off between clinical staff workload (observations sets that alerted) and prediction of the escalation events.

At a reasonable PPV of 66.7% (or 2/3 of the positive class) our proposed non-quantised ED-CEWS$_{F,2}$ system showed the lowest number of false alerts per TP observation set (17 false alerts), at an EWS threshold of 3.3 when testing on the entire ED dataset, $E_{\{1,2,3\}}$, (2.6 when assessed on the test dataset, E_3). This shows that there should be a benefit in having a score that increases in steps of 0.1 rather than integers: the number of false alerts per TP observation set would be reduced.

Chapter 6

Machine Learning methods for patient condition monitoring

6.1 Introduction

In this chapter we review the Machine Learning (ML) methods that have been proposed for patient condition monitoring in the hospital setting. "A computer program is said to learn from experience E' with respect to some class of tasks T' and performance measure P', if its performance at tasks in T', as measured by P', improves with experience E'" Mitchell (1997). In our work, we would like the ML system to process clinical data features $\mathbf{x} \in \mathbb{R}^n$, e.g. $\mathbf{x}_1 = \{\text{HR}, \text{RR}, \text{SpO}_2, \text{SBP}\}_1$ represents the vital-sign feature vector measured at time t_1 from a patient (our data E'), in order to detect abnormal physiology before an escalation event during the patient ED stay (our task T').

From the different tasks that ML algorithms can perform, we use in this thesis:

- *Novelty detection*, for which "normal" patterns $\mathbf{x}$ are available for training, whereas "abnormal" patters are rare, so that the algorithm learns the density for the normal dataset; subsequent feature vectors far away from its modes are considered to be abnormal;

- *Classification*, for which the algorithm is asked to identify which of c categories

some input belongs to, by learning a function $f : \mathbb{R}^n \rightarrow 1, ..., c$, which assigns an input described by vector $\mathbf{x}$ to a category c, i.e. $c = f(\mathbf{x})$;

- *Regression*, for which the algorithm learns a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, to predict a (continuous) numerical value (rather than a category, as in classification) given an input;

- *imputation of missing values*, which allows the estimation of the values of the missing entries;

- *denoising*, in which the algorithm learns to generate clean data from corrupted feature vectors.

The performance measure (P') is specific to the task T'. For example, in chapter 5 the AUROC metric is used in the observation-wise performance analysis, to determine the best EWS model for the detection of physiological deterioration before an escalation event in the ED.

Most ML algorithms can therefore be described as instances of the combination of a dataset specification, a cost function, an optimisation procedure and a model (Goodfellow et al., 2016).

Generalization and Capacity

To generalise from the training set to the test set, most ML algorithms assume a common structure in the data, i.e. they assume the data are independently and identically distributed (i.i.d.). In the i.i.d. assumption each exemplar feature vector is generated independently from other examples, and each example is drawn from the same distribution $p(\mathbf{x})$, formally, $p(\mathbf{x}) = \prod_i p(\mathbf{x}_i)$. Underfitting occurs when the model performs poorly on the training data, and overfitting occurs when the gap between the training error and the test error is too large, i.e. the model is overfitted to the training data, and does not generalise well on the test data. We can control these effects by controlling the model's

capacity or its hypothesis space, i.e. the set of functions that the algorithm is allowed to select as being the solution. Linear regression is a case of polynomial regression with a degree of one. If we change the order to two, the hypothesis space is now a linear combination of a linear and a quadratic function. A ML algorithm will generalise better if its capacity is appropriate for the complexity of the data, and this may guide the selection between non-parametric (complexity grows as a function of the training set) and parametric (function described by a parameter vector with finite size, fixed before any data are observed) models.

The hyperparameters control the capacity or the regularisation of a model, and to avoid overfitting these to the training set, a further independent set of examples, the validation set (typically 20% of the training set) is used to select the hyperparameters. Once the hyperparameter optimisation is complete, the model generalisation error is estimated using the test set.

Parameter Estimation and Optimisation

A common approach to learn the parameters of a ML model, when using the i.i.d. assumption, is maximum likelihood estimation, in which a probabilistic model of the data is controlled by a set of parameters θ drawn from a set Θ of possible values. This procedure selects those that maximise the probability that the model will generate the training data. The estimator has the form

$$\begin{aligned} \hat{\theta} &= \text{argmax}_{\theta \in \Theta} \prod_i p(\mathbf{x_i}, \theta) \\ &= \text{argmax}_{\theta \in \Theta} \sum_i \log p(\mathbf{x_i}, \theta), \end{aligned} \tag{6.1}$$

in which the monotonically increasing property of the logarithm is explored, and the log likelihood is optimised instead[1]. Under the i.i.d. assumption, the maximum likelihood

[1]Also it is more convenient to decompose this operation into a sum, as it avoids numerical problems caused by the product of several factors in the interval [0, 1].

estimation approach is consistent, i.e. in the limit of infinite data the estimator recovers the correct distribution of the data.

To find θ in the maximum likelihood approach the objective function

$$\ell(\theta) = \sum_i \log p(\mathbf{x_i}, \theta) \tag{6.2}$$

is defined from the log likelihood, and the optimisation problem consists in maximising $\ell(\theta)$, subject to $\theta \in \Theta$. In simple models this can be done analytically (i.e. solving the derivative $\nabla_\theta \ell(\theta) = 0$ for θ). In the cases for which there is no closed form, gradient ascent can be used. In each iteration i of this algorithm, the update value of θ is computed to ensure that θ moves towards the direction in which $\ell(\theta)$ increases most rapidly, using the equation:

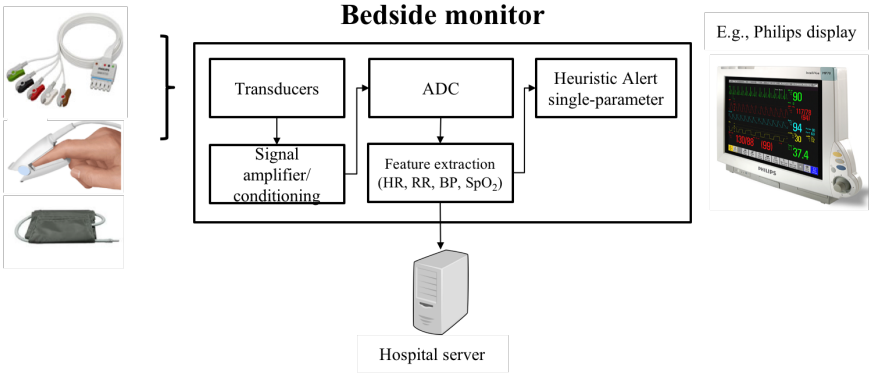
$$\theta^{(i)} = \theta^{(i-1)} + \alpha(i)\nabla_\theta \ell(\theta) \tag{6.3}$$

where $\alpha(i)$, the learning rate, is a positive scalar controlling the size of the optimisation step Bishop (2007).

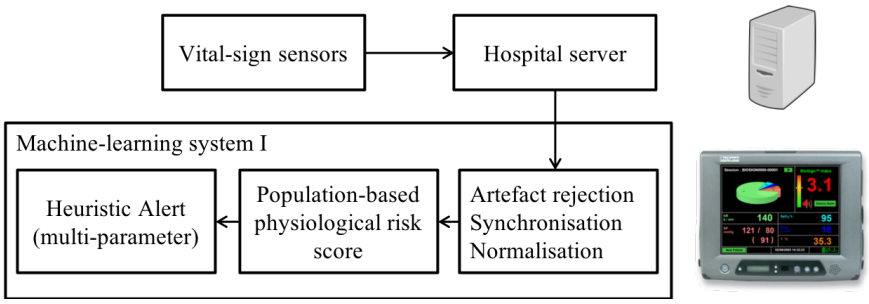
6.2 ML for patient condition monitoring

Figure 6.1 shows the main conceptual differences between (i) simple expert systems currently used for patient condition monitoring in the ED setting, the (ii) ML system deployed in the large-scale ED study, which is analysed in the next chapter, and the (iii) improved ML system proposed in the later chapters of this thesis.

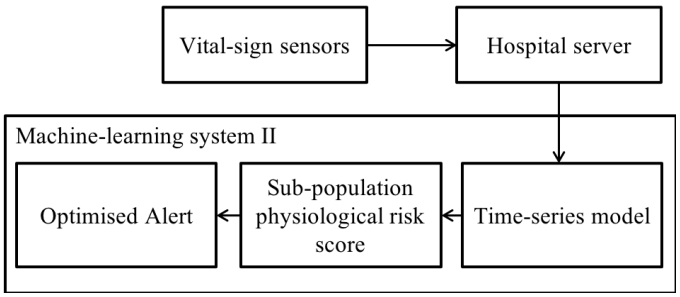
Expert systems are rule-based classifiers built from expert knowledge, similar to those EWS systems, defined heuristically, analysed in the previous chapter. Other expert systems have been described, and these can be broadly divided into those for which dynamics are not taken into account (Oberli et al., 1999), and those that make use of time-series trends, as in Charbonnier and Gentil (2007). Expert systems reported in the literature are usually trained on a small dataset, in which they perform well, and not



(a) *Current in-hospital vital-sign monitoring systems. Statistical signal processing methods and heuristics are used to filter the data and algorithms are deployed to extract numeric features (HR, RR, BP and SpO_2). Heuristic single-parameter thresholds, such as those in EWS systems are used to trigger alerts. Data can be stored in hospital servers, and made available to other systems.*



(b) *FDA-approved data-fusion system, with population-based probabilistic model to identify patient physiological deterioration, currently being deployed in hospitals. Data are synchronised using standard statistical methods and heuristics, and then scored. Heuristic multi-parameter alerting thresholds and a persistence criterion are used.*



(c) *A 2-stage machine learning approach for vital-sign monitoring systems has been proposed in recent literature to deal with the high-frequency vital-sign data available in hospital wards. The first stage learns the patients' physiological time-series dynamics, and the second stage learns a model able to identify physiological deterioration, taking the patient context into account (patient condition/disease/hospital ward, age, sex, etc...). The model hyper-parameters and alerting threshold are optimised using performance metrics appropriate to the task.*

Figure 6.1

generalised to larger datasets. Other limitations found in these approaches are: (i) rule-based systems are simple when compared with the number of possible clinical outcomes; (ii) medical knowledge is difficult to capture and translate into rules; (iii) rule formation is labour-intensive for clinicians, and (iv) re-evaluation is required when the rules change (Hann, 2008). The alerting thresholds of the expert systems currently deployed in the hospital, such as that integrated into the bedside monitor illustrated in Figure 6.1a, are often re-configured or turned off by clinical staff, depending on the patient context, to avoid a high false alert rate, especially in busy wards such as the ED (Way et al., 2014).

Current ML based systems, such the FDA-approved system described in (Tarassenko et al., 2005), illustrated in Figure 6.1b, use standard statistical signal processing methods and heuristics to filter and synchronise the multivariate vital-sign time-series data, and a population-based model of normality that alerts when test data differs considerably from that model.. The novelty detection approach used by this system is detailed in section 6.3. Heuristics are used in this systems to select a multi-parameter alert threshold as well as a persistent deterioration period to avoid alerting on intermittent artefacts.

In this thesis, we improve on recent research work that uses a 2-stage ML approach, in which first a time-series model of the patients' physiology is developed, and then a sub-population-based physiological risk assessment tool is used to identify patient deterioration (Clifton et al., 2012). In this case, the alerting criterion is optimised using performance metrics such as the AUROC. The ML approach of such systems is detailed in section 6.4.

6.3 Novelty detection

Novelty detection is well suited to the ED setting as patients come into the hospital with multiple symptoms, and only a few (fewer than 5% in a large population dataset; see chapter 3) suffer a major adverse event, while the majority of patients stay in the ED without experiencing an adverse event.

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, 2009; Markou and Singh, 2003a,b; Marsland, 2003; Pimentel et al., 2014), describing the different types of models H, methods for setting their parameters $\boldsymbol{\theta}$, and for determining the novelty thresholds.

Pimentel et al. (2014) provide a structured overview of recent studies, classifying novelty detection techniques according to five general categories: probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic techniques. **Probabilistic methods** rely on the density estimation of the "normal" class, $p(\mathbf{x}|\theta)$. Low-density areas in the training set indicate low probability of containing "normal" examples. **Domain-based methods** are typically insensitive to the distribution of the "normal" class. A domain containing "normal" data are characterised by defining a boundary around the "normal" examples, but does not provide an explicit estimate of the distribution. Classification of test data are then determined by their location with respect to the boundary. For example, the One-class Support Vector Machine (OSVM) approach (Schölkopf et al., 1999) 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 hypersphere 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. **Distance-based methods** include the concepts of nearest-neighbour and clustering analysis. Appropriate distance metrics are used to compute a similarity measure between test data and the "normal" training data. "Normal" data are assumed to be tightly clustered, while novel data occur far from the training set in the data space. **Reconstruction-based approaches** involve training a regression model using the training set. Test data are evaluated using the "reconstruction error", which is the difference between the test point and the output of the mode. "Abnormal" data give rise to a high novelty score (high re-

construction error). This category of methods includes different configurations of neural networks and principal component analysis. Finally, **Information-theoretic methods** compute the information content in the training data using information-theoretic measures, such as entropy or Kolmogorov complexity. Novel data significantly alter the information content in a dataset, while "normal" data (being similar to the training data) do not.

Many of these methods have data-specific parameters based on factors such as the availability of training data, the type of data, and the application domain. Data knowledge is usually necessary to develop an optimal novelty detection approach for real-world datasets.

6.3.1 Baseline novelty detection model

In Tarassenko et al. (2005), the following steps were taken to train and apply a probabilistic novelty detection method, able to alert on vital-sign patterns deviating from "normality":

1. **Data pre-processing for the training phase:** 3,500 hours of continuous vital-signal data were collected from 150 high-risk patients at the JR, between 2001 and 2003, as part of an observational study. Patients were monitored after having suffered a myocardial infarction or severe heart failure, acute respiratory problems, or trauma injuries. The physiological variables included in the model were HR, RR, SpO_2, and temperature sampled at a frequency of 1 Hz, and BP values measured at 30-minute intervals during the day, and at hourly intervals during the night. As the data were recorded asynchronously, they were re-sampled at five-second intervals and filtered for artefactual values using physiological thresholds (e.g. Table 4.1).

2. **Feature extraction:** For both the training and test sets, the vital signs (HR, SpO_2, RR and Temperature) were used as features directly. The BP was converted into the Systolic and Diastolic Average, $SDA = \frac{1}{2}(\text{SBP} + \text{DBP})$, to give equal

weight to the SBP and DBP.

3. **Construction of feature vectors and normalisation:** Each new vital sign (feature) was synchronised with the values of the remaining features, using the zero-order hold (ZOH) procedure, i.e. when data were missing, a feature was held out for up to 40 seconds, except for SDA, which is held for up to 30 minutes. Each vital-sign measurement, x, is then normalised to have approximately the same dynamic range, using a zero-mean, unit-variance transform, $\hat{x} = \frac{x-\mu}{\sigma}$, where μ and σ denote the mean and the standard deviation of that vital sign, respectively, determined from the training data.

4. **The Kernel Density Estimate model**: The Kernel Density Estimate (KDE), which is a non-parametric approach that makes few assumptions about the form of the data's distribution, was the method selected to build the joint probability function representing the region of "normal" data. It has the following form:

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} K_\sigma \left(\mathbf{x} - \mathbf{x}_n\right) \tag{6.4}$$

where K_σ represents the Gaussian (or Radial Basis Function, RBF) kernel defined in equation 5.3, σ represents the standard deviation of each of the N, D-dimensional, Gaussian components, located at $\mathbf{x}_n$. The density is obtained by summing the contributions of a Gaussian kernel at each data point $\mathbf{x}_n$. σ was set using the following heuristic estimator $\hat{\sigma}$ (Bishop, 1994):

$$\hat{\sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{L} \sum_{i=1}^{L} \|\mathbf{x}_n - \mathbf{x}_i\| \right) \tag{6.5}$$

where $L = 10$ and $\mathbf{x}_i$ is the ith closest feature vector to $\mathbf{x}_n$.

The KDE approach has the advantage that there is no computation involved in the "training" phase because the training set is used to define the joint Probability

Density Function (PDF). However, it has the disadvantage that the computational cost of evaluating the density grows linearly with the size of the dataset. To mitigate this problem, a sparse kernel density estimate model was created, in which the 2.4 $\times 10^5$ vectors in the training set were reduced to a subset of 500 cluster centres using the K-means algorithm. The 100 cluster centres furthest from the centroid of the 500 centres were discarded, thereby retaining the most "normal" 400 cluster centres.

5. **Novelty score**: In the test phase a short-term median filter is applied to the physiological variables to remove artefacts (five-seconds for HR, SpO_2, and TEMP, and nine-seconds for RR). During the data synchronisation step, a historical five-minute median filter is maintained to impute vital-sign values for up to 30 minutes so that novelty scores can still be determined in cases of missing data. The training set mean value for that vital sign is used for periods of missing data beyond that. The zero-mean, unit-variance normalisation is applied as in the training phase, using μ and σ determined from the training set.

   The likelihood estimates $p(\mathbf{x})$ are difficult to interpret in the clinical setting, and hence a Patient Status Index (PSI, the novelty score), was used to show how far a test vector $\mathbf{x}$ deviates from the normal model:

$$z(\mathbf{x}) = \log\left(\frac{1}{p(\mathbf{x})}\right) - \log\left(\frac{1}{p_{max}(\mathbf{x})}\right) = -\log\left(\frac{p(\mathbf{x})}{p_{max}(\mathbf{x})}\right) \tag{6.6}$$

   where $p(\mathbf{x})$ is the likelihood of $\mathbf{x}$, and $p_{max}(\mathbf{x})$ is the maximum value of $p(\mathbf{x})$. The $log(p_{max}(\mathbf{x})) = 6.08$, was estimated by gradient ascent on the KDE baseline model, in order to scale (most) of the novelty score between 0, representing the "normal" vital-sign data, and 6, representing the most "abnormal" vital-sign data. Conversely, "abnormal" data have low $p(\mathbf{x})$, and therefore high PSI, $z(\mathbf{x})$. The novelty score is only computed in cases for which at least three physiological parameters are available, out of HR, RR, SpO_2, TEMP and the SDA features.

6. **Alert generation:** alerts are generated by choosing a decision boundary that separates the "normal" from the "novel" regions, and a persistence criterion to filter intermittent artefact, as follows,

   (a) **Decision boundary:** The vital signs were considered "novel" if they produced a $z(\mathbf{x}) \geq 3$, as this threshold was shown to be effective to capture single-channel and multiple-parameter (i.e. considering multiple vital signs) abnormal patterns.

   (b) **Persistence criterion:** In the baseline data-fusion system an alert is only generated if the novelty score (i.e. $z(\mathbf{x})$) exceeds the decision boundary (threshold) for a total of four minutes in the current five-minute window of data. The alert is disabled if the novelty score decreases below the threshold for a total of two minutes in the current three-minute window of data.

As discussed in chapter 5, the presence of the oxygen mask is an important indicator of patients at risk of being escalated to the resuscitation room in the ED. In the next sections we describe the mixed data KDE and the OSVM models, that allow the inclusion of clinical data pertinent to the ED patient context, and are adequate ML approaches for the scoring of observational and continuous vital-sign data.

6.3.2 Kernel Density Estimate with mixed data

The KDE method can also be used for estimating the joint PDF defined over mixed categorical and numerical data using the concept of "generalised product kernel" (Racine, 2008). Consider $\mathbf{X}_c \in \mathbb{R}^D$, and $\mathbf{X}_d \in \{0,1\}^E$ representing the collection of D-dimensional numerical (real) variables, and the E-dimensional binary variables, respectively; each collection with the same number of N data examples. Let $k_c(\cdot)$ and $k_d(\cdot)$ be the univariate kernel functions, and let $K_c(\cdot)$ and $K_d(\cdot)$ be the product kernel functions for the real and discrete variables, respectively:

$$K_c(\mathbf{x}, \mathbf{x}_i|\sigma) = \prod_{j=1}^{D} k_c(x_{:j}, x_{ij}|\sigma), \tag{6.7}$$

where $\mathbf{x}_i$ denotes the i^{th} row of $\mathbf{x}$, with $i = 1, ..., N$, $x_{:j}$ the j^{th} column of $\mathbf{x}$, with $j = 1, ..., D$, σ is the smoothing parameter (or kernel width), and

$$K_d(\mathbf{x}, \mathbf{x_i}|\lambda) = \prod_{j=1}^{E} k_d(x_{:j}, x_{ij}|\lambda), \tag{6.8}$$

with $j = 1, ..., E$, and λ is the smoothing parameter, in this case. The RBF kernel is an exemplar $k_c(\cdot)$ kernel (equation 5.3), and we use the estimator kernel proposed by (Aitchison and Aitken, 1976), as an exemplar $k_d(\cdot)$ kernel,

$$k_d(x_{:j}, x_{ij}) = \begin{cases} 1 - \lambda & \text{if } x_{:j} = x_{ij} \\ \lambda/(d' - 1) & otherwise, \end{cases} \tag{6.9}$$

where d' is the number of categories of x, and $\lambda \in [0, (d' - 1)/d')]$. Using the general product kernel, the joint PDF $p(\mathbf{x}_c, \mathbf{x}_d|\sigma, \lambda)$ can be estimated by

$$\hat{p}(\mathbf{x}_c, \mathbf{x}_d|\sigma, \lambda) = \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x_i}|\sigma, \lambda), \tag{6.10}$$

where $K(\mathbf{x}, \mathbf{x_i}|\sigma, \lambda) = K_c(\mathbf{x_c}, \mathbf{x_{c_i}}|\sigma) K_d(\mathbf{x}_d, \mathbf{x}_{d_i}|\lambda)$.

In Pimentel (2015) a product of two kernels was used to model the joint PDF of mixed continuous and discrete vital-sign data, an isotropic multivariate RBF kernel, i.e. with one bandwidth parameter σ for all dimensions of $\mathbf{x}_c$ (e.g. HR, RR, SpO_2, and SBP), and an isotropic multivariate binomial kernel, with the same value of λ for all the dimensions of $\mathbf{x}_d$ (e.g. APVU and FiO_2 support). In our work we will allow for some correlation between the continuous vital-sign variables, i.e. an isotropic K_c as well, but a different λ_j for each categorical variable j, making use of equation 6.8 to represent their product and then equation 6.10, to estimate the joint PDF for mixed data.

The σ and λ_j parameters, can usually be found using four general approaches, (1) reference rule-of-thumb, (2) plug-in methods, (3) cross-validation, and (4) bootstrap methods (Racine, 2008). In our work the σ and λ_j of a mixed data joint PDF model are found by maximising the leave-one-out likelihood of the data (a cross-validation based approach),

$$J(\sigma, \lambda) = \frac{1}{N} \sum_{j=1}^{N} \log \left(\frac{1}{N-1} \sum_{i=1, i \neq j} K(\mathbf{x}, \mathbf{x_i} | \sigma, \lambda) \right). \tag{6.11}$$

The initial value for σ, σ_0, is given by applying the bandwidth heuristic estimator from equation 6.5, and the initial value for each λ_j, λ_{0j} , is given by the plug-in method, appropriate to the Aitchison and Aitken (1976) kernel, described in Chu et al. (2015), which, for $d = 2$ (and $n = N$), $\hat{\lambda}_j^{d=2}$ is given by,

$$\hat{\lambda}_j^{d=2} = \frac{1}{2} \left[1 + \frac{n \left(\frac{1}{2} - p(x_j = 1) \right)^2}{p(x_j = 1)(1 - p(x_j = 1))} \right]^{-1}. \tag{6.12}$$

6.3.3 One-class Support Vector Machines

To describe this approach, we first consider a binary classification problem, in which each point $\mathbf{x_i}$ is associated with a class label $y_i \in \{-1, 1\}$, indicating the class membership. The SVM algorithm can create a non-linear decision boundary by projecting the data through a non-linear function ϕ to a feature space of higher dimension, i.e. points which could not be separated linearly, in their original space $\mathbb{R}^D$, are transformed to feature space $\mathbb{F}$ where they can be separated by a linear hyperplane. The hyperplane in $\mathbb{F}$ is parametrised by $\mathbf{w}^T\mathbf{x} + b = 0$, with $\mathbf{w} \in \mathbb{F}$ and $b \in \mathbb{R}$, determining the margin between the classes $y = -1$ and, $y = +1$, which data points appear in opposite sides of the hyperplane.

To obtain a good generalisation a maximal margin between the classes is considered, to avoid similar test examples, lying very close to the hyperplane, crossing to the wrong side, and being given the incorrect label. To this end, the SVM training algorithm ensures that $y(\mathbf{w}^T\mathbf{x} + b) \geq 1$ for all training examples $\mathbf{x}$. In addition, to prevent over-fitting to

noisy data, and create a soft margin, slack variables $\xi_i \in \mathbb{R}^N$ are introduced to allow some points to lie on the "wrong side" of the decision boundary, and a constant $C > 0$ penalises these "misclassifications"; higher values of C result in less smooth decision boundaries in an attempt to classify points correctly. Thus, the objective function of the SVM classifier minimises:

$$\begin{aligned} minimise & \frac{\| \mathbf{w} \|^2}{2} + C \sum_{i=1}^{N} \xi_i \\ & \text{s.t. } y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \, \forall i \\ & \text{and } \xi_i \geq 0, \, \forall i. \end{aligned} \tag{6.13}$$

The minimisation process is typically solved using Lagrange multipliers, and the decision function rule for a data point $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^{N} \alpha_i y_i k\left(\mathbf{x}, \mathbf{x}_i\right) + b \right) \tag{6.14}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers, and $k\left(\mathbf{x}, \mathbf{x}_i\right) = \phi\left(\mathbf{x}\right)^T \phi\left(\mathbf{x}_i\right)$ is known as the kernel function. The latter exploits Mercer's theorem, which shows that if $k(\cdot)$ is a kernel function, i.e. symmetric and positive semi-definite, then there exists some Reproducing Kernel Hilbert Space $\mathbb{F}$ that can be expressed as an inner product $k\left(\mathbf{x}, \mathbf{x}_i\right) = \phi\left(\mathbf{x}\right)^T \phi\left(\mathbf{x}_i\right)$. Popular choices for the kernel function are the linear, polynomial, sigmoidal, and the Gaussian RBF (equation 5.3).

In the OSVM approach, Schölkopf et al. (2001) 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 defines regions in the input space where the PDF of the data has most support. The objective function is adapted from the

previous equation 6.13 as,

$$\begin{aligned} &\text{minimise } \frac{\| \mathbf{w} \|^2}{2} + \frac{1}{\nu N} \sum_{i=1}^{N} \xi_i - \rho \\ &\text{s.t. } (\mathbf{w} \cdot \phi(\mathbf{x}_i)) \geq \rho - \xi_i, \forall i \\ &\text{and } \xi_i \geq 0, \forall i \end{aligned} \tag{6.15}$$

where ν is the parameter that defines the smoothness of the boundary (i.e. defines the proportion of training observations that lie on the "wrong" side of the hyperplane, and is a lower bound on the number of training examples used as support vectors), with $\rho \in \mathbb{R}$ as an offset parameter of the hyperplane. The decision function becomes:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \phi(\mathbf{x}_i) - \rho) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i k(\mathbf{x}, \mathbf{x}_i) - \rho \right) \tag{6.16}$$

Thus, the OSVM creates a hyperplane characterised by $\mathbf{w}$ and ρ, which has maximal distance from the origin in the feature space $\mathbb{F}$, separating all the data points from the origin.

6.3.4 Application of novelty detection approaches to patient condition monitoring

The baseline KDE model was proposed to monitor continuous vital-sign data from patients recovering in the step-down unit wards. Current research is evaluating its feasibility (and generalisation) for different hospital wards. One way to make this clinical decision support tool more patient-specific is to consider that patients can be divided into subpopulations, and re-compute the model using data from the ward context of the patient, as different hospital wards are concerned with different patient conditions, which may present different patterns of normality and physiological deterioration. In addition, various studies have compared the baseline model with other state-of-the art novelty detection ML approaches, relevant to our work, such as the OSVM, which we summarise next.

In Wong (2011), the Weighted Kernel Density Estimate (WKDE) and the OSVM algorithms were used to estimate models of normality for the continuous (bedside monitor) vital signs of patients recruited in the ED pilot study (a 472 patients study, summarised in section 2.4.1). In the WKDE approach, each kernel applied to the centroids selected by the k-means clustering is weighted by the number of points assigned to the cluster of that centroid. This model performed marginally better than the baseline KDE (where the centroids have equal weights) to identify patient deterioration before their escalation in the ED (TPR was 51.7% and 41.3%, respectively, in a smaller subset of 217 no-event and 29 event patients with continuous vital-sign data), at the cost of an increase of the false alert rate in their continuous vital-sign data (specificity was 78.8% versus 79.7%, respectively). A patient-wise performance analysis, which we describe later in section 8.3.4, was used to assess the model performance. When cross-validation was used to re-train both the KDE and an OSVM models, the latter showed superior performance (SEN and SPEC were 69.0% and 69.6% versus 58.6% and 73.3%, respectively).

Clifton et al. (2014) analysed the use of the baseline KDE approach, a Gaussian Mixture Model (GMM) and OSVM to score continuous vital-sign time-series, which included both observational and continuous data, to detect physiological deterioration in patients discharged to the post-operative ward of the Cancer Centre, Oxford University Hospitals NHS Foundation Trust, UK, after upper-gastrointestinal cancer surgery. In the novelty detection GMM approach, the normal data are assumed to be distributed according to a mixture of M multivariate Gaussian distributions $p(\mathbf{x}) = \sum_{i=1}^{M} \pi_i p(\mathbf{x}|\boldsymbol{\theta}_i)$, with a prior probability π_i and a likelihood $p(\mathbf{x}|\boldsymbol{\theta}_i) = N(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent the mean and covariance matrix for the multivariate Gaussian i, respectively. The maximum likelihood estimates of the model parameters can be determined by using the expectation-maximisation (EM) algorithm (Dempster et al., 1977), using data from the stable patients population. As in the KDE model, test data with low likelihood are then considered novel (abnormal). The OSVM approach presented a higher mean partial

AUROC[2] to that of GMM and KDE (partial AUROCs were 0.28 ± 0.03, 0.24 ± 0.02 and 0.26 ± 0.01, respectively) in identifying physiological deterioration before an escalation to the ICU.

Finally, we give the example of the work of Pimentel (2015), in which a KDE and an OSVM model of normality that included numerical and categorical data were trained, in order to include categorical variables such as the presence of FiO_2 support and the level of consciousness (APVU scale), to better identify physiological deterioration in post-operative cancer patients from the previous example. These models were only applied to the patients' observational data, as both of these categorical parameters are not available in continuous vital-sign data, as in the case of our ED dataset. Although the KDE and OSVM models, re-trained from mixed continuous-categorical data, increased the AUROC over the baseline KDE model (AUROC was 0.818 for the KDE model re-trained using a cross-validation approach, versus 0.722 for the baseline KDE model), they were outperformed by the KDE and OSVM models trained without the categorical variables (e.g. AUROC was 0.848 for the continuous variable only KDE models). The OSVM approach also performed better in this work, with an AUROC of 0.857.

6.4 Gaussian Processes for time-series modelling

Clifton et al. (2012) proposed that observational and continuous vital-sign data, used in the hospital environment to track the patient physiology, could be represented by a Gaussian Process (GP) regression model, as it could provide a better way to estimate vital-sign data in periods of artefactual or missing vital-sign data. In the next sections we review the GP framework, and its use in modelling vital-sign time-series in the hospital environment.

Time-series usually consist of continuous measurements x_t that can be sampled at any point in time. Discrete time data are typically modelled, probabilistically, using

[2]The partial AUROC metric considers only a sub-region of the AUROC, in this case to indicate which TPR and FPR values are clinical relevant.

either auto-regressive (AR) or state-space approaches. AR approaches are more useful when only forecasting is important and finding the latent state is not (e.g. prediction of future financial observations), while state-space approaches are used in cases when the true, unobserved, state of the system, or latent state, is evolving over time and can only be observed indirectly (e.g. tracking the location of objects through the Global Positioning System). AR models model the next output of a system as a function of a number of previous outputs.

$$x_t = f(x_{t-1}, ..., x_{t-\tau_x}) + \delta_t, \tag{6.17}$$

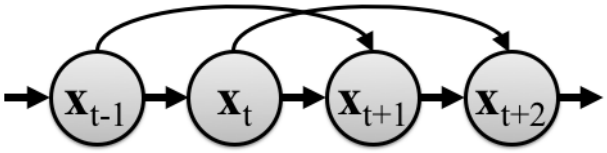
where $x_t \in \mathbb{R}^D$ are output vectors, τ_x specifies the AR model order, and δ_t represents random noise that is i.i.d. across time. A state-space model introduces latent variables states $x_t \in \mathbb{R}^D$, and is defined by the state transition function f and the measurement function g,

$$r_{t+1} = f(r_t, \mathbf{u}_t) + \mathbf{v}_t \tag{6.18}$$

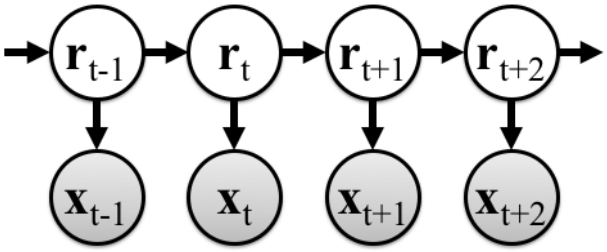
$$x_{t+1} = g(r_t, \mathbf{u}_t) + \mathbf{e}_t \tag{6.19}$$

where $\mathbf{v}_t$ and $\mathbf{e}_t$ represent additive noise known as the process noise and the measurement noise, respectively, and $\mathbf{u}_t$ is a set of control inputs. Linear dynamical systems (e.g. the Kalman filter) and Hidden Markov Models are examples of state-space models with continuous and discrete valued latent state, respectively.

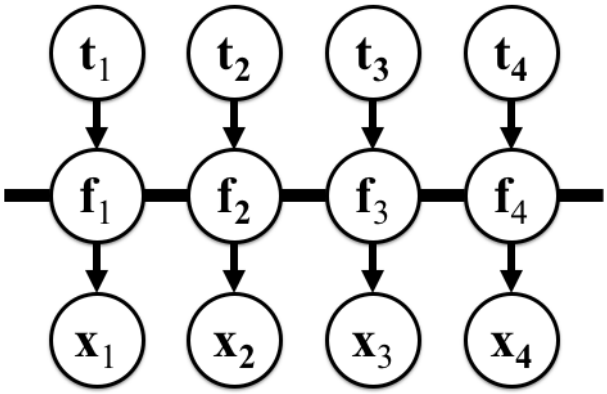
Gaussian processes are a class of stochastic processes that perform inference directly over the space of functions (a prior over the functions, Rasmussen and Williams (2006)), contrasting with models of functions defined by a parametrised class of functions and a prior over the parameters (Frigola-Alcade, 2015). In this thesis we are concerned with modelling the vital-sign time-series $\mathbf{x}$, which we consider the latent variable, estimated at time inputs $\mathbf{t}$. We therefore adapt the GP nomenclature to this problem statement, as follows.



(a) *Second order auto-regressive model.*



(b) *State-space model.*



(c) *Gaussian Process regression model.*

Figure 6.2: *Graphical model for auto-regressive (order two), state-space and Gaussian process time-series modelling approaches. These graphical models show the difference in conditional independence assumptions between the three approaches. The grey shading signifies an observed node (variable) while a white node signifies a latent variable. In the GP case the function values f are fully connected, as indicated by the bold bar, and are dependent on $\mathbf{t}$. The observations x are f plus some observation noise. Please note that the bold bar is non-standard in the graphical models notation. This representation is used in Turner (2011) and Frigola-Alcade (2015). Also note that while the input data needs to be ordered and evenly sampled in (a) and (b), the order of data is defined by the input values ($\mathbf{t}$) in (c), and the input data can be unevenly sampled in this case.*

A function $f : \mathbb{T} \to \mathbb{R}$ is distributed according to a GP if and only if $p(f(t_1, ..., f(t_N)))$, the density of that function's values at any N points $t_i \in \mathbb{T}$, is a multivariate Gaussian. This enables it to be specified as:

$$\mathbf{x} = f(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), k(\mathbf{t}, \mathbf{t}_*)), \tag{6.20}$$

where $m(\mathbf{t}) \in \mathbb{R}^D \to \mathbb{R}$ and $k(\mathbf{t}, \mathbf{t}_*) \in \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ represent the mean and covariance functions, respectively. Hereafter, the index set is assumed to correspond to the time (at which vital-sign observations are measured). Figure 6.2 illustrates the differences between AR, state-space, and GP models, in graphical model representation.

GP prior specification

The prior mean is typically set to zero: $m(\mathbf{t}) = 0$, and the covariance function or kernel, $k_{\boldsymbol{\theta}}(\cdot)$, encodes the "similarity" between inputs t_i and t_j, with possibly some Gaussian observation noise. The kernel is described by hyperparameters $\boldsymbol{\theta}$, which induce general properties of the GP functions, such as smoothness, input scale and output scale.

There exists a large class of well-suited covariance functions (see chapter 4 of Rasmussen and Williams (2006)), the most commonly used being the Squared-Exponential, SE (also known as exponentiated-quadratic, or Gaussian kernel function):

$$k_{SE}(t_i, t_j) = \theta_h^2 \exp\left\{-\frac{d^2}{2\theta_l^2}\right\}, \tag{6.21}$$

where $d = \|t_i - t_j\|$, and $\boldsymbol{\theta} = \{\theta_h, \theta_l\}$ are hyperparameters modelling the typical amplitude of deviation from the mean, and typical time-scale on which the function varies, respectively. In the previous section, the OSVM kernel was applied to our vital-sign data values $\mathbf{x}$. In this section $\mathbf{x}$ represents the latent vital-sign variable, being estimated by the GP, from inputs $\mathbf{t}$, representing the observation times (the covariance function kernel being applied to the latter). The SE covariance function is said to be *stationary* because it only depends on the difference in $\|t_i - t_j\|$, rather than on their absolute value. As seen

before, for the OSVM case, the kernel covariance functions have to fulfil Mercer's theorem, meaning that they have to be symmetric and positive semi-definite; hence $k_{SE}(\cdot)$[3] is a valid kernel.

GP posterior calculation

Conditional on observed data, point-wise predictions can be made about the function values $\mathbf{x}_*$ at any "test" location of the index set $\mathbf{t}_*$. The posterior density for a test point $\mathbf{t}_*$ is Gaussian,

$$\mathbf{x}_*|\mathbf{t}_*, \mathbf{x}, \mathbf{t} \sim \mathcal{N}\left(\hat{\mathbf{x}}_*, var(\mathbf{x}_*)\right), \tag{6.22}$$

where the mean and variance are given by (assuming the mean function $m(\mathbf{t})$ to be zero):

$$\hat{\mathbf{x}}_* = k(\mathbf{t}_*, \mathbf{t}_*)^T k(\mathbf{t}, \mathbf{t})^{-1} \mathbf{x}, \tag{6.23}$$

$$var(\mathbf{x}_*) = k(\mathbf{t}_*, \mathbf{t}_*) - k(\mathbf{t}, \mathbf{t}_*)^T k(\mathbf{t}, \mathbf{t})^{-1} k(\mathbf{t}, \mathbf{t}_*). \tag{6.24}$$

To facilitate our notation we use $\mathbf{K} = k(\mathbf{t}, \mathbf{t})$ to encode the covariance between training inputs $\mathbf{t}$; the covariance function $\mathbf{K}_* = k(\mathbf{t}, \mathbf{t}_*)$ to represent the covariance between the test inputs $\mathbf{t}_*$ and the training inputs $\mathbf{t}$; and $\mathbf{K}_{**} = k(\mathbf{t}_*, \mathbf{t}_*)$ to define the prior variance of $\mathbf{t}_*$.

GP prediction using noisy observations

In realistic modelling scenarios we only have access to noisy versions of the functions we are trying to model. Thus, a noise parameter σ, representing Gaussian noise, is usually considered to model noisy datasets (assuming additive i.i.d. Gaussian noise). Adding this term changes the mean and covariance functions, equations 6.23 and 6.24, respectively, to (assuming, again, the mean function $m(\mathbf{t})$ to be zero):

[3]The sub-index in k_{SE} is used to identify the type of covariance function, and this acronym strategy is used for the remaining kernels.

$$\hat{\mathbf{x}}_* = \mathbf{K}_*[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}\mathbf{x} \tag{6.25}$$

$$var(\mathbf{x}_*) = \mathbf{K}_{**} - \mathbf{K}_*[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}\mathbf{K}_* \tag{6.26}$$

where $\mathbf{I}$ is the identity matrix.

Hyperparameter optimisation

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

$$-\ell = -\log p(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{x}^T\mathbf{K}^{-1}\mathbf{x} + \frac{1}{2}log|\mathbf{K}| + \frac{N}{2}log(2\pi), \tag{6.27}$$

where N is the number of observations. The NLML can be interpreted as a cost function, in which the first term penalises low data likelihood (low data fitness) and the second term penalises model complexity (or model capacity). This function can be minimised using using gradient descent. In a full Bayesian treatment, the hyperparameters should be integrated out, but this often leads to analytically intractable models, and sampling methods or other approximations are usually used to estimate the integrals. In the light of recent research an alternative, more advantageous, optimisation procedure is introduced next.

6.4.1 Bayesian Parameter Estimation and Optimisation

Bayesian optimisation (BO) is a useful strategy for finding the minimum of an objective function that does not have a closed-form expression, when the optimisation problem is non-convex.

The BO is summarised in Algorithm 1, and illustrated in Figure 6.3. The exemplar

Algorithm 1 Bayesian Optimisation Algorithm

1: $\boldsymbol{\Theta} = \{(\boldsymbol{\theta}_{init}, g(\boldsymbol{\theta}_{init}))\}$
2: **for** $i = 1, 2, ..., N$ **do**
3: Estimate $\mu(\hat{\boldsymbol{\theta}})$ and $\sigma(\hat{\boldsymbol{\theta}})$ from the GP posterior $\tilde{g}(\hat{\boldsymbol{\theta}})$, computed over $\boldsymbol{\Theta}$
4: Estimate $u(\hat{\boldsymbol{\theta}})$
5: Sample $g(\boldsymbol{\theta}^*)$ at $\boldsymbol{\theta}^* = \text{argmax}_{\hat{\boldsymbol{\theta}}}\ u(\hat{\boldsymbol{\theta}})$
6: Augment the data $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}, (\boldsymbol{\theta}^*, g(\boldsymbol{\theta}^*))\}$
7: **end for**
8: optimal solution $\boldsymbol{\theta} = \text{argmin}_{\boldsymbol{\theta}^*}\ g(\boldsymbol{\theta}^*)$

target function we wish to minimise in this figure is:

$$g(\theta) = -\left\{\left(\exp(-(\theta-2)^2\right) + \exp\left(\frac{-(\theta-6)^2}{10}\right) + \frac{1}{\theta^2+1}\right\}. \tag{6.28}$$

The objective function is modelled as a GP, i.e. $g(\boldsymbol{\theta}) \sim \mathcal{GP}(\mu(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}'))$. At each iteration the GP posterior distribution, $\tilde{g}(\hat{\boldsymbol{\theta}})$, approximates the unknown objective function, at candidate points $\hat{\boldsymbol{\theta}}$, using observations from $\boldsymbol{\Theta}$ (Figure 6.3.a, Algorithm 1, line 3). The most promising candidate parameters $\hat{\boldsymbol{\theta}}$ for evaluation are selected via an acquisition function $u(\hat{\boldsymbol{\theta}})$ that makes use of the objective function mean function $\mu(\hat{\boldsymbol{\theta}})$, and its uncertainty $\sigma(\hat{\boldsymbol{\theta}})$, derived from the posterior (Figures 6.3.c, and 6.3.e, Algorithm 1, line 4). The next best query sample from the objective function is at the parameter $\boldsymbol{\theta}^*$, that maximises $u(\hat{\boldsymbol{\theta}})$ (Figures 6.3.a, Algorithm 1, line 5). The new sampled point, observation $(\boldsymbol{\theta}^*, g(\boldsymbol{\theta}^*))$, is accumulated to the set $\boldsymbol{\Theta}$ (Algorithm 1, line 6), and the algorithm loops up to the maximum number of iterations. The optimal $\boldsymbol{\theta}$ is that minimising the objective function evaluated at the sampled points (Algorithm 1, line 8).

With this approach the expensive objective function is only evaluated at the sampled locations $\boldsymbol{\theta}^*$, and the Gaussian process is relatively easy to evaluate at each iteration, if the number of observations remains small.

GP Prior

In Gardner et al. (2014a), a squared-exponential kernel with a vector of automatic relevance determination (ARD) hyperparameters, was used as the prior over the objective

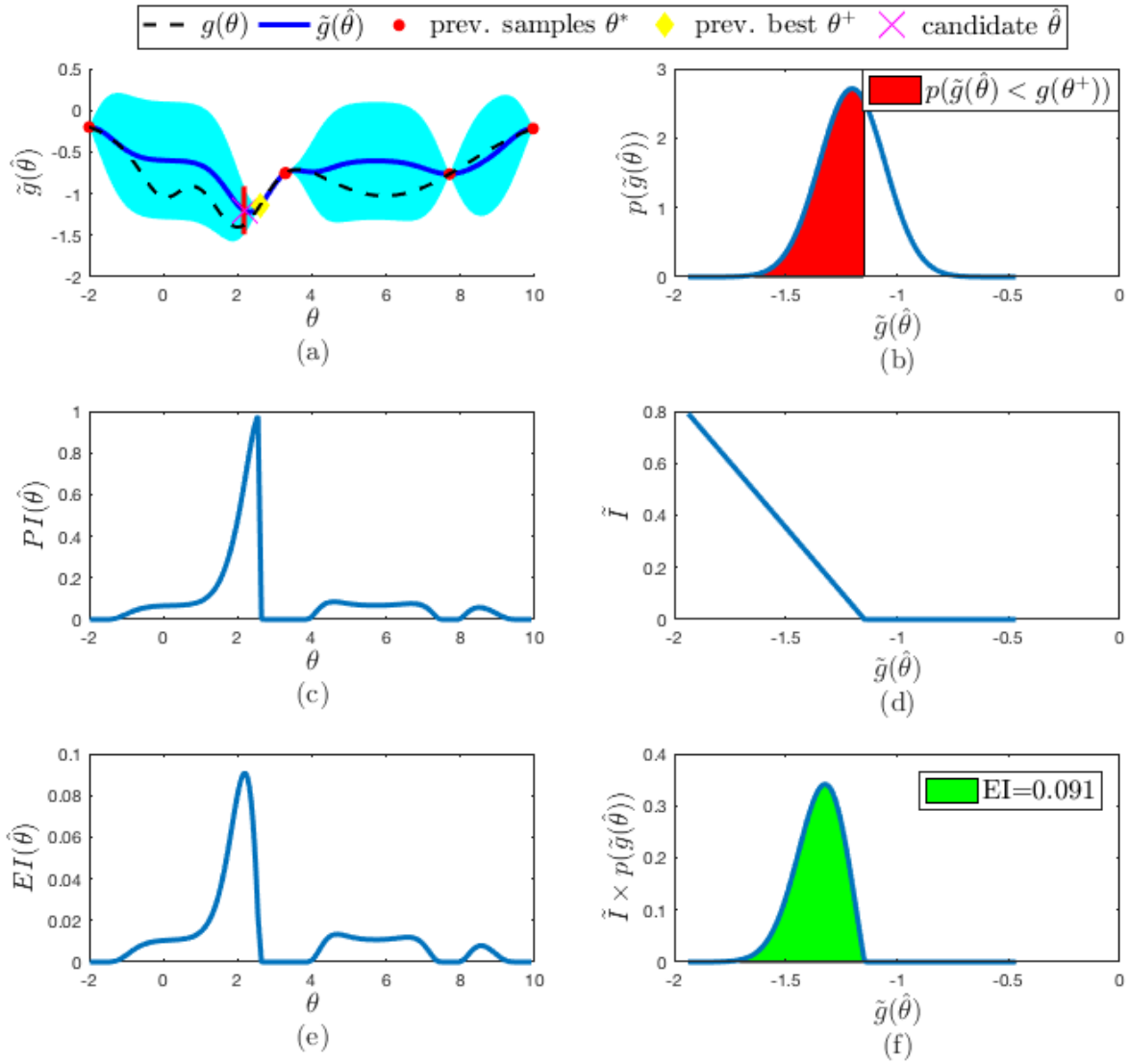


Figure 6.3: *This BO illustration is adapted from Colopy et al. (2017), and formulated for finding the minimum of an objective function, the approach used in our work. In (a) a GP is fitted to 5 observations, and its posterior $\tilde{g}(\hat{\theta})$ is shown. The latter is used to approximate the objective function $g(\theta)$, in the dashed black line, equation 6.28, which is unknown. The current optimal minimum is represented by the yellow point at $g(\theta^+) = -1.149$. The next query point, the magenta cross, is selected such that it maximises an acquisition function such as the Probability of Improvement, PI, in (c), or the Expected Improvement, EI, in (e), when compared to the current best point. Please note that in our configuration of this problem, maximising the acquisition function corresponds to finding the candidate minimum of our cost function $g(\theta^+)$. The components of each acquisition function, evaluated at $\theta^* = 2.172$, are shown in (b), (d) and (f). The posterior marginal distribution of $\tilde{g}(\theta^* = 2.172)$ is marked in red in (a) and plotted in (b). The probability that this point will be lower than the current best value (as estimated by the GP) is the area under the curve (AUC) in red (i.e. the integral below $g(\theta^+) = -1.149$). The improvement $\tilde{I}$ over the current best value versus any possible realisation of $\tilde{g}(\theta^* = 2.172)$ is shown in (d). $EI(\hat{\theta})$ is the AUC of the product of (b) and (d), shown in green in (f).*

function. This kernel is defined as (Rasmussen and Williams, 2006):

$$k_{SE\text{-}ARD}(\boldsymbol{\theta}, \boldsymbol{\theta}') = c_h^2 \exp\left\{-\frac{1}{2}\sum_{m=1}^{M}\frac{(\boldsymbol{\theta}_m - \boldsymbol{\theta}'_{l_m})^2}{\nu_m^2}\right\}, \tag{6.29}$$

where the hyperparameter c_h is the signal variance and the length-scales hyperparameters $\nu_{m=1}^{M}$ are allowed to be different for each feature (sub-index m). For most applications we wish to automatically remove those parameters that have small contributions to the model (small θ). However, in the work of this thesis, it is advantageous to use the SE-ARD hyperparameters in BO objective function prior, to correctly model the relationship between hyperparameters with very different units (such as the vital-sign length-scales, and signal-variances).

Acquisition Function

Let $\boldsymbol{\theta}^+$ be the best point in sample set $\boldsymbol{\Theta}$ evaluated thus far. The improvement of $\hat{\boldsymbol{\theta}}$ is defined as the decrease of $\tilde{g}(\hat{\boldsymbol{\theta}})$, a Gaussian random variable, against $g(\boldsymbol{\theta}^+)$, which due to the GP model, is itself a random quantity:

$$\tilde{I} = \max\{0, g(\boldsymbol{\theta}^+) - \tilde{g}(\hat{\boldsymbol{\theta}})\}. \tag{6.30}$$

Figure 6.3d shows an example of the improvement evaluated for a candidate point. The Expected Improvement acquisition function, EI, is the expectation over this truncated Gaussian variable, defined as $EI(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\tilde{I}(\hat{\boldsymbol{\theta}})|\hat{\boldsymbol{\theta}}\right]$, which can be computed analytically:

$$EI(\hat{\boldsymbol{\theta}}) = \begin{cases} \sigma(\hat{\boldsymbol{\theta}})(Z\Phi(Z) + \phi(Z)) & if \quad \sigma(\hat{\boldsymbol{\theta}}) > 0 \\ 0 & if \quad \sigma(\hat{\boldsymbol{\theta}}) = 0, \end{cases} \tag{6.31}$$

$$\text{with: } Z = \frac{\mu(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}^+)}{\sigma(\hat{\boldsymbol{\theta}})}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF of the standard normal distribution, respectively, and $\mu(\hat{\boldsymbol{\theta}})$ and $\sigma(\hat{\boldsymbol{\theta}})$, are the mean and standard deviation derived from the posterior $\tilde{g}(\hat{\boldsymbol{\theta}})$. The EI function balances the trade-off between exploiting and exploring. When exploring, points where the GP variance is large are chosen (global search), and when exploiting, points where the GP mean is low are selected (local optimisation, for a minimisation task). Figure 6.3e shows an example of the $EI(\hat{\boldsymbol{\theta}})$ for the BO iteration, and compares it with an alternative, simpler, acquisition function, the Probability of Improvement, $PI = \Phi\left(\frac{\mu(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}^+)}{\sigma(\hat{\boldsymbol{\theta}})}\right)$, shown in Figure 6.3c.

6.4.2 Application to vital-sign time-series modelling

A preliminary description of the 2-stage ML based system illustrated in the introductory Figure 6.1 is described in Clifton et al. (2012). In this work the nurse's observation sets and the beside monitors continuous data were mixed into one data-stream to train patient-specific GP regression models for each vital sign of 200 patients recovering from upper-gastrointestinal cancer surgery in a Oxford University Hospitals NHS Foundation Trust, Oxford, postoperative ward. Ten-fold cross-validation was used to set the signal variance and length-scale hyperparameters of the Squared Exponential covariance function for each vital sign of each patient. The GP regression model was then used to smooth the vital-sign data of each patient, which provided a lower mean absolute error between the model's estimates, and the test data, than using (i) the population mean or (ii) the patient-specific mean of each vital-sign channel. Better estimation of the vital-sign time-series led to early detection of abnormal (or novel) patterns in seven patients in the dataset [4]. The novelty detection model (a multivariate extreme value theory model) was built from data from this ward, making the model more specific to these patients. However, the time-series model used all the observations to generate the time-series (offline estimation), and the authors suggested an online estimation implementation in future work, so it would be

[4]The total number of abnormal patients is not provided, however the authors referred that it was statistically insignificant, given the incidence of ICU readmission.

appropriate to be used in real-time settings. The single-task GP time-series model used in the previous work was extended in Dürichen et al. (2015), with the use of a Multi-Task Gaussian Processes (MTGP). This model learns the cross-correlation between pairs of vital-sign SE covariance functions, each with a different length-scale parameter. The premise was that information about the HR may help generate a better estimate of RR, including in periods of missing RR data, each data-channel varying with specific dynamics. We note that in this latter work only continuous vital-sign data was used to demonstrate the benefits of the model.

More recently, Alaa et al. (2016) developed a personalised risk score algorithm, based in the GP framework, to monitor patient condition on hospital wards. First, physiological data streams (i.e. observational HR, RR, SpO_2, TEMP, SBP, and lab-tests data) from the stable hospital population were modelled using a mixture of MTGPs (a mixture of experts, each encoding a relevant time-series feature). The MTGP hyperparameters, for a number of mixtures, were learned using the EM algorithm and the number of mixtures M were learned using the Bayesian Information Criterion. Stable patients were used to allow learning stationary MTGP covariance functions, since those of deteriorating patients are less likely to be stationary. Then transfer learning was carried out, using linear regression to map the MTGP mixture to the patients' admissions features (age, sex, race, ethnicity, transfer status, stem cell transplantation, and admission unit). The M mixture model configurations were learned from the stable patients and the admissions features mapping were learned for the unstable patients. At test time t, a risk score is computed for every GP expert pair (stable and unstable patient versions) in the mixture, and the average of these scores is set as the final risk score. In this work the PPV and the TPR metrics were used to assess the alerting performance for abnormal vital signs prior to the patient's unscheduled ICU admission. The personalised risk score from Alaa et al. (2016) outperformed the MEWS, the Rothman index (a logistic regression model for physiological risk scoring of ICU patients), and the LASSO[5] logistic regression model.

[5]Least Absolute Shrinkage and Selection Operator.

Colopy et al. (2016) used a GP model to perform change-point detection on continuous HR data from ICU patients discharged to the step-down unit wards. The objective of their model was to detect physiological deterioration before a clinical emergency event. In this approach a GP was fitted to a 7-hour window of patient data, and subsequently advanced every 5 minutes. 60 points for the last hour and 120 data points from 1 to 7 hours before the forecast window, were used to fit the GP with zero mean function and a covariance kernel modelled as:

$$k_{MAT(5/2)+SE+WN}(t_i, t_j) = \theta_{h_1}^2 \left(1 + \frac{d\sqrt{3}}{\theta_{l_1}}\right) \exp\left(-\frac{d\sqrt{3}}{\theta_{l_1}}\right) + \theta_{h_2}^2 \exp\left(\frac{-d^2}{2\theta_{l_2}^2}\right) + \theta_{WN}^2 \delta(t_i, t_j) \tag{6.32}$$

where $\theta_{MAT(5/2)+SE+WN} = [\theta_{h_1}, \theta_{l_1}, \theta_{h_2}, \theta_{l_2}, \theta_{WN}]$ refers to the hyperparameters of the final covariance function, the sub-indexes $MAT(5/2)$ and WN refer to the $Matérn(5/2)$ and Gaussian noise (or white noise, identified with sub-index WN also, in the respective hyperparameters) covariance functions, respectively. This kernel indicates that short-term variations in HR are governed by the twice-differentiable $MAT(5/2)$ kernel, while hourly variations in HR are governed by the smooth SE kernel, and the measurements are corrupted by noise with a Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

The GPs were fitted using Markov Chain Monte Carlo (MCMC). In this case the HR scoring function is the mean of the log likelihoods of data in a 5-min forecast window. Note that in this case the score is personalised as it is learned online (sequentially) from the HR data collected from each patient. The performance metric was defined as the Time-to-event (TTE) versus FPR, i.e. the best model is the one that detects abnormal HR at an earlier TTE for those patients with events, and at the lowest FPR for those patients without events. The performance of this method in detecting patient HR deterioration was superior to the reference EWS system and comparable to the baseline KDE method (described in section 6.3.1).

Colopy et al. (2017) improved the performance of the previous GP-forecast model by using Baysean Optimisation. In their most recent work, patient-specific hyperparameters

for the covariance functions encoding short ($MAT(3/2)$) and long-term (SE) dynamic changes in HR data, were optimised. The hyperparameter optimisation process was based on maximising the 15-min forecast log marginal likelihood by iteratively computing the GP posterior, over the accumulated objective function values, and using the $EI(\cdot)$ acquisition function to select the next best point up to 200 iterations. The GP model used in the BO objective function was specified with the $MAT(3/2)$-ARD covariance function to allow learning hyperparameters, for the forecast GP, operating in substantially different units, e.g. the length-scale, and the signal variance operated in minutes and log(beats per minute), respectively.

6.5 Conclusion

Recent research on ML methods applied to in-hospital patient vital-sign monitoring, to identify physiological deterioration, has been reviewed in this chapter. Three main topics are being investigated in this area: (i) the development of time-series models that represent the patient's physiological data, which may be considered latent, i.e. they may be corrupted by artefacts or periods of missing data, that are correctly re-estimated by the model; (ii) patient-specific models to identify the risk of physiological deterioration, trained from data collected from the patient's clinical context, and (iii) appropriate performance metrics to select the optimal algorithm configuration for this task.

Chapter 7

Evaluation of the continuous data-fusion system intervention

7.1 Introduction

Failure to act on or recognise deterioration, has been found to be the cause of in-hospital patient mortality for 5.2% of 1000 adults who died in 2009 in 10 acute hospitals in England (Hogan et al., 2012), and 35% of 2000 deaths reported from 1 June 2010 to 31 October 2012 in the NHS database of patient safety incidents (Donaldson et al., 2014). The aim of continuous data-fusion systems is to improve patient surveillance and safety in critical hospital settings, such as High-Dependency Units and the Emergency Department, as early identification of physiological deterioration in these settings should prompt early treatment by skilled clinical staff in the expectation of changing the deterioration trajectory back to normal physiology.

Acutely-ill adult patients from the ED majors area are connected to bedside monitors which sample their vital signs continuously. Relevant alerts generated from the outputs of these monitors between nurses' observations should indicate when the patient physiology deteriorates. In this chapter, we analyse the clinical staff response to the alerts triggered by an FDA-approved data-fusion system, Visensia, connected to the outputs of

the patient bedside monitors in use during phase 3 of the large-scale ED study.

7.2 Previous clinical studies using Visensia

Watkinson et al. (2006) conducted a randomised clinical trial that compared the use of continuous (electronic) monitoring of vital signs for reducing the frequency of adverse events in high-risk patients outside of critical care areas with standard ward care. An early version of Visensia was used to record the continuous monitoring data. The novelty scores of the patients in the continuous electronic monitoring group were assessed retrospectively by two senior clinicians once patients had completed the study. The software was not used to alert nurses. 94% of 690 transitions from normal to abnormal physiological activity were considered true episodes. The development of an abnormal physiological episode into a major event could be predicted using Visensia's recording of vital signs, with a sensitivity of 63% and specificity of 52%.

Sen et al. (2009) reported a retrospective analysis of 117 patients admitted to a level 1 trauma centre over 6 months. Vital-sign data were collected both pre-hospital (e.g. trauma triage) and from the ED and the novelty score calculated retrospectively. The study found that, in the pre-hospital setting, a novelty score over 3 was predictive of patients who needed life-saving interventions (OR 1.8, 95% CI $= [1.1, 4.2]$). The novelty score had statistically significant likelihood ratios for life-saving interventions including endotracheal intubation, blood transfusion, CPR and use of resuscitation drugs.

Hravnak et al. (2011) conducted a prospective, single-centre, before-and-after study in the USA to assess whether the Visensia data-fusion novelty score correlated with single-parameter cardio-respiratory instability concern criteria[1], and whether nurse response to the system's alerts was associated with a reduction in patient instability.

The study had 3 sequential stages. In stage 1 patients had continuous single-channel monitoring and standard care. In stage 2 the novelty score was displayed at the bedside

[1]These criteria were a HR of <40 or >140 bpm, a RR of < 8 or > 36 rpm, SBP of <80 or >200 mmHg, DBP >110 mmHg, and SpO2 of <85%.

and on central monitors, and staff were trained to use the software. In stage 3 staff responded to Patient Status Index alerts using a pre-defined process developed for this purpose. For each stage the incidence and duration of cardio-respiratory instability versus no instability periods were determined.

Stage 3 showed a statistically significant reduction in the average duration of instability episodes per admission, average duration of physiologically plausible instability episodes per admission, and average number of full instability episodes per admission, when compared with stage 1.

Choukalas et al. (2011) described a retrospective cohort study in a mixed medical-surgical-cardiac ICU in an urban tertiary-care hospital, including 20 consecutive patients requiring advanced cardiac life support (ACLS), due to in-hospital cardiac arrest. The novelty scores were calculated at 5-minute intervals for the 20 hours before the cardiac arrest. Six of the patients did not need ACLS care. For the remaining 14, the mean lead-time of the PSI alert before the cardiac arrest was 15.1 hours. Nurses documented patient instability an average of 9.3 hours before the cardiac arrest.

Choukalas et al. (2015) also reported a retrospective controlled cohort study in an 18-bed ICU in an urban hospital, including 61 patients who had a cardiac arrest while in hospital and 729 controls. Novelty scores were calculated at 1-minute intervals for the 24 hours before cardiac arrest. The study found that there was no difference in novelty scores between the two groups at the beginning of the observation period, but that the score became significantly higher for patients with cardiac arrest starting from 10 hours prior to the cardiac arrest (p-value < 0.05). The study used a version of the Visensia software without pre-set alert levels and which is not currently commercially available.

In summary, the studies show the performance of the Visensia to detect physiological instability prior to cardio-respiratory adverse events or escalation procedures on hospital wards.

7.3 Technical alerts

Following the Hravnak et al. (2011) study, Hravnak et al. (2015) analysed the use of machine learning approaches to remove artefacts present in non-invasive vital-sign high frequency data streams. These artefacts can cause alert events which are not directly correlated with patient physiology, defined as "technical alerts" ("physiological alerts", if the alert is related with abnormal physiology"). In this study, continuous vital-sign data (HR, RR, SpO_2 and BP) were recorded from 634 admissions to a step-down unit ward. Data were divided into Block 1, as the training/cross-validation set, and Block 2, the test set. Expert clinicians annotated Blocks 1 and 2 events as real or artefact. After feature extraction, algorithms were trained to create and validate models automatically classifying events as real or artefact. The models were tested on Block 2 data.

Block 1 yielded 812 alert events (with a persistence requirement, such as that described in section 6.3.1), with 39% judged by experts as artefact (RR 43%, SpO_2 40%, BP 15%, HR 2%), and Block 2 yielded 1521 alert events, with 40% judged by experts as artefact (RR 28%, SpO_2 58%, BP 13%, HR 1%). The best performing algorithm applied to Block 2 test data had an AUROC of 0.94 for RR artefacts using the Random Forest algorithm. The equivalent figure for BP artefacts was 0.84 (using Logistic Regression), and 0.72 for SpO_2 artefacts (using the Naive Bayes algorithm). HR did not have enough examples to build a model.

This study shows that the number of artefacts present in continuous data captured from the bedside monitors, is high, especially for the RR and SpO_2 data-streams (Figure 7.1 shows and exemplar SpO_2 related artefact from this study).

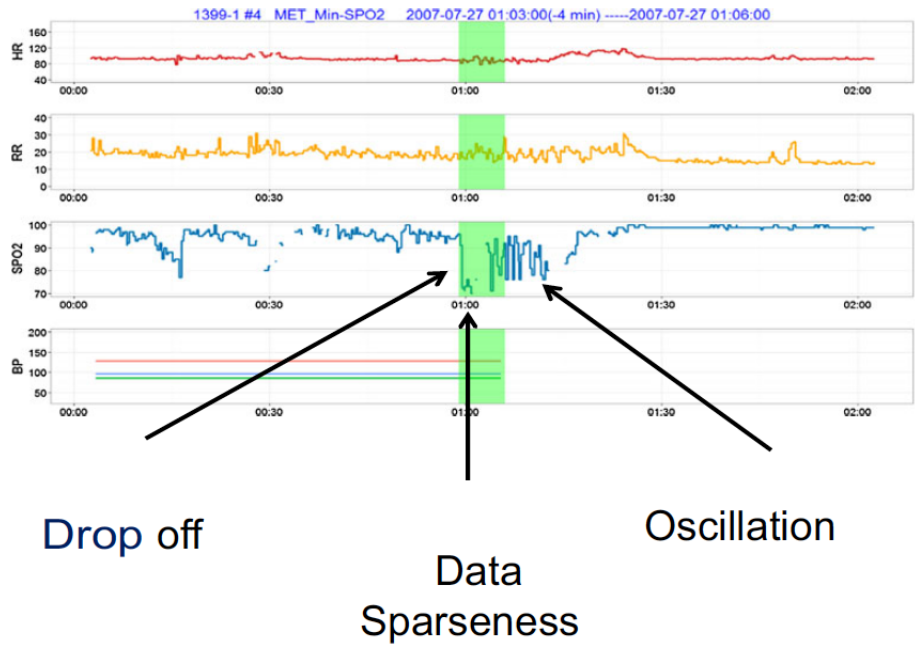


Figure 7.1: *Technical alert example from Hravnak et al. (2015). Three types of artefacts are present: (i) SpO_2 probe drop-off; (ii) as a consequence the SpO_2 is intermittent for a 5-minute period (data sparseness); and (iii) as the staff check the equipment on the patient, oscillatory patterns appear as a consequence of patient movement (as reported in the paper).*

7.4 Evaluation of the data-fusion system

7.4.1 Data preprocessing

All patients with full documentation from phases 2 and 3 were used in this analysis, i.e. groups C_2 (3,113 ED admissions) and C_3 (3,204 ED admissions) in the consort diagram in Figure 3.4.

Observational data

Application of the EWS: The nurses' physiological measurements were processed as described in section 4.2.2, and their CEWS re-computed.

Application of the PSI: The baseline KDE model was used to compute likelihood estimates (novelty scores) and alert, on each observation set, as described in section 6.3.1 (no persistence criterion was required in this case). In this case each feature vector included HR, RR, SDA, SpO_2 and TEMP. The missing values were set to their mean value

in the training data, 83.8 bpm, 18.3 rpm, 94.7 mmHg, 95.2 %, and 36.0 ^{o}C, respectively, as reported in Tarassenko et al. (2005).

Continuous data

We use superscript prime to distinguish the observation from the continuous vital-sign data acronyms. E.g. HR and HR$'$ refer to the observational (i.e. nurses' physiological measurements) and the continuous (i.e. from the bedside monitors) heart rate data, respectively, and the same nomenclature is applied to the remaining vital signs, thereafter.

Application of the EWS: First, the vital signs were synchronised by holding out HR$'$, RR$'$ and SpO_2' for 40 seconds and BP$'$ for 30 minutes at each new vital-sign value that becomes available on any of the four channels. Then, the synchronised data were then down-sampled to one-minute windows by taking the median of the data in each window. Finally, the CEWS (Table 3.1) was computed for each one-minute window.

PSI$_1$ model: The baseline KDE model output by the Visensia modules is referred to as PSI_1 model, thereafter. The PSI and the alerts were generated by the Visensia modules, from the continuous vital-sign data collected from each bedside monitor, described in section 6.3.1.

Visensia alerts: There were three alert states implemented in the Visensia data-fusion system: no alert (alert state = 0), when the novelty score was low; alerting (alert state = 1), when the persistence criterion was met; and silenced (alert state = 2), when the system was set to pre-defined periods (15-min, the default 30-min, or 1-hour windows), during which the alerts were not audible. The silencing functionality ensures that the alert sound is off as clinical staff intervene to treat the abnormal condition identified by the data-fusion system.

Technical alerts: Continuous data from phase 3 were further analysed, and a set of heuristics was designed to identify technical alerts. This was conducted using the following steps:

1. A set of rules to identify physiological or technical alerts were initially formalised

after analysis of the alert periods in the data from phase 3.

2. As clinical staff reported technical alerts due to SpO_2 related artefacts were most common, the initial heuristics were used to find the first 100 physiological alerts, and the first 100 SpO_2 related technical alerts;

3. Two research nurses independently annotated these alerts, using the coding system in Table 7.1, and when the annotations differed, consensus was achieved by a research assistant providing a further annotation in those cases.

4. The set of heuristics was improved after review of the annotated data, and then applied to the rest of the continuous data from phase 3. The final rules can be seen in Table 7.1. In this table, PA_i and TA_i refer to the i^{th} type of physiological or technical alert codes, respectively.

Figure 7.2 (top) shows an example of the intermittent (nurse observations) and continuous vital-sign data measured for one patient from phase 3. The bottom plot shows the PSI (or the *novelty score*) determined at each newly acquired vital sign (magenta line), and the correspondent alert state (black line). This notation is used for the remainder of the thesis.

The labelling process for the 200 alerts resulted in 49% technical alerts (42% for SpO_2', 4.5% for RR′, 0.5% for BP′, 2% for data hold-out related artefacts). Most artefacts were associated with the SpO_2' data, as expected, as the labelling process conducted by the clinical experts, was set to identify those. Cohen's (unweighted) κ coefficient (Cohen, 1960) is a statistic which measures inter-rater agreement for qualitative (categorical) items, and was used to evaluate inter-rater agreement in labelling the technical alerts. Only two labels were used for this calculation, the technical alerts versus the physiological alerts labels. Agreement was classified as follows: $\kappa < 0$ is less than chance agreement, $\kappa \in [0.01, 0.40]$ is fair agreement, $\kappa \in [0.41, 0.80]$ is moderate agreement, and $\kappa \geq 0.81$ is good agreement (Viera et al., 2005). The κ coefficient between the expert nurses who labelled the data was 0.84 (good agreement).

Table 7.1: *Rules and labels used to annotate physiological and technical alerts. SDA - Systolic-Diastolic Average. PSI - Patient Status Index (the novelty score).* [a] *CEWS criterion (Table 3.1).*

Alert type	Cause	Labels	Rules
Physiological	Low SDA	PA_1	SDA $<$ 41 mmHg.
	High SDA	PA_2	SDA $>$ 185 mmHg.
	Low SpO_2	PA_3	$SpO_2 \leq 84\%$ [a].
	High RR	PA_4	RR $\geq$ 34 rpm [a].
	Low RR	PA_5	RR $\leq$ 7 rpm [a].
	High HR	PA_6	HR $\geq$ 128 bpm [a].
	Low HR	PA_7	HR $\leq$ 42 bpm[a].
Technical	SpO_2	TA_1	- High change rate ($>$ 10% SpO_2 in 30-sec). - Difference between nurse observation within 10 min of the alert and median of 5-min continuous data before the alert $>$ 10% SpO_2. - Extreme low SpO_2 values ($<$ 85%) for at least 50% of the 5-min window, with sharp return to normal values ($>$10% SpO_2 in 1-min) within 5-min of the alert or no data after the alert due to probe disconnection.
	RR	TA_3	- High change rate ($>$ 12 rpm in 30-sec) and duration of RR abnormality is $<$ 4-min. - Difference between nurse observation within 10-min of alert and median RR 5-min window before the alert of $>$ 10 rpm. - Duration of extreme RR ($<$ 3 rpm or $>$ 44 rpm) is $\geq$ 2.5-min.
	HR	TA_6	- High change rate ($\geq$ 40 bpm in 30-sec) and duration of HR abnormality is $<$ 4-min. - Duration of extreme HR (HR $<$ 30 bpm) is $>$ 3-min.
	BP	TA_8	BP outside of physiological ranges (see Table 4.1).
	Other	TA_2 TA_4	Absent PSI: PSI $<$ 3 for $>$20% of the 5-min window before the alert. # channels $<$ 3: PSI values not present when $<$ 3 channels exist.
	Disconnected from monitor	TA_5	At most 1 data point per channel during the 5-min window previous to the alert.
Silenced	-	TA_7	Occurrence of an alert state = 2 without an alert state = 1 in the previous timestamp.

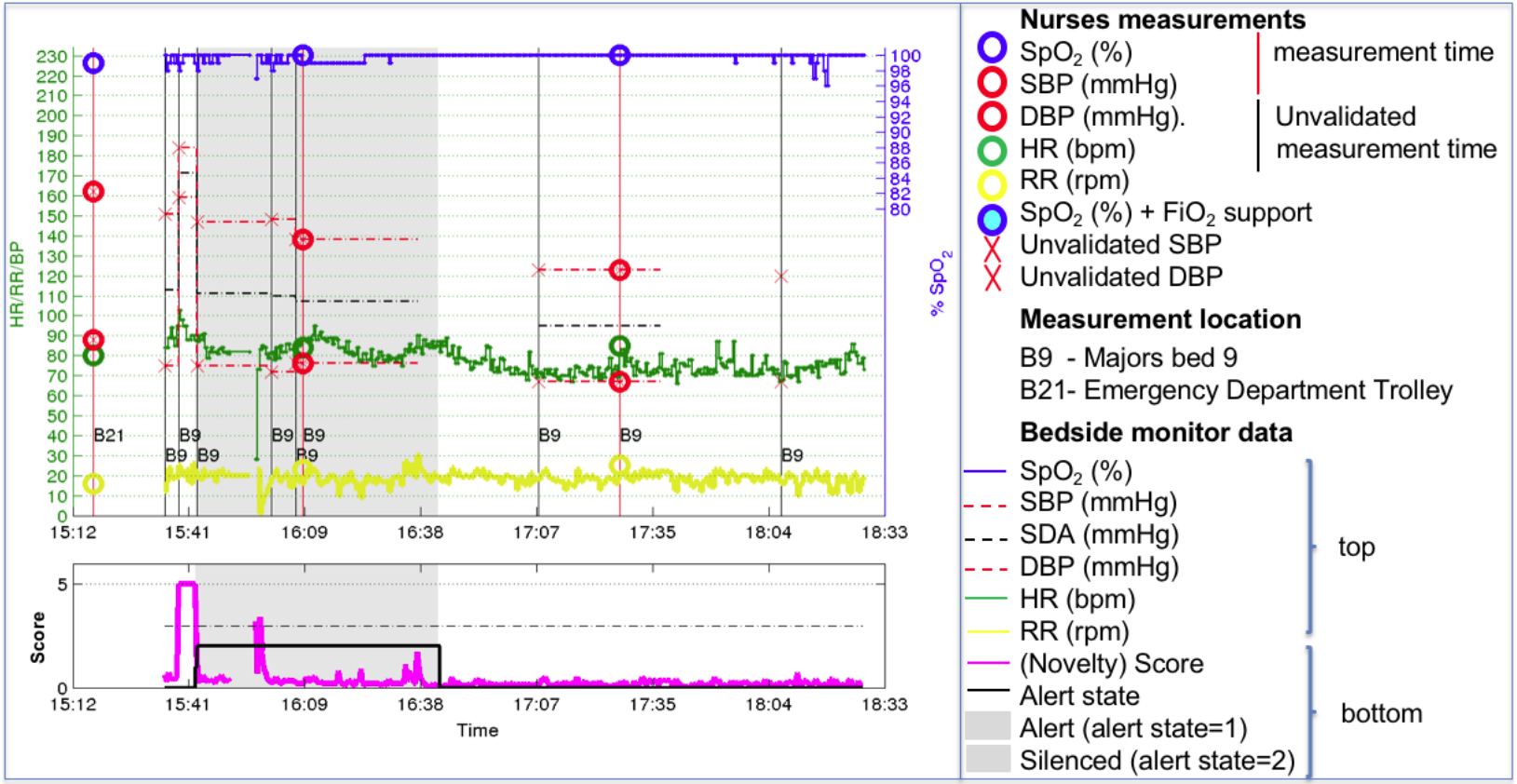


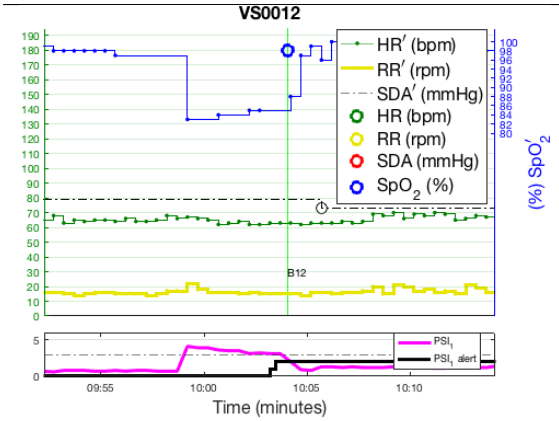
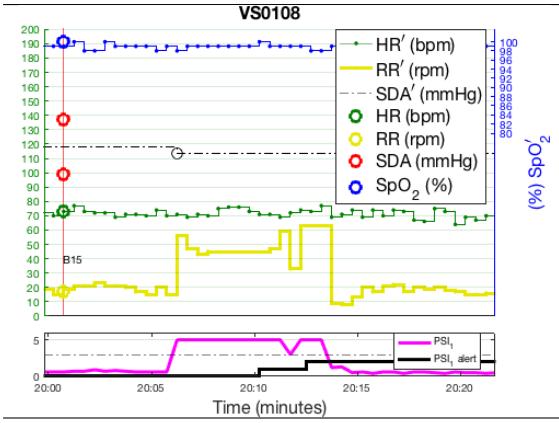
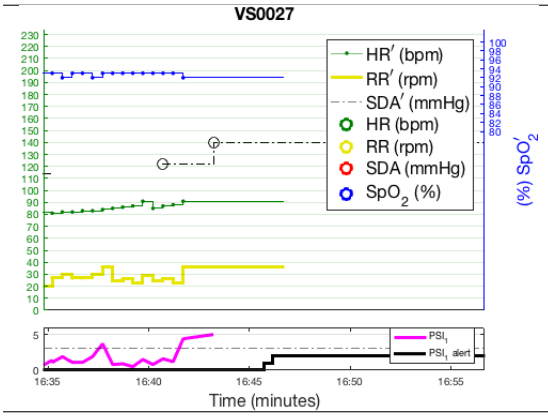
Figure 7.2: *Example of observational and continuous data for one patient from phase 3. One alert generated by the Visensia algorithm can be seen in the plot at the bottom. This alert was immediately silenced (after 2 min) by clinical staff, and hence it is followed by a 30-min period in which Visensia was silenced. The GCS (Glasgow Comma Scale) and the temperature are not shown in this plot, as they are used in EWS but not in this version of the PSI generated by the data-fusion system. The unvalidated observations (black vertical solid lines) are explained in section 3.1, and result from VitalPAC showing the physiological measurements from the bedside monitor at BP measurement times.*

Examples of the rules derived for the final automatic labelling process, and of the resulting technical alerts, are shown in Table 7.2. The first example results from a fast decrease of more than 10% SpO_2' in a 30-sec window, which occurred when the probe attached to the finger became loose enough to bias the estimate for 4 minutes, which is the time required to generate an alert in the data-fusion system. Staff then fixed the probe, and a fast recovery in the SpO_2' can be seen within 5 minutes of the alert. The other two examples in the same table are related to artefacts on the RR′ signal, and artefacts created by holding-out noisy RR′ data resulting from disconnecting the patient from the bed side monitor (while holding-out HR′ and SpO_2' in parallel).

Alerts that did not present PSI scores higher than 3, for 4-min, or for which at least 3 channels of data were available in the 5-minute window before the alert, were also considered technical alerts (coded as type "other", and labelled as TA_2 and TA_4, respectively, in Table 7.1).

The optimised heuristics (Table 7.1) yielded an accuracy of 97.5% when applied to the labelled data. The κ coefficient between the computed results and the labelled data was 0.95.

Table 7.2: *Three rules to identify technical alerts. See Table 7.1 for the full technical alert codes.*

SpO_2 artefact example (TA_1)	
	Loss of signal caused by loose probe: (i) step increase/decrease of $> 10\%$ SpO_2' in less than 30-sec; (ii) discrepancy between obs. and cont. data within 10-min of the alert window data, of $> 10\%$ SpO_2'.
Respiratory rate artefact example (TA_3)	
	Poor impedance signal causing: a step increase and decrease, or instantaneous oscillatory step increase up and down of $RR' > 15$ rpm.
Sensor disconnection example (TA_5)	
	At most 1 data point per channel during the 5-min window previous to the alert.

7.4.2 Physiological instability

The physiological instability in phase 3 was evaluated on observational data using three metrics (also used in the analysis conducted in chapter 4): (i) the percentage of patients in each "patient acuity" group; for this metric, CEWS was calculated for each observation set and the patients were categorised by their maximum CEWS, in four acuity groups, $EWS_{max} = 0$, $EWS_{max} = 1$, $EWS_{max} = 2$ and $EWS_{max} \geq 3$; (ii) the percentage of patients with at least two observations per hour, in each "patient acuity" group; (iii) the percentage of observation sets in each "patient acuity" group. Histograms were used to compare these metrics between phases 2 and 3.

For the continuous data, physiological instability was assessed by determining all periods with a CEWS ≥ 3. Boxplots were used to compare the median duration of physiological instability between phases 2 and 3, for (a) all patients, (b) no-event and (c) event patients, normalised by the amount of data for each ED admission.

7.4.3 Time from arrival to escalation in the ED

The median time from patient arrival to escalation to the resus area during the ED stay, was compared between phases 2 and 3, for those patients escalated after ED arrival.

7.4.4 Clinical staff response to the data-fusion system alerts

Figure 7.3 shows the most common staff allocation during adult patients treatment at ED of the John Radcliffe Hospital, Oxford. This ward has 20 cubicles, 16 in the majors area and 4 in the resus area. We note that phase 2 differed from phase 3 in the triage process, the latter using nurse assessment, rather than the Manchester Triage System (see chapter 3). The data-fusion system was active on all majors beds except those used for nurse assessment (majors beds 1 to 6), and those in the resus area. To evaluate staff response, one must take into consideration the fact that nurses conduct the physiological measurements required to evaluate patient condition, and that doctors are responsible

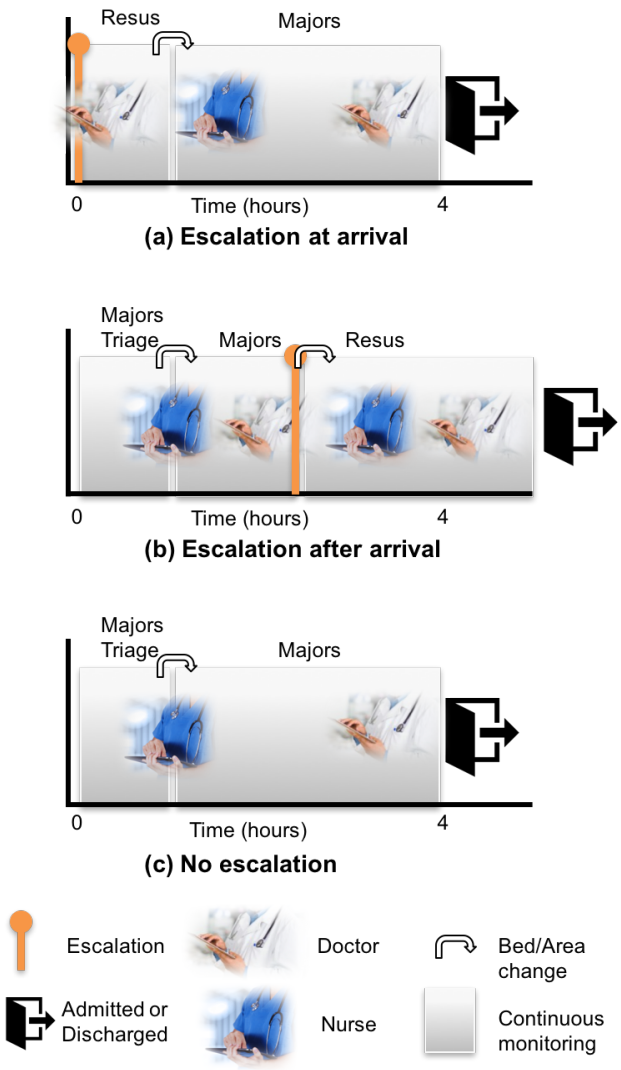


Figure 7.3: *Illustration of staff allocation during an adult patient ED stay in phase 3 of the large-scale study. (a) Patients arriving by ambulance may be admitted to the resus area (four beds in total), and a doctor evaluates their condition in order to initiate treatment. A nurse then evaluates their response to initial treatment, and more stable patients are moved to the majors area to give room to unattended or more severe cases in the resus area. The doctor then returns at a protocolled time to discharge the patient; (b) Non-critical cases are first assessed in the majors triage area, beds 1 to 6 or triage room, by a senior nurse. If patients are deemed at risk of deteriorating they are moved to majors beds 7 to 16, for further monitoring by nursing staff. In case their condition deteriorates they are evaluated by a doctor and they can be escalated to the resus area. While recovering from treatment they are further cared for by a nurse. If their condition does not improve, the doctor admits them to the hospital. (c) Most patients attending an ED and at risk of deterioration move from triage to the majors area, recover under nursing staff care, and are then evaluated by doctors. The ED doctors then decide whether to discharge or admit the patient to a hospital ward, within 4 hours of the patient arrival.*

for initiating an intervention to correct the underlying problem. Therefore, one must consider those cases in which the absence of physiological measurement after an alert may be justified because another type of clinical intervention has taken place.

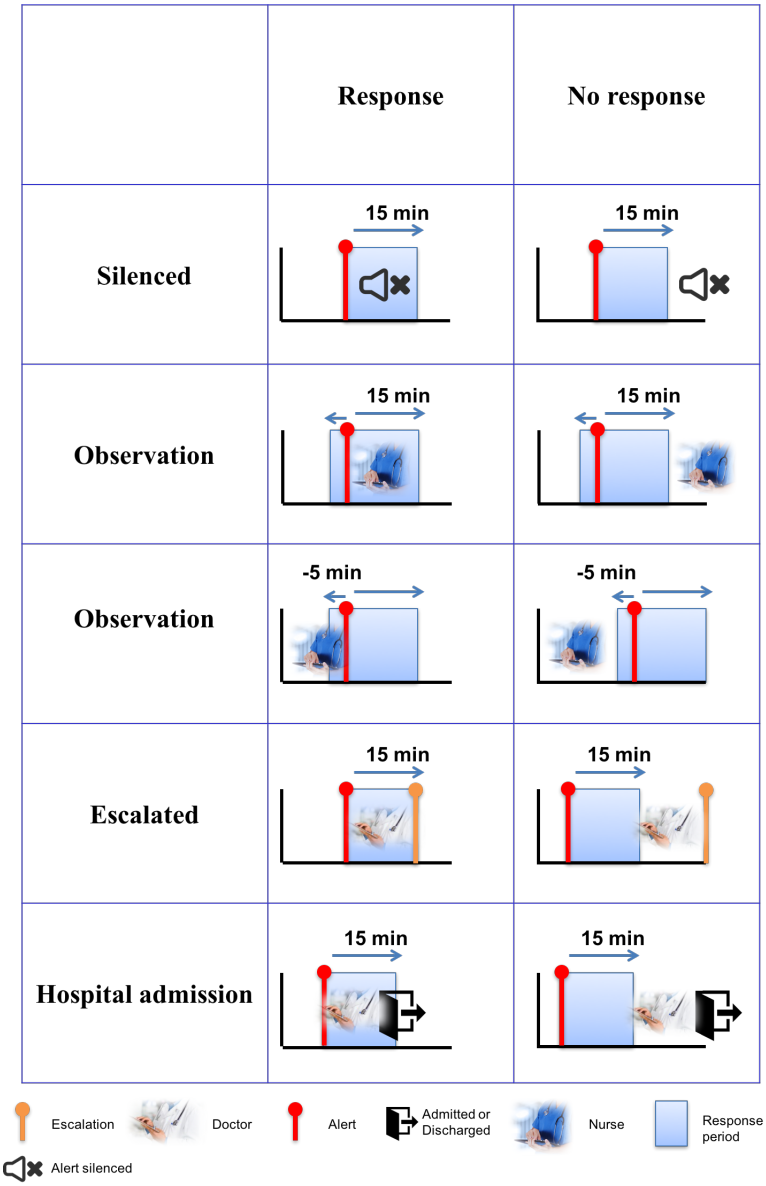


Figure 7.4: *Clinical staff response was considered when (i) the alert was silenced within 15 minutes, there was (ii) an observation set within -5 and 15 minutes from the alert, (iii) the patient was escalated within 15 minutes of the alert, or when (iv) the patient was discharged from the ED within 15 minutes of the alert (home or to the next hospital ward). (ii), (iii) and (iv) were selected by whichever occurred first, and (i) was selected only if (ii), (iii) and (iv) did not occur. Note that physiological measurements are made by nurses, while the decision to escalate/discharge patients is made by doctors.*

Figure 7.4 illustrates the four cases in which the clinical staff response to the data-fusion system alerts during phase 3, were grouped.Taking this model into account, the percentage of the data-fusion system physiological alerts followed by a staff response in phase 3, was analysed. The technical alerts, caused by artefacts, were removed from this metric.

7.4.5 Statistical analysis

The Kolmogorov-Smirnov test was used to evaluate the difference between vital-sign distributions. The Wilcoxon test was used to evaluate the difference between the medians of time-related distributions. The χ^2 test of independence was used to assess the difference between sets of categorical unpaired data.

7.5 Results

Figure 7.5 compares the vital-sign distributions between phases 2 and 3. All, except observational RR data, present statistically significantly different distributions, but note that there are observational and continuous measurements on the order of 10^4 and 10^5 data points, respectively, and the differences are not clinically relevant, for example the median HR in phase 3 is higher than that in phase 2 by 1 bpm. The other vital signs present similar differences.

7.5.1 Physiological instability

Figure 7.6 compares the percentage of (a) all patients, and (b) patients with more than two observations per hour, in each "patient acuity" group, between phases 2 and 3. Phase 3 showed a higher proportion of patients in the critical group ($EWS_{max} \geq 3$, p-value $<$ 0.001) as more observations were scored in this critical group (see Figure 7.7, p-value $<$ 0.001). The number of patients with at least two observations per hour, per group, is comparable between phases 2 and 3. This is, on any given shift, there is a comparable

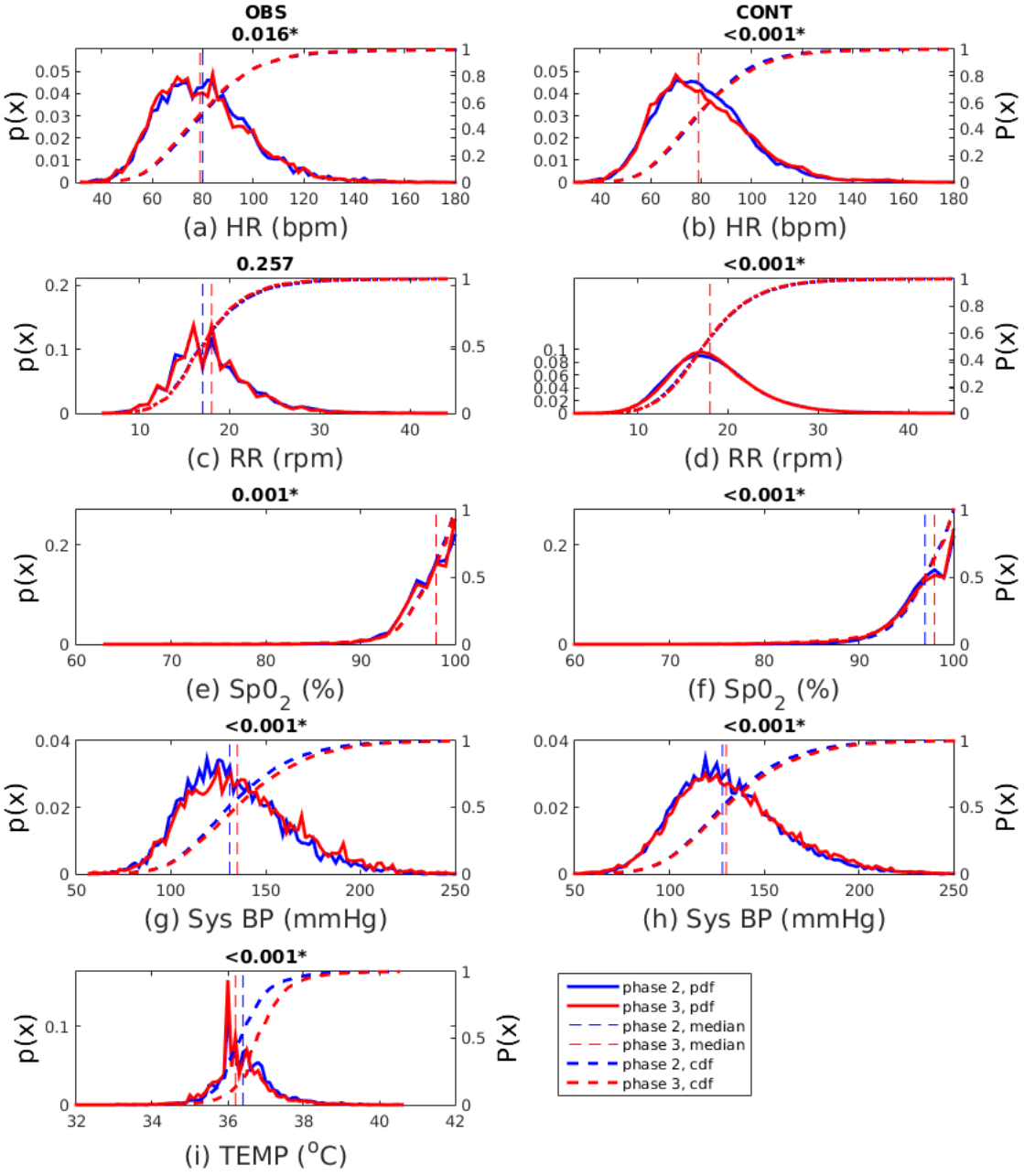


Figure 7.5: *left) PDFs, CDFs and medians of phases 2 and 3 observational data; right) PDFs, CDFs and medians of phases 2 and 3 continuous data; *Indicates the presence of a statistical significant difference between phase 2 and phase 3 CDFs, using the Kolmogorov-Smirnov Test (k-test).*

amount of time available to conduct the observations. The use of Visensia in phase 3 may have guided the nurses to focus more on the critical patients ($EWS_{max} \geq 3$).

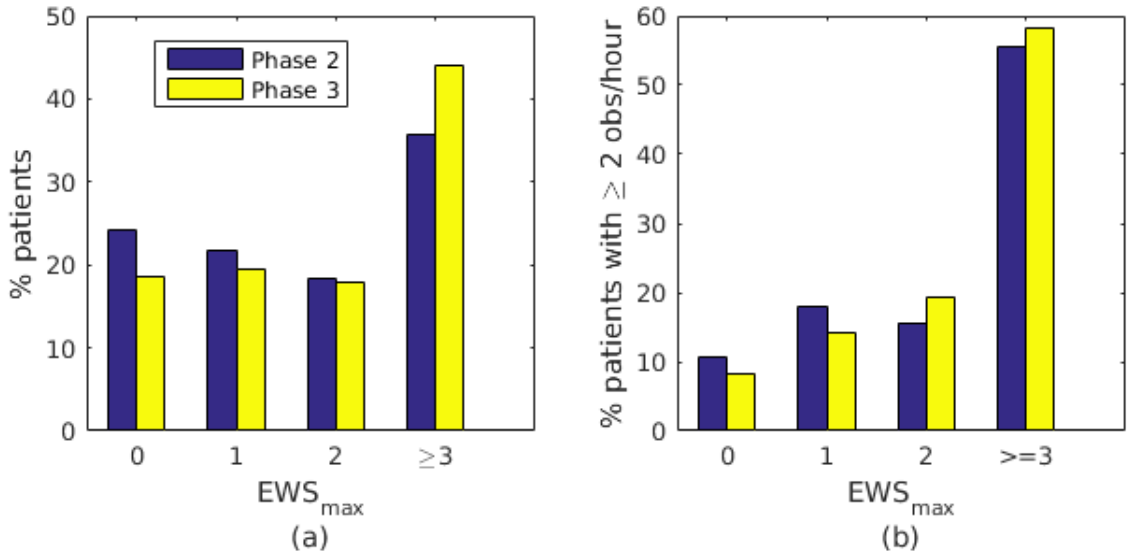


Figure 7.6: *(a) Comparison of the percentage of patients in each acuity group, between phases 2 and 3 (p-value $<$ 0.001); (b) Comparison of the percentage of patients with more than 2 observations per hour in each acuity group, between phase 2 and 3.*

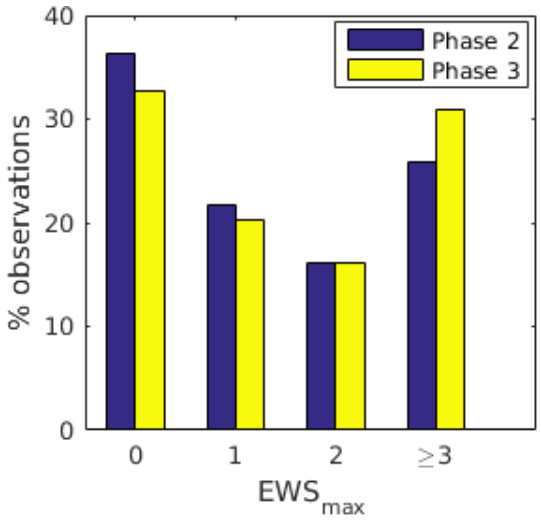


Figure 7.7: *Comparison of the percentage of observations in each acuity group between phases 2 and 3 (p-value <0.001).*

For the continuous vital-sign data, our definition of physiological instability differs from that used in Hravnak et al. (2011), not only in the thresholds used to compute abnormal values for each vital sign, but also in the fact that a multi-parameter score is also used in our study to determine periods of physiological instability. Also the median

duration is analysed in our study, rather than the average duration, as the distribution of duration of physiological instability is highly skewed (see boxplots in Figure 7.8).

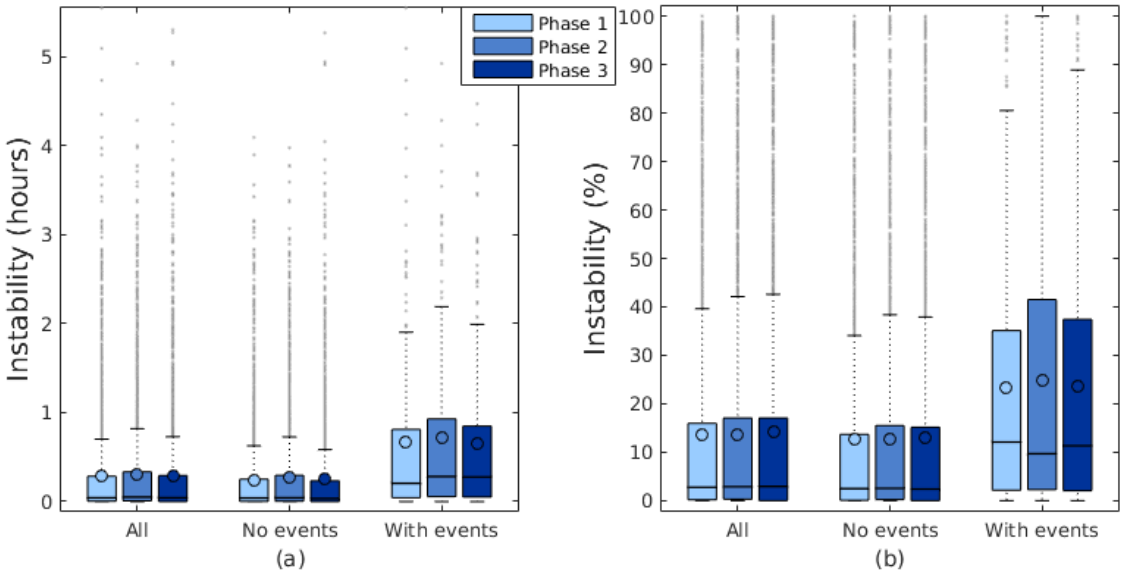


Figure 7.8: *Boxplots of the duration of physiological instability for each phase. (a) in hours; (b) normalised by the amount of data per admission.*

Figure 7.8 compares the distribution of the duration of physiological instability for all, no-events and with-events patients, between phases 2 and 3. For this section, the events included any of (i) death, (ii) cardiopulmonary resuscitation, (iii) unscheduled ICU admission within 30-days of admission, and escalation to resus (iv) at or (v) after ED arrival (191, 221, and 296 documented event patients with continuous data, in total, for phases 1, 2, and 3, respectively).

The median instability time was about three minutes for patients without events, and 12.3, 16.8 and 16.6 minutes, for patients with events for phases 1, 2 and 3 (10% of the continuous data available on average per patient, respectively). When normalising the instability periods by the amount of data available for each patient admission, the median was not significantly different between phases 2 and 3 (or phases 1 and 2), for all the subgroups in the figure (i.e. “All”, “No events” or “With events”).

This result shows that the introduction of the data-fusion system did not alter the duration of physiological instability occurring in adult patients attending this ED. This is not surprising since the major part of the treatment which would correct the physiological

instability is usually given after the ED stay (for example, on hospital admission to the ICU).

7.5.2 Time from arrival to escalation in the ED

Figure 7.9 compares histograms of the times from arrival to an escalation to the resus area in the ED, between phases 2 and 3. The median times to escalation are not significantly different, 1.4 and 1.7 hours for phases 2 and 3, respectively (p-value was 0.083), which would support the idea that the data-fusion system intervention did not alter the time to initiate treatment on patients that deteriorated in the ED.

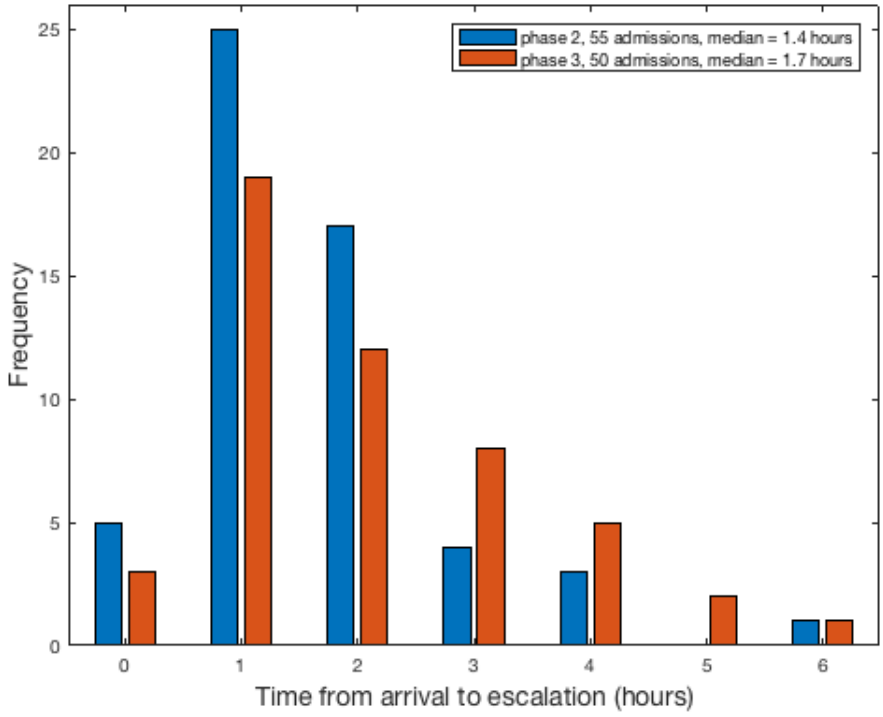


Figure 7.9: *Comparison of the time from arrival to an escalation to resus in the ED, between phases 2 and 3, for those patients with complete documentation (i.e. for the C_2 and C_3 groups in the consort diagram in Figure 3.4). The p-value for the difference between the medians was 0.084 (not significant), and was determined using the two-tail Wilcoxon test.*

Other variables can affect the median time to escalation (after arrival in the ED): the fact that phase 3 was at the end of the year (September to November 2014), on colder months than phase 2 (June to July 2013), may have increased the the number of attendances and the waiting time; also, the change in the triage process between phases

2 and 3 may have increased the length of ED stay for those patients in phase 3 with data-fusion monitoring (undergoing nurse assessment in beds 1 to 6 and then transferred for further assessment to beds 7 to 16).

7.5.3 Clinical staff response to the data-fusion system alerts

Figure 7.10 shows a tree-diagram with the breakdown of the number and types of responses that followed the data-fusion system alerts, within 15 minutes of the alert. Each level can be described as follows:

1. Documented patients: There were 3,204 patients for whom electronic Track and Trigger data was complete in phase 3. A total of 50 (of the documented) patients were escalated to the resus during their ED stay.

2. Patients with active Visensia: For 127 patients it was not possible to assign continuous data, mostly because of periods of system down-time (i.e. the bedside monitors or the data-fusion monitors were not operational, or shut down). For 179 patients there were no continuous data in active Visensia beds (7 to 16), i.e. they either remained on beds 1 to 6 (the triage beds) during their ED stay, or were escalated from the triage beds directly to the resus area, or returned from resus to the triage beds. For 50 patients there were less than 5 minutes of data, which might have occurred because these patients were being evaluated because of symptoms such as the level of consciousness, which is not captured by continuous monitoring data. The remaining data included 2,848 patients with 37 escalations to the resus area, after arrival to the ED. From the 50 escalations to resus area, in the documented patients (see the previous point), for 14 patients there were no data from beds with an active data-fusion system, i.e. they were escalated from the triage beds directly to the resus area beds.

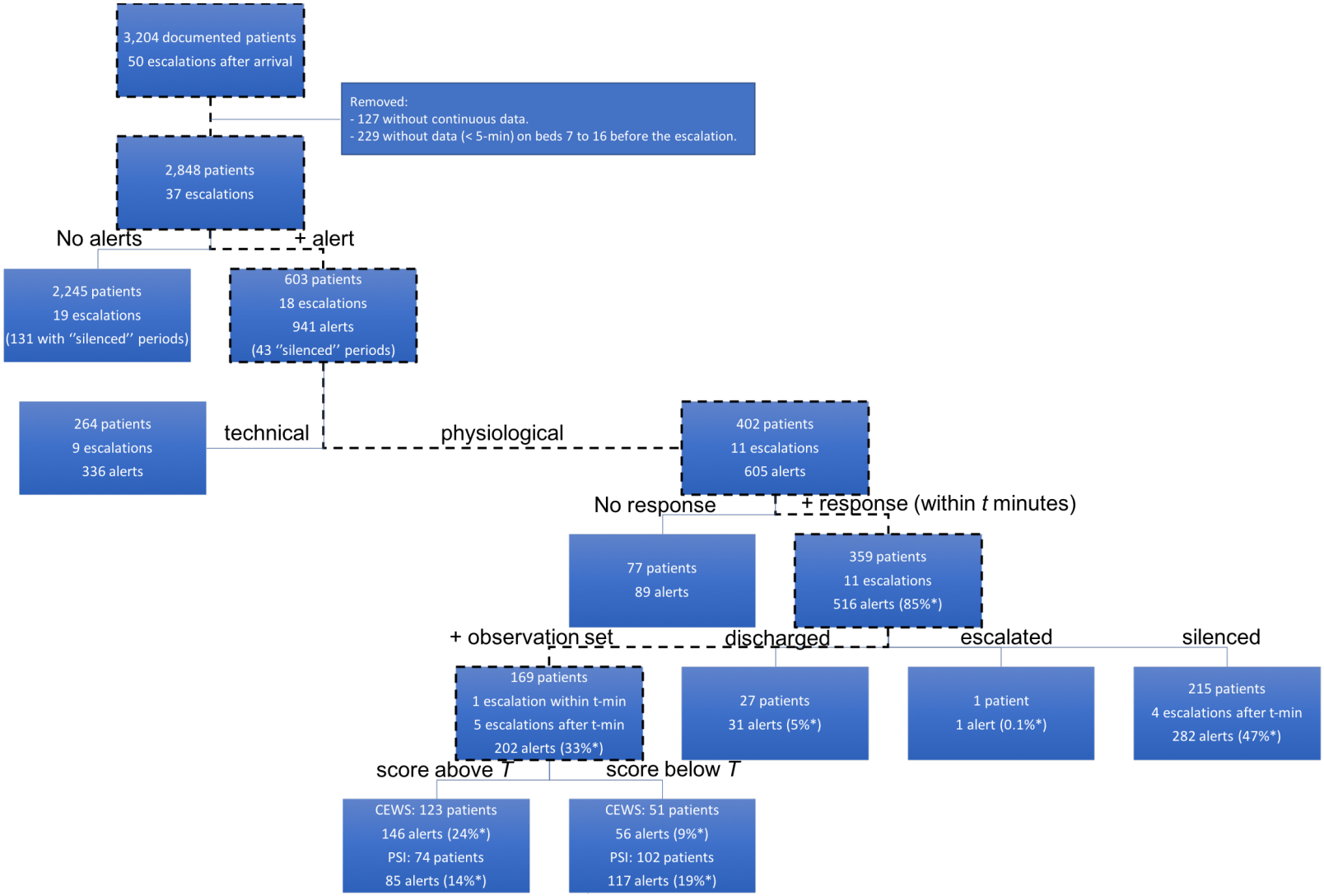


Figure 7.10: *Breakdown of clinical staff response to the data-fusion system alerts when removing technical alerts. "Score above T" means EWS* $\geq$ *3 and PSI* $>$ *3, where* $T = 3$ *in this case. *The percentage is calculated over the total number of physiological alerts (605).*

3. Patients with and without alerts: From the total of 2,848 patients, 603 patients had at least one alert (941 alerts in total). From those patients without alerts (2,245) 131 had periods during which the data-fusion system was silenced. For 115 of those patients the silence period was activated before the patient was attached to the monitor, either because the silenced period had remained activated from the previous patient or had been immediately activated after the new patient was connected to the bedside monitor; and for 16 patients the clinical staff silenced the machine during the patient condition monitoring period. Figure 7.11 shows an example in which the nurse adjusted the SpO_2 probe (an oscillatory SpO_2 signal can be observed) and silenced the alert (30 minutes) during this period.

From those 603 patients with alerts only 18 escalations to the resus area occurred, and 19 (51% of the total of 37 escalations from the 2,848 patients) had no alerts before the escalation. Figure 7.12 shows an example in which the patient presented a low novelty score (i.e. no alerts), before the escalation. The lack of RR′ data from 3:30 to 4:05 am set the RR′ value to the mean of the training set (18 rpm) in the data-fusion system, which depressed the novelty score and hence no alert was generated. In the presence of high RR, the high HR′, just after 5:00 am, and again at 5:45 am, briefly causes the PSI to go above 3.0.

Other patterns, that may influence the decision of escalating the patients to the resus area, were found, such as low temperature ($< 36^oC$) and GCS (< 15), which occurred for seven patients and three patients of the patients without alerts, respectively. These parameters are only present in the clinical observations data and not captured in the continuous data.

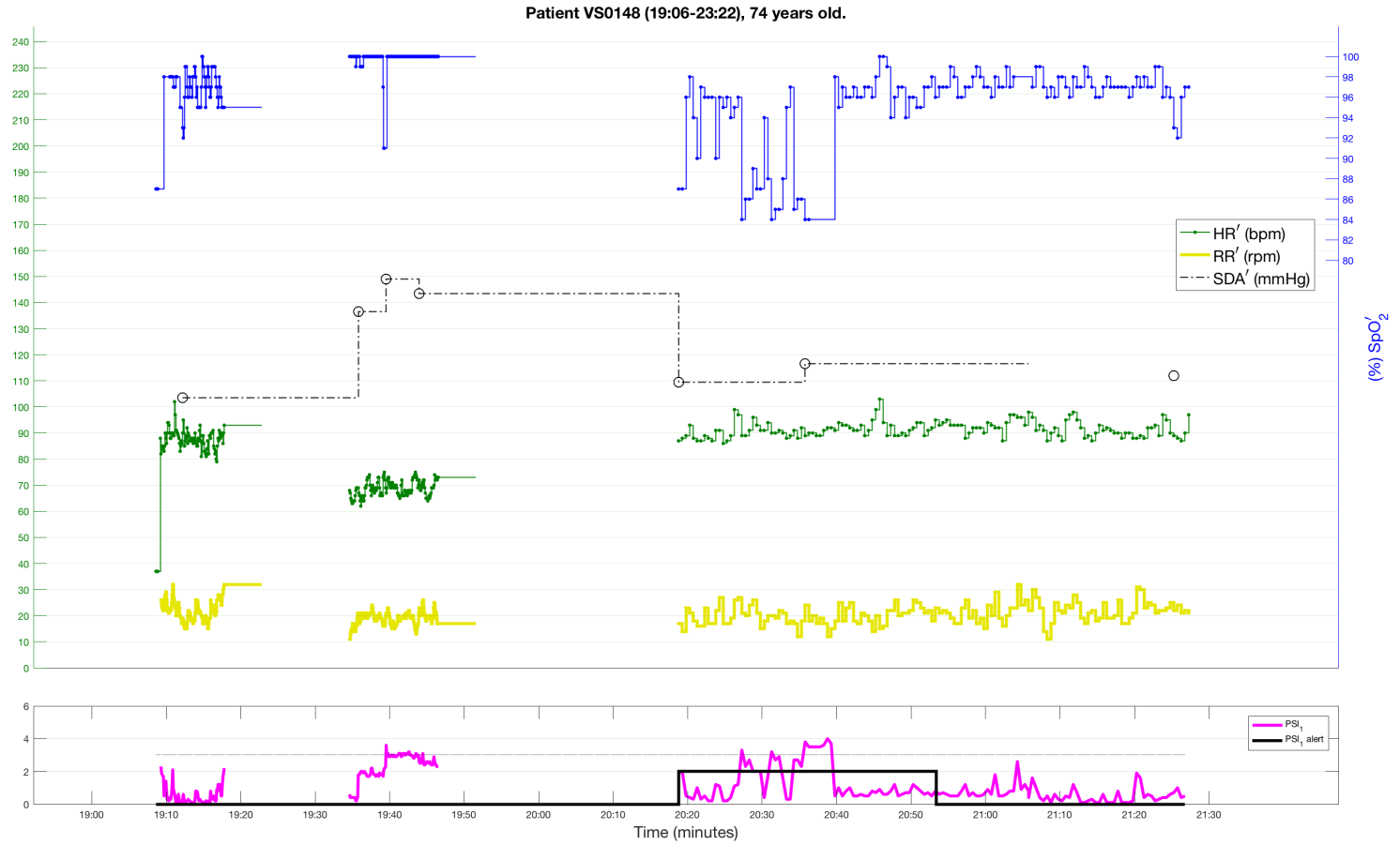


Figure 7.11: *Example of patient without alerts but with a period in which the system is silenced so the nurse can correct the SpO_2 probe without causing an alert. The black line represents the alert state: 0 = no alert (normal physiology), 1 = alert (abnormal physiology), and 2 = system is silenced for a pre-defined period.*

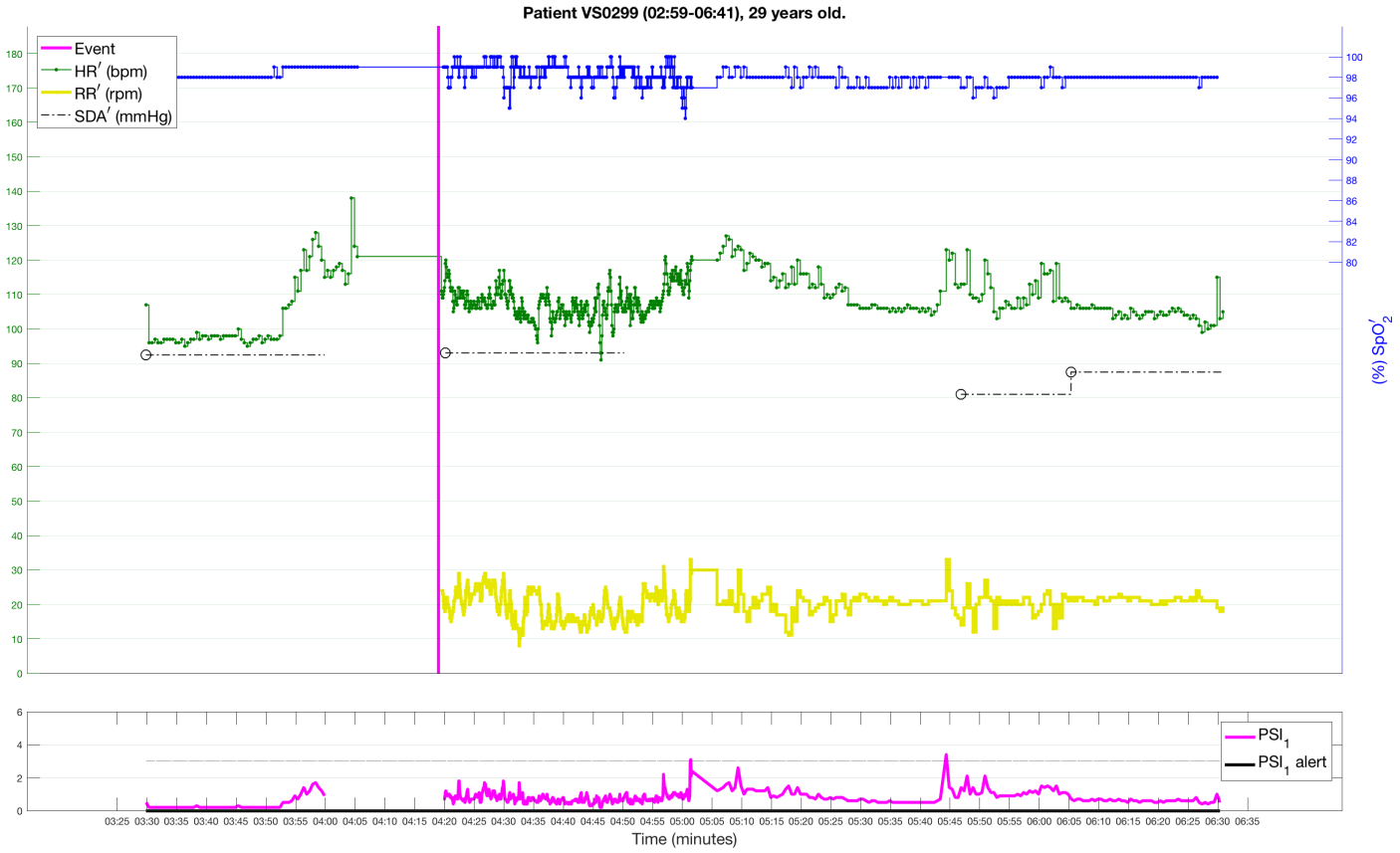


Figure 7.12: *Example of an escalation to the resus area that was not preceded by a data-fusion system alert. The lack of RR′ data from 3:30 to 4:05 am set the RR′ value to the mean of the training set (18 rpm) in the data-fusion system, which depressed the novelty score and hence no alert was generated. Black line represents the alert state: 0 = no alert (normal physiology), 1 = alert (abnormal physiology), and 2 = system is silenced for a pre-defined period.*

4. Technical alerts: The final heuristics, defined to remove technical alerts, resulted in 336 (36%) of the 941 alerts being categorised as technical alerts (20.2% SpO_2' related, 1.9% HR′ related, 6.2% RR′ related, 0.6% BP′ related, 1% caused by data hold-out, and 5.9% of type "other"). Most of the artefacts were present in the continuous SpO_2 data, which is in agreement with the findings in Hravnak et al. (2015).

In the group of patients with technical alerts, seven escalations followed only technical alerts, one caused by extreme RR values (< 3 rpm), three caused by changes higher than 10% SpO_2' in 30 seconds, one generated due to extreme blood pressure values (outside physiological limits, 49/11 mmHg), and two caused by SpO_2' below 85% for 5 minutes before the alert, and with rapid recovery to normal values (increase $>$ 10% SpO_2' in 1 minute) within 5 minutes of the alert.

5. Physiological alerts with and without clinical staff response: From the 605 physiological alerts in 402 patients, 516 (85%) were followed by a response from the clinical staff within 15 minutes of the alert, preceding 11 escalation to resus events (19% of the 37 escalations). 89 alerts (77 patients) had no response from clinical staff, within 15 minutes of the alert. However, for half of these (43 alerts) those same patients generated at least one other alert, and one of those alerts caused a staff response. Reasons for the lack of response in the remaining 46 alerts, include (i) three patients that had already received care previously being escalated to the resus area on arrival to the ED; (ii) intermittent alerts that may have been caused by a transient abnormality that resolved itself; or a response beyond the 15-minute window, which can happen for example if the patient is waiting to be moved to another ward.

6. Types of clinical staff response within 15 minutes of the physiological alerts: From the 516 alerts with a response, for 282 (46.6% of the 605 alerts) the only response was silencing the alert. One alert resulted in the patient being escalated to the resus area (Figure 7.13); 202 alerts (33%) were followed by an extra observation set; and for 31 alerts (5%) the patients were discharged from the ED. Those with an extra observation set preceded six escalations, one of these occurring immediately after the physiological

measurement (within 15 minutes of the alert, see Figure 7.14). The remaining escalations (4 of the 11) were only preceded by "silencing" actions. The clinical staff were probably providing therapy to the patients in these cases.

7. Risk scores of extra observation sets: While 24% of the staff responses (those with observation sets that followed an alert) showed a high-risk EWS (EWS ≥ 3), only 14% had an abnormal PSI (PSI > 3). The abnormal vital signs that caused the data-fusion system (PSI) alert stabilised as a result of the treatment initiated, but the EWS includes other vital signs (temperature and GCS) that may have increased the EWS score for those patients.

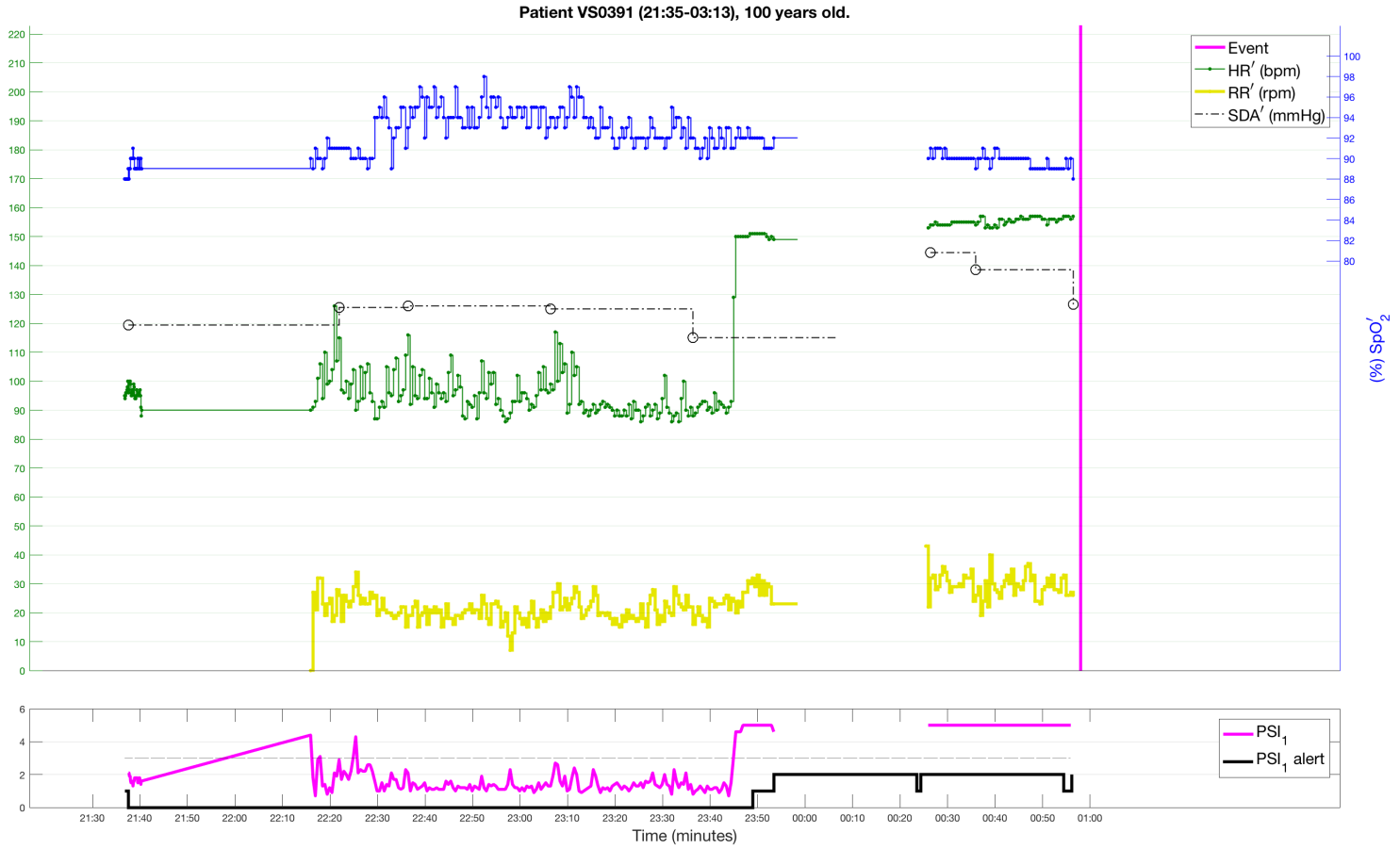


Figure 7.13: *This example shows the alert followed by the escalation event. It is the last alert from a group of 3 alerts that occurred in the same patient. In this case a physiological measurement had already been made 18 minutes after the first alert. Hence, by the time of the last alert the clinical team had already made a number of interventions. Black line represents the alert state: 0 = no alert (normal physiology), 1 = alert (abnormal physiology), and 2 = system is silenced for a pre-defined period.*

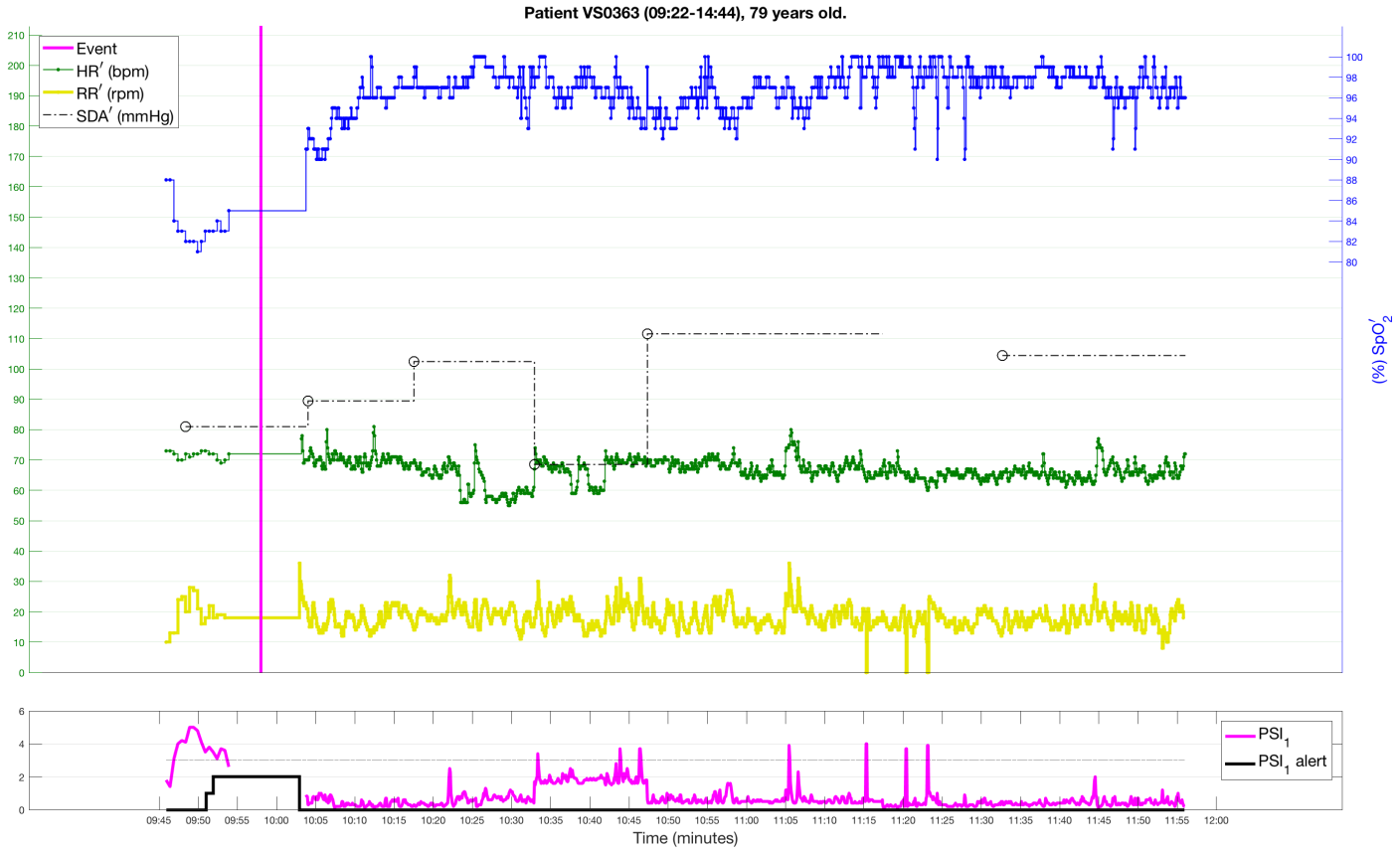


Figure 7.14: *This example shows a case for which both the nurse physiological measurement and the escalation to the resus area fall within 15 minutes of the first alert, clearly due to hypoxia (and low diastolic blood pressure), as the patient is moved to a resus bed and treated with an oxygen mask, afterwards. The black line represents the alert state: 0 = no alert (normal physiology), 1 = alert (abnormal physiology), and 2 = system is silenced for a pre-defined period.*

7.6 Discussion

In this chapter the performance of the data-fusion system, used in phase 3, was analysed. During phase 3 clinical staff were not only asked to respond to electronic-T&T (using the CEWS system) notifications, but also to alerts generated by the data-fusion system, by taking an extra observation set.

A higher percentage of observation sets was observed for patients with EWS_{max} ≥ 3 in phase 3 than in phase 2 (p-value < 0.001), which may indicate that nurses were able to concentrate on the more acutely-ill patients. This effect was already observed in chapter 4, with the introduction of the e-T&T system in the ED ward in phase 2, which was also guiding the nurses in phase 3. We note however, that no significant difference was found in the median time to escalate a patient to the resus area, or the median (normalised) duration of physiological deterioration present in the continuous vital-sign data, between phases 2 and 3 of the large-scale study. This indicates that the introduction of the data-fusion system in phase 3, did not change these metrics significantly.

About 36% (336) of the data-fusion system alerts were categorised as technical alerts. Various artefacts contributed to their occurrence, such as (i) vital-sign values outside of the physiological range, (ii) implausible vital-sign change rates caused by patient movement or probe disconnection, and (iii) inadequate algorithm design in holding-out vital-sign values influenced by artefacts. Alerts that (iv) did not present PSI scores higher than 3 for 4-min, or for which at least 3 channels of data were available in the 5-minute window before the alert (5.9% of the cases), were also considered technical alerts (of type "other").

Our analysis shows that clinical staff responded to 85% of the physiological alerts. 33% and 47% of the total alerts were followed by an extra observation set and a silencing action, respectively. For most of the latter alerts, staff intervened and stabilised the patient before taking an extra observation set (often after 15 minutes since the original alert). Only 49% of the escalations to resus were preceded by data-fusion system

alerts, as there are patterns of deterioration that the system may not be identifying (e.g. neurological deterioration, or other patterns identified by nurse concern, etc.).

7.7 Conclusion

We conclude that ED clinical staff responded to most (85%) of the "physiological" alerts generated by the data-fusion system, but that there is still a large number of alerts (at least 1/3 of the total alerts, in our analysis) being generated due to data artefacts. In addition, half of the escalations were not preceded by an alert. The next chapters will discuss improvements that can be made to the data-fusion system, such as a time-series model that is able to cope with the artefacts present in these data, and the analysis of features that can be added to the data-fusion system, to improve the detection of physiological deterioration in the ED setting.

Chapter 8

Time-series modelling of the vital signs for ED patients

8.1 Introduction

In this chapter we propose a time-series model that is able to (i) fuse continuous bedside monitor data and intermittent physiological measurements recorded by clinical staff in e-T&T systems, and (ii) cope with the data artefacts. We have shown that the latter are caused mainly by poorly-positioned sensors and patient movement, with which the signal processing techniques used by the data-fusion system described in the previous chapter, were not able to cope, creating unnecessary technical alerts. Figure 8.1 illustrates the components on which this chapter is focused.

We hypothesise that the fusion of different data-sources will improve the detection of physiological deterioration in the ED setting, as complementary information may exist in observational and continuous data. Furthermore, if the vital-sign time-series is modelled as being a stochastic process with known characteristics, artefactual periods may be identifiable as deviations from the expected process. As described in Rasmussen and Williams (2006) we make use of the GP regression models to represent the time-series data.

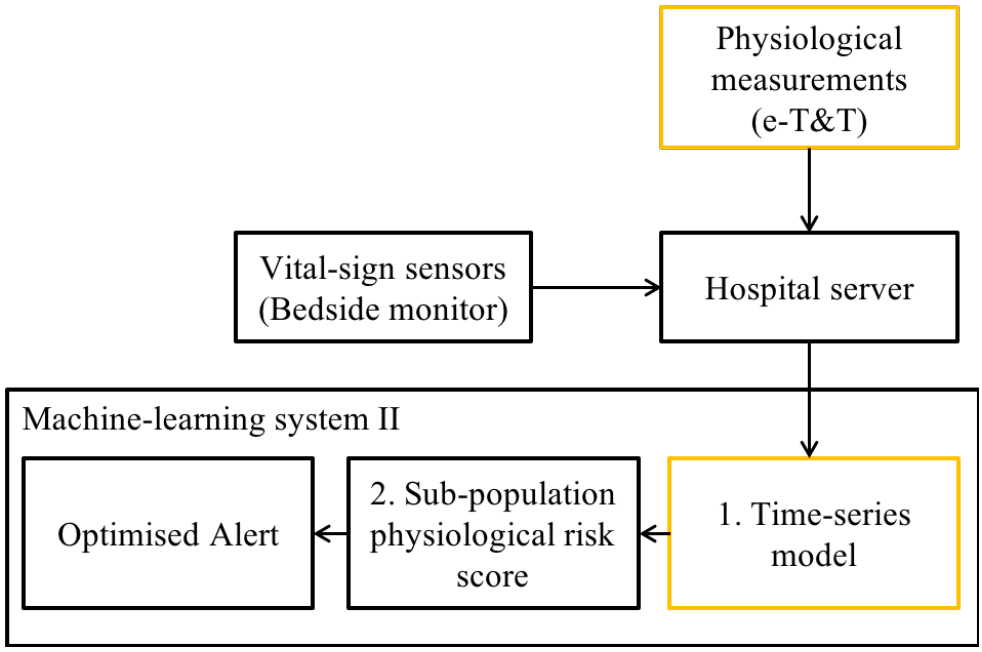


Figure 8.1: *In this 2-stage ML approach, the time-series model is estimated at each step using a Bayesian Gaussian process model which makes use of both continuous and observational data observed up to the time at which a prediction is made. The second stage, the physiological risk score model, is discussed in the next chapter.*

8.2 Fusing intermittent and continuous vital-sign data

The continuous data (HR, RR, and SpO_2), are characterised by uniform, high-frequency, and asynchronous sampling; i.e. the data from the different channels are captured at different times and hence have different timestamps, but with a fixed sampling rate. The physiological data acquired from sensors are typically reliable estimates of the vital signs, although the sensors are susceptible to corruption by artefacts from movement of the patient, etc. The dynamics of the resulting time-series data acquired from sensors depend on patient-specific characteristics, such as age and sex, and the patient context, such as the presence of disease and comorbidities.

In contrast to the continuous data, the physiological measurements are multivariate time-series characterised by non-uniform, intermittent, and synchronised sampling; i.e. the observation timings depend on whenever the nursing staff observe the patient, and are synchronised because a full set of vital-sign observations are typically acquired each time, as part of the hospital's clinical protocol. We note that these measurements made

by nurses may have a small associated bias; e.g. SpO_2 can be read from a nearby bedside monitor (because nurses can observe a pulse oximeter to record SpO_2), but a temperature reading requires waiting for the thermometer measurement to be stable. Clinical judgement is therefore required to record the most appropriate vital-sign value and its timing; furthermore, it has been suggested that clinicians may reject abnormal values in favour of more "normal" values (Taenzer et al., 2014). Furthermore, the authors in (Clifton et al., 2015) discuss the observation that these "corrections" may be made because the clinicians sense information within the measured vital signs that the EWS system does not encapsulate in their overall assessment of a patient's risk status, allowing them to outperform the EWS. It is therefore advantageous to analyse the benefits of fusing the information encoded in the observational and continuous data in estimating the patient health status.

8.2.1 Multi-instance time-series modelling

The multi-instance GP (MIGP) model is here proposed for fusing the observational and the continuous data. Our model assumes that the patient's vital signs can be accurately estimated by taking into account the statistical properties of each source of that vital sign, and fusing them in such a way that the result is a more accurate representation of the true latent physiology of the patient than would otherwise be obtained by treating the sources independently.

Figure 8.2 illustrates the MIGP approach. In various clinical settings a vital sign may be sampled by k different sensors with differing precisions and biases, at different times, and operating at different sampling frequencies. Figure 8.2a shows an example latent "ground truth" signal, $\mathbf{x} = \sin(\mathbf{t})$; and then two time-series of observations of that vital sign, taken from two different exemplar sensing modalities, $\mathbf{x_1} = f(\mathbf{t_1}) + \epsilon_1$, and $\mathbf{x_2} = f(\mathbf{t_2}) + \epsilon_2$. These two time-series were observed at different times, and for our preliminary work, we will assume that they have the same bias, but different precisions, i.e. the error of each sensor is represented by $\epsilon_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, and $\epsilon_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, with

$\mu_1 = \mu_2 = 0$ and $\sigma_1 < \sigma_2$, respectively. We note that, in our example, only the (noisy) samples of the latent signal f can be observed, and we seek to estimate $\mathbf{x}$ using $\mathbf{x_1}$ and $\mathbf{x_2}$.

Suppose that, a vital sign is observed by k sensors (or other sources), with available dataset $\mathbf{D} = [\{\mathbf{t}_1, \mathbf{x}_1\}, ..., \{\mathbf{t}_k, \mathbf{x}_k\}]$, where $\mathbf{t}_k = [t_1, ..., t_{N_k}]_k$ are observation times, and $\mathbf{x}_k = [x_1, ..., x_{N_k}]_k$ are the corresponding observations. N_k is the number of observations provided by source k. We will assume that noise hyperparameters are permitted to vary according to source, yielding k noise terms in the diagonal of the covariance matrix:

$$\mathbf{K_{MIGP}} = \begin{bmatrix} K(\mathbf{t}_1, \mathbf{t}_1) + \sigma_1 \mathbf{I} & K(\mathbf{t}_1, \mathbf{t}_2) & \dots & K(\mathbf{t}_1, \mathbf{t}_k) \\ K(\mathbf{t}_2, \mathbf{t}_1) & K(\mathbf{t}_2, \mathbf{t}_2) + \sigma_2 \mathbf{I} & \dots & K(\mathbf{t}_2, \mathbf{t}_k) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{t}_k, \mathbf{t}_1) & K(\mathbf{t}_k, \mathbf{t}_2) & \dots & K(\mathbf{t}_k, \mathbf{t}_k) + \sigma_k \mathbf{I} \end{bmatrix}, \tag{8.1}$$

i.e. $X_t \sim MIGP(\boldsymbol{\theta})$, is a random variable that follows a GP with a covariance function,

$$\mathbf{K}_{MIGP} = \mathbf{K}_\theta + \mathbf{K}_{WN}^k \mathbf{I}, \tag{8.2}$$

that adds k Gaussian noise terms $\mathbf{K}_{WN}^k$, associated with the indices of each k^{th} instance of X_t, to the GP hyperparameter set $\boldsymbol{\theta}$. The remaining hyperparameters of the set $\boldsymbol{\theta}$, are assumed to be shared between the instances of X_t, to produce one, fused, data-stream. The GP mean is also assumed to be the same between the instances, and is set to zero in our model. To specify the times and observations, $\{\mathbf{t}_k, \mathbf{x}_k\}$, belong to an instance k, a labels vector is added as additional input, specifying the indices corresponding to the different instances (this labels vector is omitted in our notation).

Finally, at prediction time, the noise term σ, from equations 6.25 and 6.26, is replaced by the weighted average of the noise values over the data sources:

$$\hat{\sigma} = \frac{\sum_{k=1}^{n} N_k \sigma_k^2}{\sum_{k=1}^{n} N_k}. \tag{8.3}$$

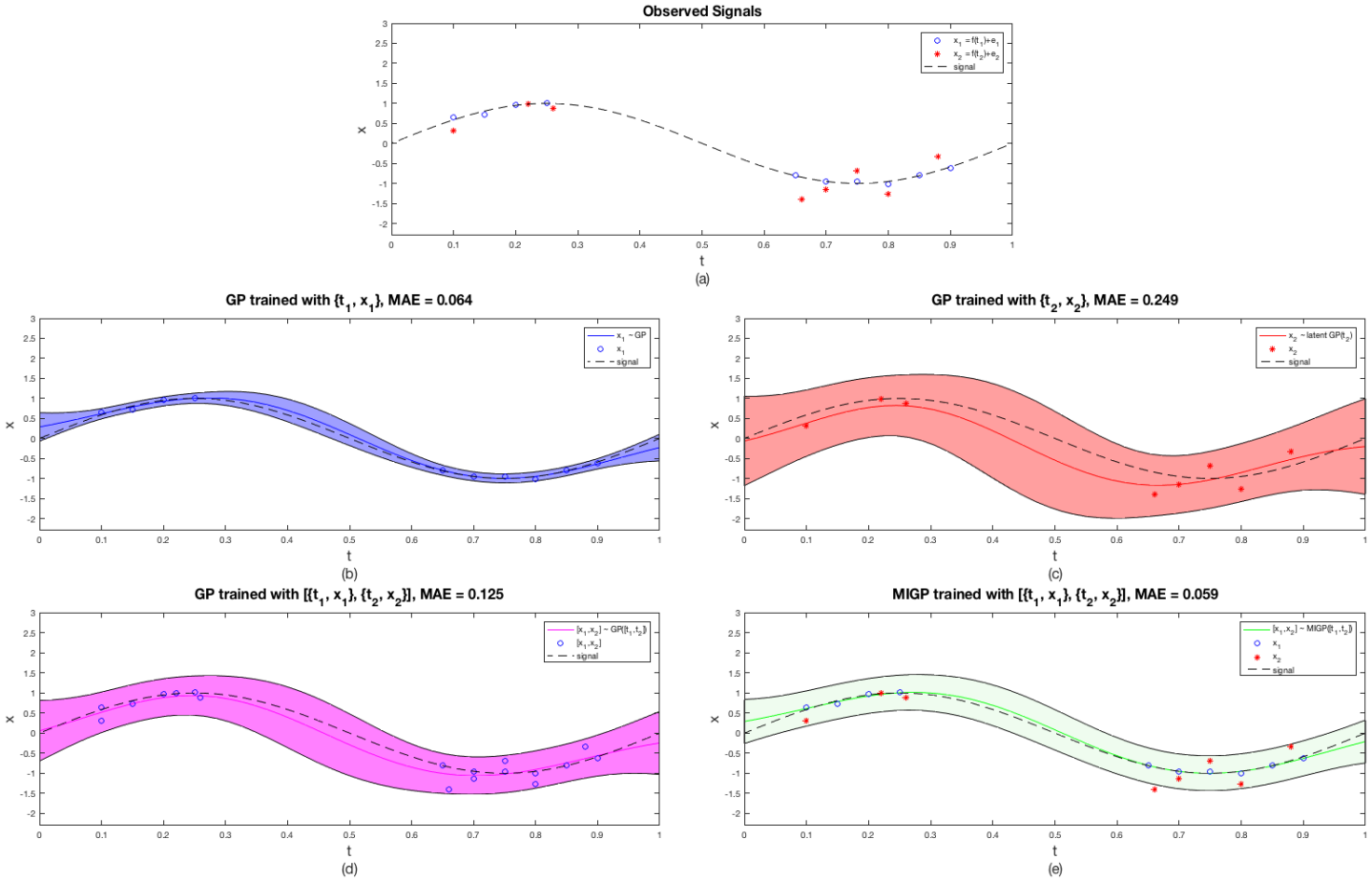


Figure 8.2: *GP versus MIGP example for an artificial dataset. The hyperparameters were trained on data (circles) in each plot. (a) For this example,* $\mathbf{x} = \sin(\mathbf{t})$*, in dashed black line, represents the true latent signal, but only samples from* $\mathbf{x_1} = f(\mathbf{t_1}) + \epsilon_1$ *and* $\mathbf{x_2} = f(\mathbf{t_2}) + \epsilon_2$*, are observed; (b) GP posterior (solid blue line) trained on data from* $\mathbf{x_1}$ *; (c) GP posterior (solid red line) trained on data from* $\mathbf{x_2}$ *; (d) GP posterior (solid magenta line) trained on data from both* $\mathbf{x_1}$ *and* $\mathbf{x_2}$*; (e) MIGP posterior (solid green line) trained on data* $\mathbf{x_1}$ *and* $\mathbf{x_2}$*. The MIGP in (e) gives a lower mean absolute error (MAE) than the GP trained on the mixture of datasets in (d).*

To complete our example from Figure 8.2, we note that if the hyperparameters of the GP model are estimated using data from each sensor, the model obtained for data with a high precision will approximate the target signal with low error and high confidence. For the examples in this Figure, the GP hyperparameters were estimated by minimising the NLML using the gradient descent based approach described in section 6.4. As illustrated in Figure 8.2b, confidence in the model's estimation of the latent function is evaluated by representing the GP distribution up to two standard deviations (solid black lines) around the estimated means (dashed black line).

However, a GP optimised with data from a sensor with low precision can be expected to estimate the latent signal with higher error and lower confidence, as shown in Figure 8.2c. If the data from the two sensors are straightforwardly combined and then a GP is fitted to the resulting superset, then it could be expected that the error will be lower than that obtained using the model from the high noise in the data. However, the result remains a suboptimal estimator of the latent function because the high noise in the data have introduced error, as show in Figure 8.2d. We note that the proposed method yields an MAE of 0.059, which is substantially lower than the MAE of 0.125 obtained by naively using the superset of all data.

In practice we have no knowledge of which of the available datasets has the highest precision[1], because the latent "ground truth" is not known. Hence, we might not know which of several available sources to trust. We suggest that the proposed method performs well when no such prior knowledge exists concerning which of the available sources is not trustworthy for the estimation of an unseen latent function, as in the example shown by the figure.

[1]We also have no knowledge of which signal has the lowest bias, although we note that our preliminary model only optimises the different precisions (noise levels) associated with each instance, as we have assumed the bias to be the same.

8.3 Time-series modelling methodology for the vital signs of ED patients

8.3.1 Data preprocessing

Continuous and observational vital-sign data from the large-scale ED study (chapter 3) are used in the investigations reported in this chapter. As in chapter 5, we use dataset $E_{\{1,2,3\}}$, 2,803 admissions in total, with 2,707 and 96 admissions without and with escalation events, respectively. Data from phases 1 and 2 are used as training data, and data from phase 3 are used as test data.

Vital-sign data outside of the physiological limits were removed (see physiological limits in Table 4.1). The HR, RR, and SpO_2 time-series data were down-sampled to 1-minute windows, using the median of each minute window. For the MIGP models the log of the SpO_2 data is used, i.e. $x' = \log(105 - x)$, to make the data more normally distributed (we refer to this transform as $\log(SpO_2)$, thereafter). For the MIGP model training phase, each patient's time-series data are detrended by subtracting the mean of that channel.

8.3.2 MIGP regression model

Each patient's vital-sign time-series is assumed to be modelled using a patient-specific MIGP function, i.e. $X_t^{i,j} \sim MIGP(\boldsymbol{\theta}^{i,j})$ is a stochastic process representing the time-varying physiological data stream j, each resulting from the combination of data-sources k, measured from patient i, and estimated at times $\mathbf{t}$.

The methodology for modelling the vital-sign time-series, consists in three steps:

1. **MIGP prior specification and hyperparameter optimisation**[2]: Based on similar work in the literature (Pimentel, 2015; Colopy et al., 2016), the sum of

[2]The Gaussian Process Regression Toolbox version 4.0, from Rasmussen and Nickisch (2010), and the Bayesian optimisation software package from Gardner et al. (2014b), are used.

two squared-exponential covariance functions, modelling short-term and long-term variations in the vital signs, is used: $\mathbf{K}^{i,j} = \mathbf{K}^{i,j}_{SE_1} + \mathbf{K}^{i,j}_{SE_2} + \mathbf{K}^{i,j,k}_{WN}\mathbf{I}$. Indices i, j and k identify the patient, the data channel, and the data channel instance (or source), for which the kernel, $\mathbf{K}^{i,j}$ is trained. In our model, the instance k contributes with the Gaussian noise term, $\mathbf{K}^{i,j,k}_{WN}$, e.g. observational and continuous HR data are two instances of the patient HR, measured in the ED setting. The signal-variance hyperparameter, $\theta^{i,j}_{SE_2,h_2}$, is fixed to be the patient-specific variance for the particular channel, and the MIGP mean function is set to zero. The total number of patient-specific hyperparameters to tune per vital-sign channel is therefore: $\boldsymbol{\theta}^{i,j} = \{\theta_{SE_1,h_1}, \theta_{SE_1,l_1}, \theta_{SE_2,l_2}, \theta^k_{WN}\}^{i,j}$, i.e. 5 hyperparameters per patient and channel, as $k \in \{1, 2\}$, in our dataset. Finally, the MIGP likelihood function was assumed to be Gaussian.

The Bayesian optimisation problem is formulated as estimating the hyperparameters $\boldsymbol{\theta}^{i,j}$ that minimise:

$$\begin{aligned} argmin_{(\boldsymbol{\theta}^{i,j})} \quad & -\ell = -g(\boldsymbol{\theta}^{i,j}) \\ s.t. \quad & l_m \leq \boldsymbol{\theta}^{i,j}_m \leq u_m, \end{aligned} \tag{8.4}$$

where $-\ell$ is the NLML objective function, derived from equation 6.27. l_m and u_m represent the lower and upper bounds placed on the m^{th} element of $\boldsymbol{\theta}^{i,j}$. The following bounds are used for each $\boldsymbol{\theta}^{i,j}$:

$$\begin{aligned} \{l_1, u_1, l_2, u_2, l_3, u_3, l_4, u_4, l_5, u_5\} = \\ \left\{\frac{2.5}{60}, \frac{40}{60}, \sigma^j_{min}, \sigma^{i,j}, 1, 6, \sigma^j_{min}, \frac{\sigma^{i,j}}{\sqrt{2}}, \sigma^j_{min}, \frac{\sigma^{i,j}}{\sqrt{2}}\right\} \end{aligned} \tag{8.5}$$

The bounds l_1, u_1, l_3, and u_3, are defined in hours, and l_2, u_2, l_4, u_4, l_5, and u_5 are defined in the units of the data-channel (e.g. bpm for the HR channel), except for SpO_2, which is defined in units of $log(\%)$. σ^j_{min} corresponds to the minimum

signal-variance allowed for each vital-sign channel, and is set at $\sigma^{j}_{min} = 0.01$ for the $log(\mathrm{SpO_2})$ transform and $\sigma^{j}_{min} = 1.0$ for the remaining vital-sign channels. $\sigma^{i,j}$ is the standard deviation of the each patient's data channel.

We note that, if the MIGP model is used with only one data instance, it converts into the GP model. In the case of the observational temperature data, only the observational noise term will be assigned temperature data, and at prediction time the weighted noise term converts to the observational temperature data noise term, i.e. $\hat{\sigma} = \theta^{i,j=6,k=2}_{WN}$ (where $j = 6$, and $k = 2$, represent the temperature and the observational-instance indices, respectively).

A SE-ARD covariance function is used as a prior for the posterior of the objective function, and the $EI(\cdot)$ acquisition function is used to sample the next optimal point, i.e. the one providing a better exploration-exploitation trade-off (see section 6.4.1). The Bayesian optimisation procedure is allowed to search for the optimal hyperparameters for 100 iterations.

2. **Estimation of hyperparameters for new (unseen) time-series data**: A multi-response linear regression is used to learn estimates of the five MIGP hyperparameters, from the stable patients' demographics information (e.g. age and sex), for each data channel. The training data includes 1,670 patients from phases 1 and 2 without escalation events. The zero-mean unit-variance transform is used to normalise the predictor and the response continuous variables.

   The model, trained for each data channel, is of the form $\hat{\boldsymbol{\theta}}^{i,j} = \mathbf{V}_i\mathbf{W}_j + \boldsymbol{\epsilon}_j$, where, as before, index i represents a patient, j represents a data channel, and we have collapsed the two noise terms k, from the observational and continuous data instances, under j. $\hat{\boldsymbol{\theta}}^{i,j}$ represents the vector of estimated MIGP hyperparameters. $\mathbf{V}_i$ represents the design matrix; $\mathbf{W}_j$ is the vector of unknown model coefficients and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_j)$, a multivariate Gaussian noise term.

   The unknown parameters $\mathbf{W_j}$ and $\boldsymbol{\Sigma_j}$ are obtained by maximising the log-likelihood

objective function of the multi-response linear regression model. The algorithm terminates when the changes in the coefficients or log-likelihood are below pre-specified tolerances or when the maximum number of iterations is reached (the "mvregress" implementation from Matlab was used). The performance of this model is compared with that of the median estimator by computing the root-mean-squared-error (RMSE) between the ground truth and each of the models (regression, and the median). The ground truth is derived from applying Bayesian optimisation to obtain the patient-specific MIGP hyperparameters for the test data (as done for the training data).

3. **MIGP sequential synchronisation**: A GP filtering process is used to synchronise the vital-sign time-series. The synchronisation process starts from the time of the first data-point collected on any channel instance including, HR, RR, SpO_2, SBP, DBP, and TEMP. The weighted noise term from equation 8.3 is used to determine the process noise at prediction time, balancing the contributions from the observed intermittent and continuous data. MIGP minute-by-minute step-ahead estimates (predictions) are made for each channel using all of the continuous data up to 4 hours[3] before that minute. The updated mean is used at each step-ahead estimate, i.e. the data is detrended with a mean determined from all data available up to that step. The MIGP is allowed to predict values for up to five minutes of missing data in the HR, RR, and SpO_2 channels and for up to 30 minutes in the case of SBP, DBP and TEMP, matching the configuration of the Visensia system, operating during the ED study, in holding-out the SBP and DBP. The MIGP sequential synchronisation procedure is represented as the SMIGP model, thereafter.

Figure 8.3 illustrates the overall modelling methodology.

[3]This window is selected as most patients are in the ED for about 4 hours, and computing the GP posterior involves inverting a $n \times n$ matrix, $\mathbf{K}^{-1}$, which has complexity of the order of $\mathcal{O}(n^3)$, i.e. there is a computational limit in its use in real-time GP applications.

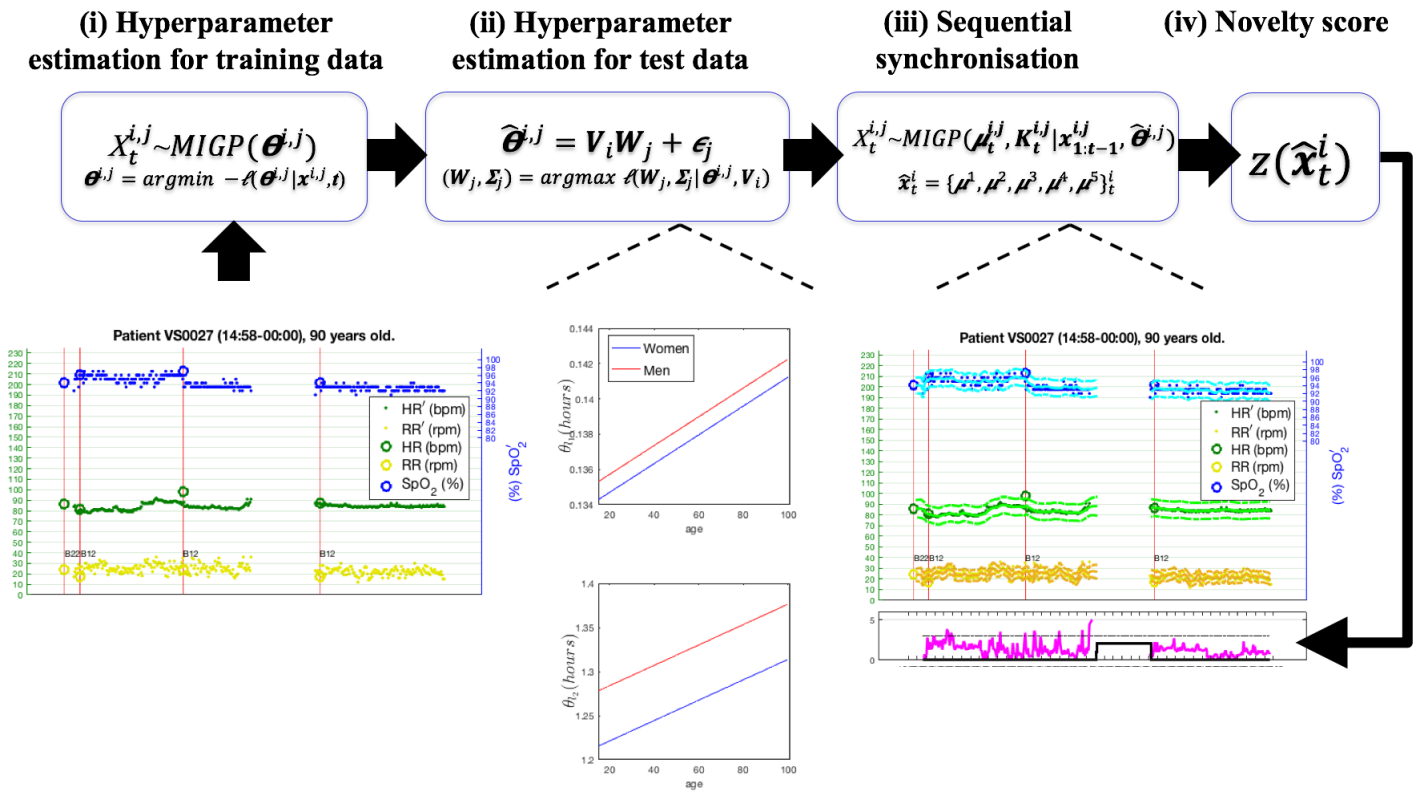


Figure 8.3: *From left to right: (i) $\boldsymbol{\theta}^{i,j}$ are estimated by minimising the NLML, $-\ell(\boldsymbol{\theta}^{i,j}|\mathbf{x}^{i,j}, \mathbf{t})$, for each vital sign j, of each patient i in the training dataset. Exemplar patient data, $\mathbf{x}^{i,j}$, are shown in the leftmost plot. (ii) A multi-output linear regression model, $\hat{\boldsymbol{\theta}}^{i,j} = \mathbf{V}_i\mathbf{W}_j + \boldsymbol{\epsilon}_j$, learns the association (via $\mathbf{W_j}$ and $\boldsymbol{\epsilon}_j$) between the patient demographics data (e.g. age and sex, encoded in $\mathbf{V}_i$) and the hyperparameters, $\boldsymbol{\theta}^{i,j}$, from (i), using maximum likelihood estimation, to estimate hyperparameters, $\hat{\boldsymbol{\theta}}^{i,j}$, for new time-series data in the test dataset. Examples of the use of age and sex to predict short and long-term length-scales θ_{l_1}, and θ_{l_2}, respectively, for the SBP data, are shown in the middle plot. (iii) The MIGP models from (ii) are used to estimate the vital-sign vector $\hat{\mathbf{x}}_j^i = \{\mu^1, \mu^2, \mu^3, \mu^4, \mu^5\}_t^i$, using minute-by-minute step-ahead predictions for each vital sign, at synchronised times t (the SMIGP model). E.g. μ_t^1, represents the MIGP mean HR (index $j = 1$) estimated at time t, using equation 6.23 adapted with the multi-instance noise term from equation 8.3. (iv) The novelty score, $z(\hat{\mathbf{x}})$, is determined over the concatenation of the SMIGP means, estimated for each vital-sign channel at synchronous times, an example being shown in the rightmost plot.*

8.3.3 Novelty score

The novelty score $z(\hat{\mathbf{x}}) = -log\left(\frac{p(\hat{\mathbf{x}})}{p_{max}(\mathbf{x})}\right)$ of the baseline KDE model is computed over the vector $\hat{\mathbf{x}}$, resulting from the means of the vital-sign time-series, synchronised using the SMIGP model described in the previous section.

8.3.4 Patient-wise performance analysis

To evaluate the performance of ML models applied to continuous time-series data, the patient-wise performance analysis, discussed in Clifton et al. (2011), is used instead of the observation-wise performance approach (used in chapter 5 to optimise EWS systems), in order to avoid multiple positive tests for a single patient. This reflects the clinical need to alert for only the first event, thereby ensuring that each patient contributes at most one positive event to the analysis.

As shown in Figure 8.4, an alert is deemed to match an escalation (TP) if it occurs within some time t_1 before the clinical escalation or some (shorter) time t_2 after the escalation event. For the analysis in this chapter the entire period before the alert is considered (most patients are in the ED for a maximum of 4 hours, the escalation occurring within this time) and no tolerance is allowed, i.e. $t_2 = 0$. A FN occurs when there is no alert generated within the window (t_1 and t_2) considered for the abnormal vital signs to be correlated with the escalation of care. A TN occurs when no alert is generated for a patient without escalations. An FP occurs when an alert is generated for a patient without escalations. The persistence criterion (defined in section 6.3.1) is used to generate alerts on continuous data.

8.3.5 Models considered

1. PSI_1: The baseline novelty score algorithm - the Patient Status Index - computed by the Visensia modules during phase 3. The time-series were synchronised using the zero-order hold procedure.

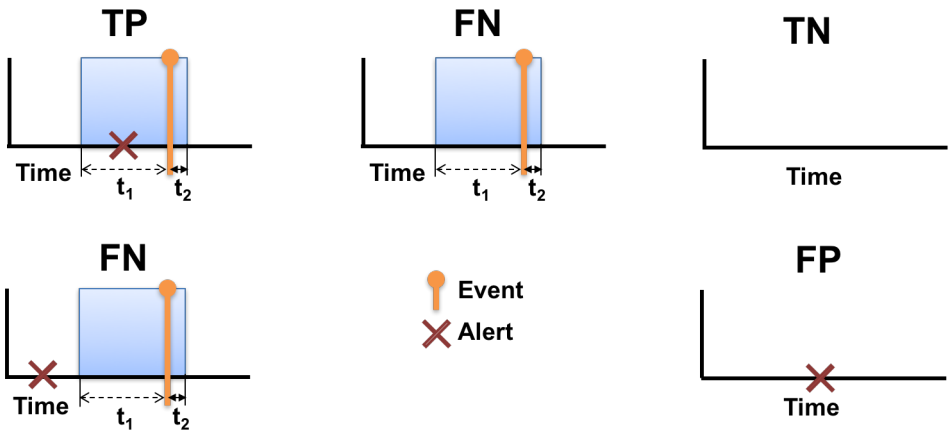


Figure 8.4: *Illustration of the patient-wise performance analysis framework for patient scoring systems to identify the first clinical escalation in the ED. "×" marks the alert time generated by the software algorithm. TP - True Positive, FP - False Positive, FN - False Negative, TN - True Negative. t_1 and t_2 are the start and end of the time window within which the vital signs are deemed to correlate with the escalation event. t_2 coincides with the escalation time, with a small tolerance, as the escalation times are often imprecisely recorded.*

2. PSI_2: The previous model with the technical alerts removed by the heuristics defined in Table 7.1.

3. $\text{PSI}_3 = z(\hat{\mathbf{x}})$: The SMIGP modelling steps are adapted and applied to the single-instance model case. For this case, the novelty score is computed for each $\hat{\mathbf{x}}$, the vector of means estimated sequentially and at synchronised times from a GP model applied to each patient's vital-sign channel, represented as SGP.

4. $\text{PSI}_4 = z(\hat{\mathbf{x}})$: The novelty score is computed over $\hat{\mathbf{x}}$, the vector of means estimated by applying the SMIGP model to each patient's multi-instance vital-sign channel.

The patient-wise performance analysis is used on the test dataset E_3 (1,070 patients, 33 of whom had an escalation to the resus area event). As the normal and abnormal classes from the dataset were unbalanced, in addition to the accuracy, the F1-score (F1-score $= \frac{2TP}{2TP+FP+FN}$) is also adopted as a performance metric.

8.4 Results

8.4.1 MIGP hyperparameters

Table 8.1 shows the median of the MIGP model hyperparameters for each vital sign, estimated from the 1,670 stable patients in the training data. Only the $\theta^{i,j}_{SE_2,h_2}$ hyperparameter is fixed, with the remaining hyperparameters found using Bayesian optimisation.

The covariance function hyperparameters maintained the expected relationships, as a result of selecting appropriate bounds for the hyperparameters. We note that the relationship between the short- and long-term signal-variance parameters was learnt by the Bayesian optimisation approach, as the long-term parameter was maintained fixed, and the short-term parameter allowed to vary from a minimum value up to the long-term hyperparameter value.

The short-term θ_{l_1} hyperparameter median is the lowest possible value (2.5 minutes or 0.0417 hours), for the RR and SpO_2 data channels, twice that value for the HR and BP, and about 15 minutes for the Temperature, which, we note, was trained with observational data only, and hence a higher length-scale was expected. Note also that TEMP observations were only available for about 60% of the training data (see observation sets completeness, in chapter 4). A minimum bound of 2.5 minutes was selected so that the MIGP would not overfit to shorter variations. We note that while the median length-scale of the long-term kernel (θ_{l_2}) is around 1.5 hours, the median is 2.58 hours for the RR, reflecting that it takes longer for this vital-sign value to change significantly.

Table 8.1: *Median and IQR of each hyper-parameter of the patient-specific MIGPs, for 1670 stable patients. NA - Not Available.* $k \in \{1, 2\}$*, represents the continuous and observational data instances for each channel, respectively.*

Channel	$\theta_{SE_1,l1}$(hours)	$\theta_{SE_1,h1}$	$\theta_{SE_2,l2}$(hours)	$\theta_{SE_2,h2}$	$\theta_{WN}^{k=1}$	$\theta_{WN}^{k=2}$
HR	0.08 [0.05, 0.08]	3.38 [2.43, 4.88]	1.39 [1.03, 2.58]	5.32 [3.80, 7.48]	3.76 [2.68, 5.26]	3.27 [2.34, 4.63]
RR	0.05 [0.04, 0.05]	2.17 [1.85, 2.49]	2.58 [1.33, 2.91]	3.11 [2.61, 3.68]	2.19 [1.84, 2.60]	1.93 [1.63, 2.25]
SpO_2	0.06 [0.05, 0.08]	1.12 [1.08, 1.19]	1.78 [1.03, 2.61]	1.20 [1.15, 1.28]	1.14 [1.10, 1.19]	1.10 [1.07, 1.14]
SBP	0.08 [0.08, 0.25]	6.40 [3.21, 9.62]	1.39 [1.03, 2.60]	11.5 [8.09, 15.91]	8.03 [5.66, 11.08]	7.63 [5.41, 10.68]
DBP	0.08 [0.08, 0.21]	5.59 [3.15, 8.50]	1.46 [1.03, 2.61]	9.22 [6.37, 13.07]	6.45 [4.44, 9.12]	6.17 [4.24, 8.71]
TEMP	0.25 [0.13, 0.61]	0.38 [0.24, 0.64]	1.31 [1.16, 4.09]	0.38 [0.23, 0.64]	NA	1.00 [0.99, 1.00]

The observational data related noise terms ($\theta_{WN}^{k=2}$) presented a lower median than that of the continuous data instance noise parameter. This was caused by the existence of much more continuous data to fit $\theta_{WN}^{k=1}$, than observational data, and the objective function - optimising complexity and fitness to the data, as reviewed in chapter 6 - forces most of the variability to be explained by the continuous data instance noise term.

Figure 8.5 presents the scatter plots and Kendall's correlation coefficients (τ) (Kendall, 1948) between the SBP hyperparameters, and between its hyperparameters and the patient age, sex, and clinical outcome (i.e. patients with or without events). In this example, the following effects are observed:

- Correlations between SBP hyperparameters: The significant positive correlation between the SBP signal-variances (θ_{h_1} and θ_{h_2}) and the noise terms, indicates that the hyperparameters preserve covariance structure between them;

- SBP hyperparameters v.s. Age: Positive correlations are found between both instance-related noise terms, θ_{WN_1} and θ_{WN_2} and both signal-variance terms, θ_{h_1} and θ_{h_2}, and the age (significant $\tau = 0.19$, $\tau = 0.18$, $\tau = 0.13$, and $\tau = 0.19$, respectively). This result follows from that in chapter 5, where the positive correlation between the age and the SBP distributions was also significant.[4]

- SBP hyperparameters v.s. patient outcomes: We can observe that significant positive correlation also exists between the SBP long-term signal-variance and those patients with events, which may reflect the fact that higher signal-variations are to be expected for these patients' SBP time-series data.

[4]Significance assessed at p-value < 0.05, thereafter; note that for Figure 5.4 the Spearman correlation was used, and in our current chapter the Kendall' τ is used as it also allows capturing the correlation between a continuous and a categorical variable, which breaks the assumptions needed for applying the Spearman correlation.

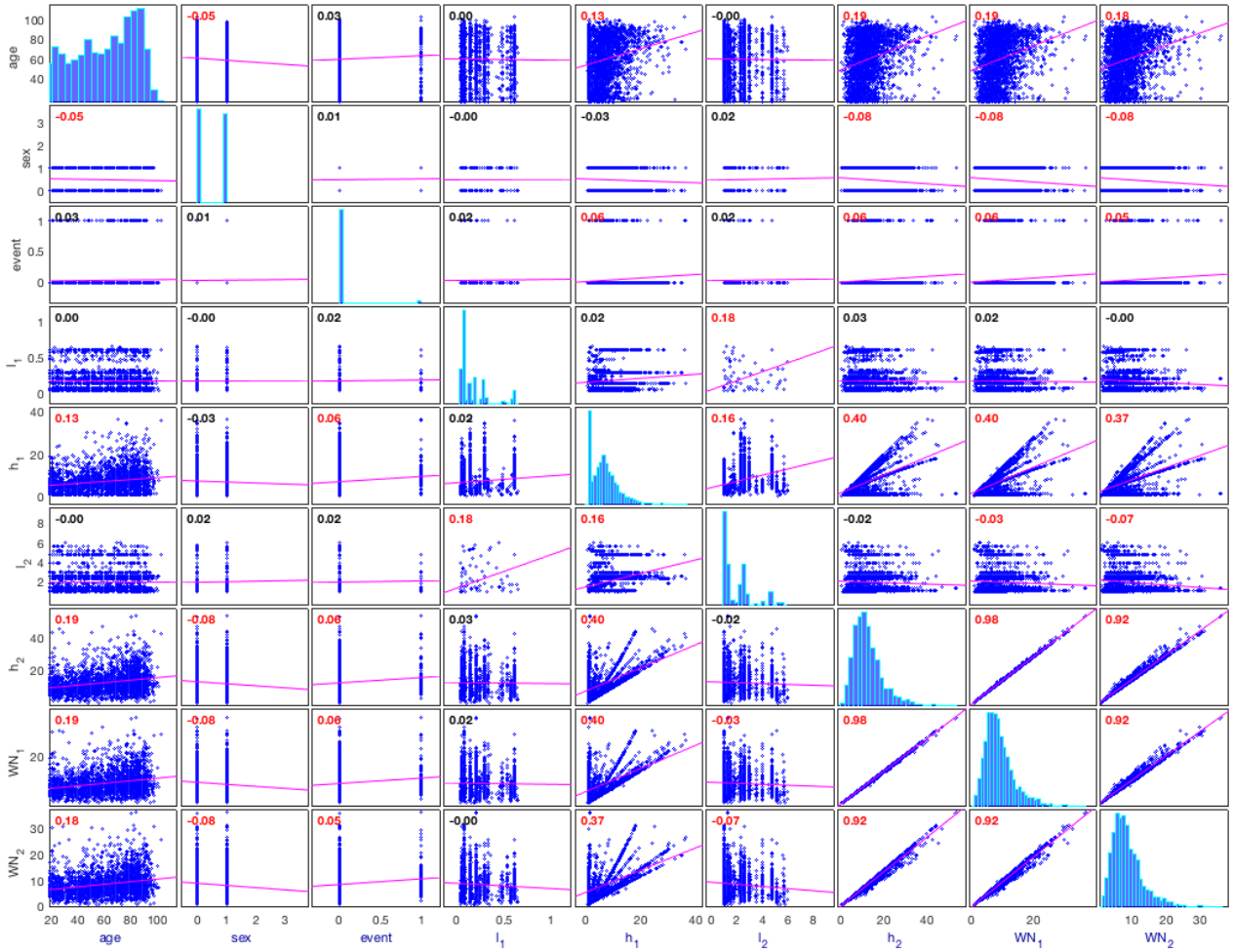


Figure 8.5: *Matlab's corrplot function is used to produce this correlation matrix. The correlation between the MIGP hyperparameters, and ED patients' age, sex, and no-event and event patient outcome is shown for the SBP. The Kendall's rank correlations significantly different from zero are shown in red. A significant correlation can be seen between age and the SBP signal-variances and noise hyperparameters, related to the effect of age on SBP, also discussed in chapter 5.*

8.4.2 Estimation of hyperparameters for test data

Significant correlations were found between the distributions of the patient demographics information and the MIGP hyperparameters (e.g. the age versus the SBP output-scale hyperparameters). Hence, the hyperparameter multi-response linear model estimator, $\hat{\boldsymbol{\theta}}^{i,j} = \mathbf{V}_i \mathbf{W}_j + \boldsymbol{\epsilon}_j$, for each channel j, incorporates the patient age and sex as independent variables. A full multivariate Gaussian covariance matrix is used for the model noise term $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_j)$, so the covariance structure between the MIGP hyperparameters is learnt. The same procedure was applied to the case of the GP model, used in PSI_3. Table 8.2 shows that the multi-response linear model gives a significantly lower RMSE than that of the median estimator, for all the MIGP hyperparameters (Wilcoxon test, p-value $<$ 0.05).

Table 8.2: *Root mean squared error (RMSE) between the ground truth and the MIGP hyperparameters estimated by the multi-response linear regression model, and their median value. The ground truth is obtained by running the Bayesian optimisation procedure to estimate the patient-specific hyper-parameters for the test data. *Significantly different RMSE distributions between the regression models and the median estimators are emphasised, by colouring the Median estimator rows in red (Wilcoxon test, p-value $<$ 0.05).*

Channel	Model	θ_{SE_1,l_1}	θ_{SE_1,h_1}	θ_{SE_2,l_2}	θ_{SE_2,h_2}	θ_{SE,WN_1}	θ_{SE,WN_2}
HR	Regression	0.100	3.563	0.927	4.471	3.119	2.658
	Median	0.105*	3.67*	1.018*	4.582*	3.195*	2.733*
RR	Regression	0.060	0.675	1.098	0.972	0.683	0.610
	Median	0.061*	0.676	1.100*	0.975*	0.685*	0.611*
log(SpO_2)	Regression	0.079	0.086	1.153	0.096	0.065	0.057
	Median	0.083*	0.093*	1.170*	0.101*	0.068*	0.058*
SBP	Regression	0.149	5.612	1.137	7.032	4.883	4.774
	Median	0.170*	5.891*	1.306*	7.559*	5.251*	5.140*
DBP	Regression	0.144	5.228	1.164	6.805	4.730	4.571
	Median	0.166*	5.504*	1.299*	7.194*	4.996*	4.827*
Temp	Regression	0.218	0.280	1.687	1.147	0.628	0.615
	Median	0.290*	0.294*	2.004*	1.201*	0.629*	0.636*

8.4.3 Effect of the time-series model in the performance of the baseline novelty detection model

Table 8.3 shows the results for the patient-wise performance analysis, when the baseline novelty score is computed over different time-series models, for the test data. The improvement in F1-score, when using the SGP and SMIGP time-series models, is explained by a significant drop in the number of FP alerts, when compared to the zero-order hold synchronisation procedure. There is also a marginal gain in the number of TP alerts for the MIGP model versus the GP model. Specific examples are given next.

8.4.3.1 Artefact removal, and technical alerts suppression

We note first that, when the technical alerts are removed by the heuristics defined in Table 7.1, in the PSI_2 model, although they help decreasing the number of FP alerts, from 331 to 232, the sensitivity decreases from 0.576 (0.395, 0.735) to 0.364 (0.209, 0.542).

Table 7.2 from chapter 7, described an example caused by a poorly-positioned SpO_2 probe (artefact TA_1), an example of poor respiratory impedance signal (artefact TA_3), and an example of disconnecting the patient from the bedside monitor sensors (artefact TA_5). The latter is not considered as a technical alert in this chapter, as a sensor disconnection flag could easily inform the data-fusion system of that state. Therefore we focus on analysing the benefits of the time-series model for artefacts TA_1 and TA_3, which are illustrated in Figures 8.6 and 8.7, respectively. To make explicit that the vital signs generated by the SMIGP model are latent variables, we make use of new notation, and we now consider that HR, HR' and $\hat{\text{HR}}$, denote the observational, the continuous (the measurements), and the posterior mean of the latent heart rate variable, generated from the SMIGP model, respectively. We use this notation for the remaining vital signs.

In the leftmost plot of Figure 8.6a, artefacts in the SpO_2' generated a period of extremely low SpO_2' value of 84%. This value was held out for at least 4-minutes, by the data-fusion system, causing an alert that did not precede an escalation event.

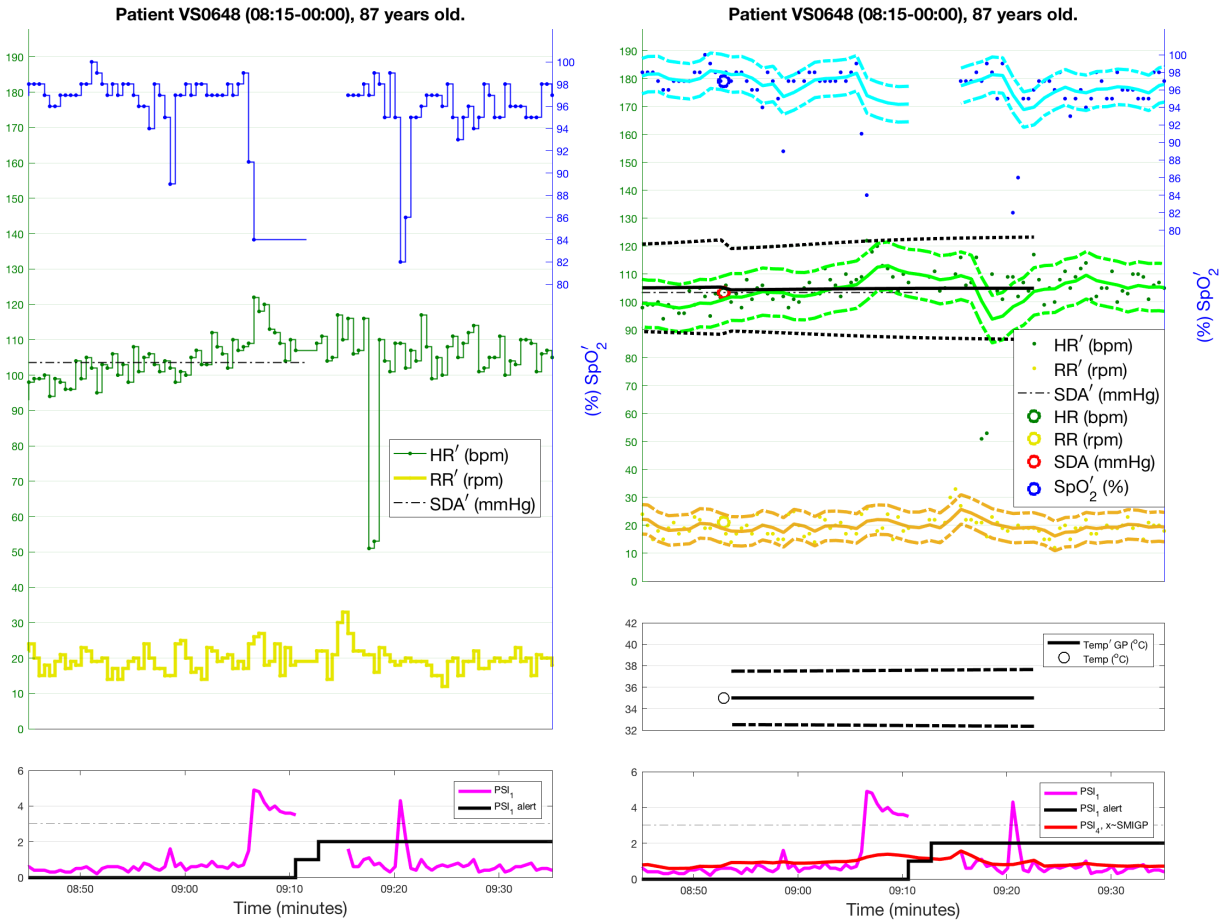
Table 8.3: *Performance of novelty detection enhanced by probabilistic time-series modelling. 1,070 patients from phase 3 are evaluated, 33 of whom had an escalation event. Red emphasises those models with the best value in each metric. TS - Time-series; ACC - Accuracy; PPV - Positive Predictive Value; SEN - sensitivity; SPEC - specificity. The 95% CI (determined by the bias-corrected and accelerated (BCa) bootstrap approach, with 2000 resamples) is shown for the relevant metrics.*

score	TS model	FP	SEN (CI)	SPEC (CI)	PPV (CI)	ACC (CI)	F1-score (CI)
PSI_1	ZOH	331	0.576 (0.395, 0.735)	0.681 (0.652, 0.710)	0.054 (0.034, 0.082)	0.678 (0.647, 0.704)	0.099 (0.061,0.145)
PSI_2	ZOH	232	0.364 (0.209, 0.542)	0.776 (0.751, 0.800)	0.049 (0.027, 0.085)	0.764 (0.737, 0.789)	0.087 (0.046, 0.141)
PSI_3	SGP	181	0.485 (0.324, 0.669)	0.831 (0.808, 0.853)	0.084 (0.049, 0.132)	0.821 (0.796, 0.842)	0.143 (0.087, 0.214)
PSI_4	SMIGP	159	0.515 (0.333, 0.688)	0.852 (0.830, 0.874)	0.099 (0.060, 0.151)	0.841 (0.818, 0.862)	0.167 (0.104, 0.246)

In the right-hand plot (Figure 8.6b), the SMIGP model is shown to attenuate the effect of the low SpO_2' values, generating higher estimates, and avoiding an FP alert. This patient did not require an oxygen mask and was discharged home at the end of the ED stay, without an escalation of care. It is therefore unlikely that the sharp decreases in SpO_2' (from 99% to $\leq$ 84% in 1 minute, in two occasions), are due to physiological deterioration, but rather they were artefactual. Although it does not give rise to a technical alert, we note that the SMIGP model is also able to cope with the sharp artefactual decrease in the HR$'$ value, also shown in the same plot.

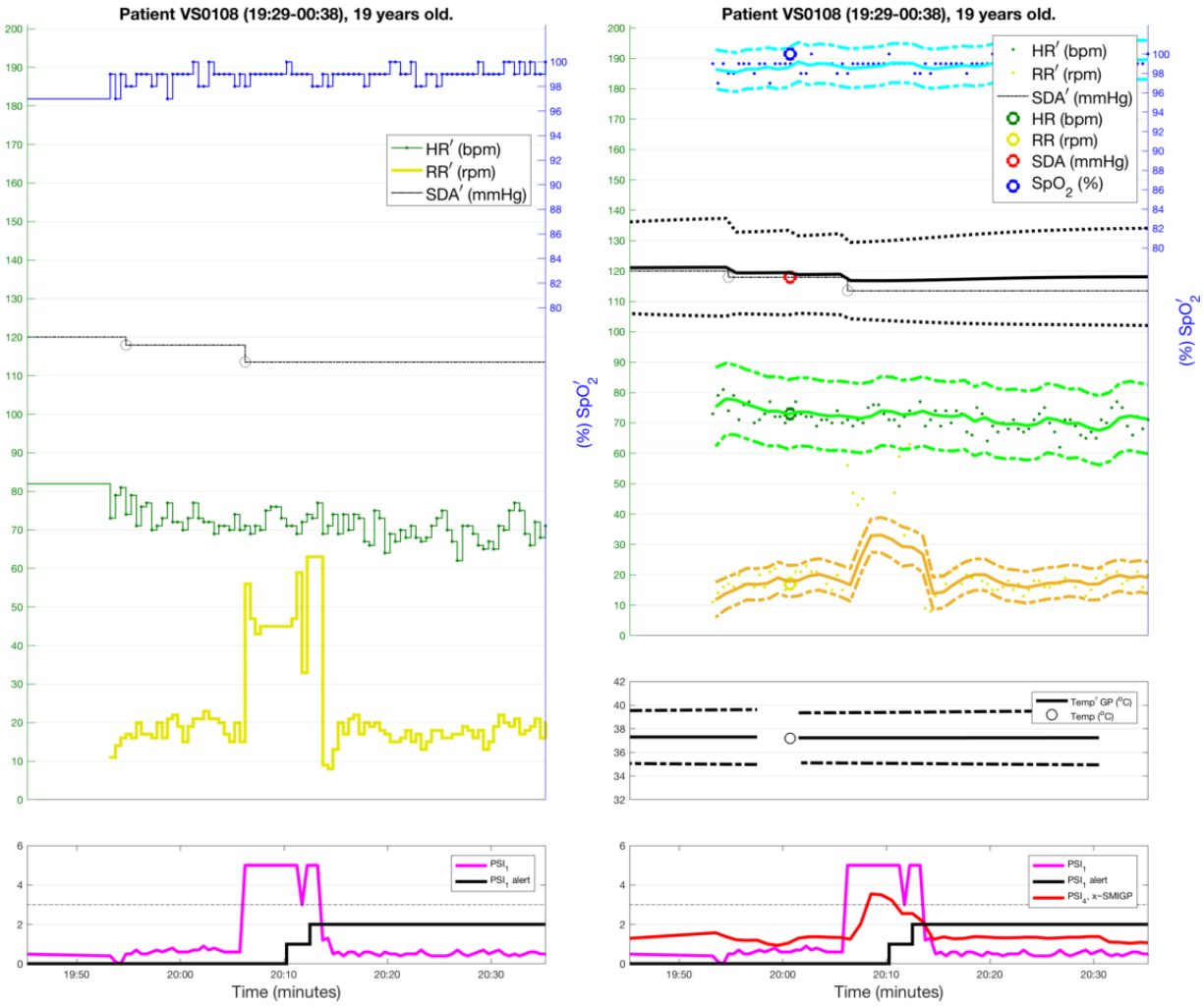
In Figure 8.7a, we show how the SMIGP model mitigates the technical alert caused by the exemplar artefact TA_3, shown previously, in Table 7.2. It is unlikely that a physiologically valid increase from 16 rpm to 55 rpm in 30 seconds occurs in the RR$'$, given that the remaining vital signs, of this 19-year old patient, were stable during the same period. Furthermore, we note that values are missing during part of this period. As a result the high RR$'$ values are held and a technical alert is generated. The SMIGP model shown in Figure 8.7b, estimates lower $\hat{\text{RR}}$ values, in accordance with the patient dynamics and previously observed data, avoiding the FP alert.

These two examples show how the GP-based time-series models cope with technical alerts. The number of FP alerts decreases from 331, for the PSI_1 model, to 181 and 159 for the PSI_3 (SGP) and PSI_4 (SMIGP) models, respectively.



(a) (b)

Figure 8.6: *(a) Shows a common technical alert in the ED dataset, caused by low SpO_2' values, due to loss of attachment of the SpO_2 probe. The zero-order hold procedure holds on to artefactual low values of SpO_2'. (b) The SMIGP model estimates higher $\hat{SpO}_2$ values in accordance with previously seen values (most at 98% for this patient), avoiding an FP alert.*



(a) (b)

Figure 8.7: *(a) Shows problems with the respiratory impedance signal, causing extremely high RR' values to be estimated by the bedside monitor, and trigger an alert in the data-fusion system. (b) The SMIGP model estimates estimates lower $\hat{RR}$ values (18 rpm, versus 55 rpm, the latter estimated by the bedside monitor at the start of the artefactual period). The FP alert is avoided in the SMIGP model.*

8.4.3.2 Missed escalations

On the basis of the F1-score metric, the system using the SMIGP model (PSI_4) performs best, followed by PSI_3 (the SGP model), with F1-scores of 0.167 (0.104, 0.246) and 0.143 (0.087, 0.214), respectively. Both of these models achieve a higher sensitivity (0.515 and 0.485, respectively) than that achieved by the heuristic artefact removal approach (PSI_2), in avoiding technical alerts. However, the PSI_3 and PSI_4 models still miss four patients and three patients with an event, respectively, when compared with the baseline model, PSI_1. We analyse next the three Visensia TP alerts that are missed by the PSI_4 model.

Patient VS0095 (Figure 8.8):

The first and the second PSI_1 alerts, shown in Figure 8.8a, are caused by persistent extremely low RR′ values of 0 rpm (non-physiological). The clinical staff reported RR measurements above 20 rpm near both of these periods, consistent with the values from of RR′ outside of the artefactual periods. Given the discrepancy between the nurse's observations and the beside monitor output, these periods should be considered artefacts, and these Visensia alerts considered as technical alerts rather than TP alerts. As these alerts preceded an escalation of care, they have been erroneously labelled as TP alerts in our performance analysis so far. In addition, they occurred 5.5 and 4 hours before the escalation event, and therefore may not be related with the physiological deterioration leading to the escalation event in the ED setting.

The SMIGP model, on the other hand, estimates higher $\hat{\text{RR}}$ values (the lowest estimates being between 5 and 10 rpm) during these artefactual periods, thereby avoiding both technical alerts (illustrated in Figure 8.8b). We note again that RR < 3 rpm and RR′ < 3 rpm were removed in the preprocessing step used in the retrospective analysis carried out in this thesis (see section 4.2.2), and hence are not observed by any of the GP related models.

Patient VS2461 (Figure 8.9):

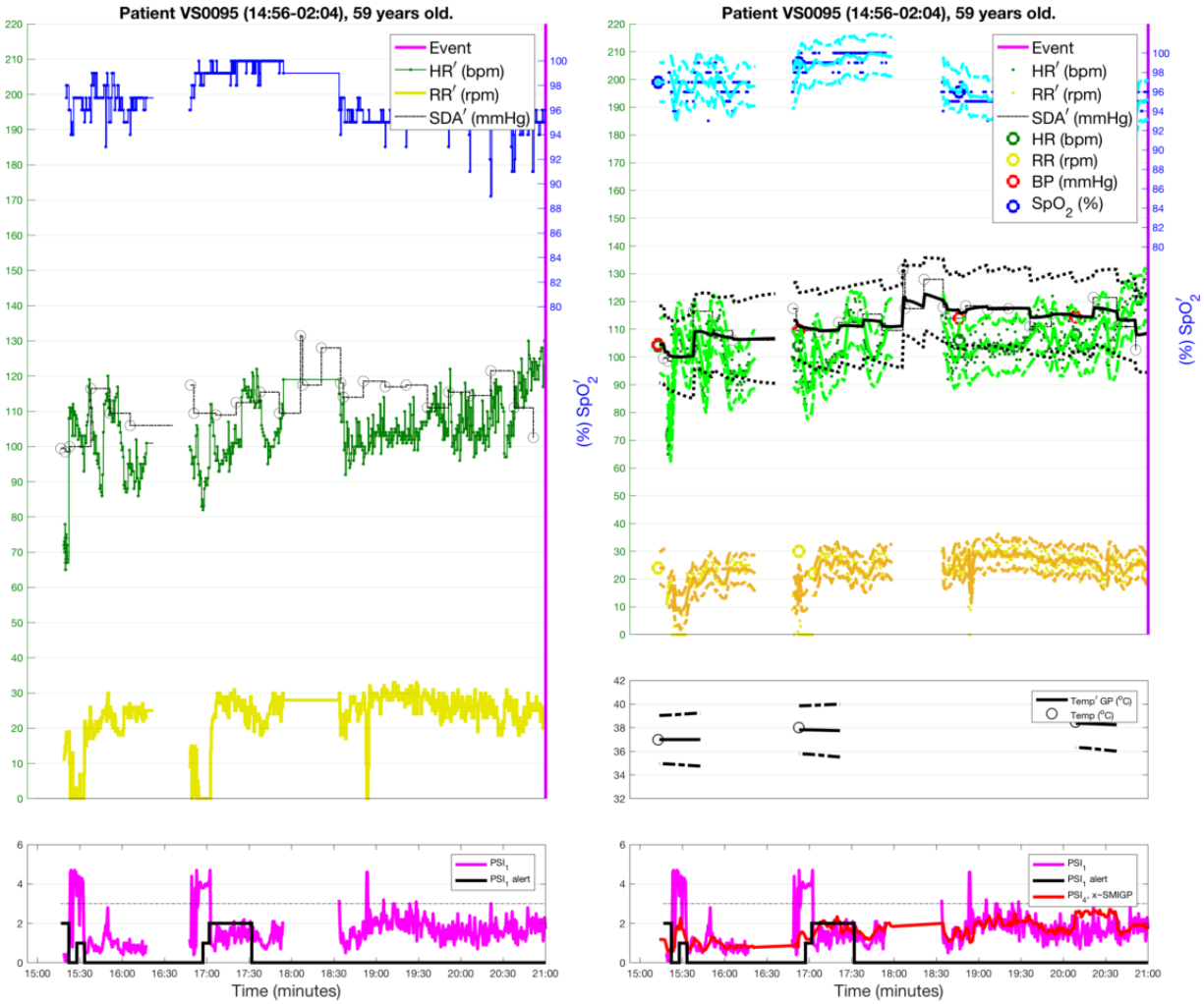
The PSI_1 model alerts on a DBP′ measurement of 10 mmHg (shown in Figure 8.9a). This value was filtered by the preprocessing step used in the retrospective analysis carried

out in this thesis, and therefore the low SDA′ is not observed by the SMIGP model (in Figure 8.9b), thereby avoiding the alert. In this case the patient presented periods of low SpO_2' during the period of low BP′, but it was mainly the low BP′ values that increased the novelty score. We consider that the technical alert generated by the PSI_1 model as a result of the low DBP′ value was also erroneously used as a TP in our performance analysis.

Patient VS2995 (Figure 8.10):

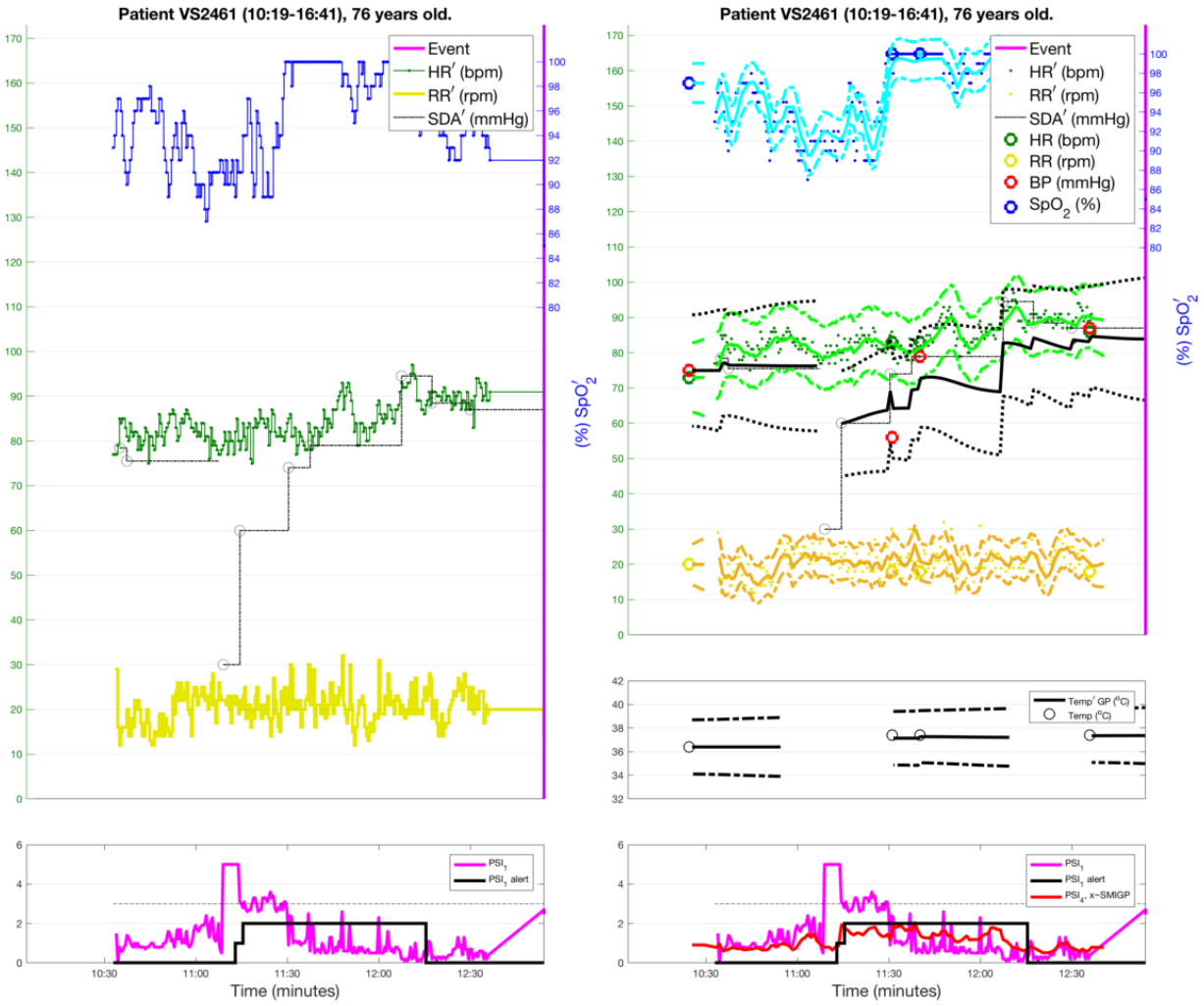
Figure 8.10a shows that the SpO_2 probe was mostly disconnected between 17:45 and 18:30, but temporarily connected (for about 30 seconds) to the patient at 18:20. This temporary attachment generated a SpO_2' value of 79% which was held for 4 minutes, creating an alert in PSI_1. This alert was also erroneously considered a TP alert as the patient was escalated (50 minutes afterwards). The SMIGP model (Figure 8.10b) corrects for this low SpO_2' value, estimating a much higher $\hat{SpO}_2$ value (91%).

In summary, all three Visensia alerts, that were missed by the PSI_4 (MIGP model), were shown to have been technical alerts that coincided with escalation events and were erroneously considered as TP alerts for the reasons given above.



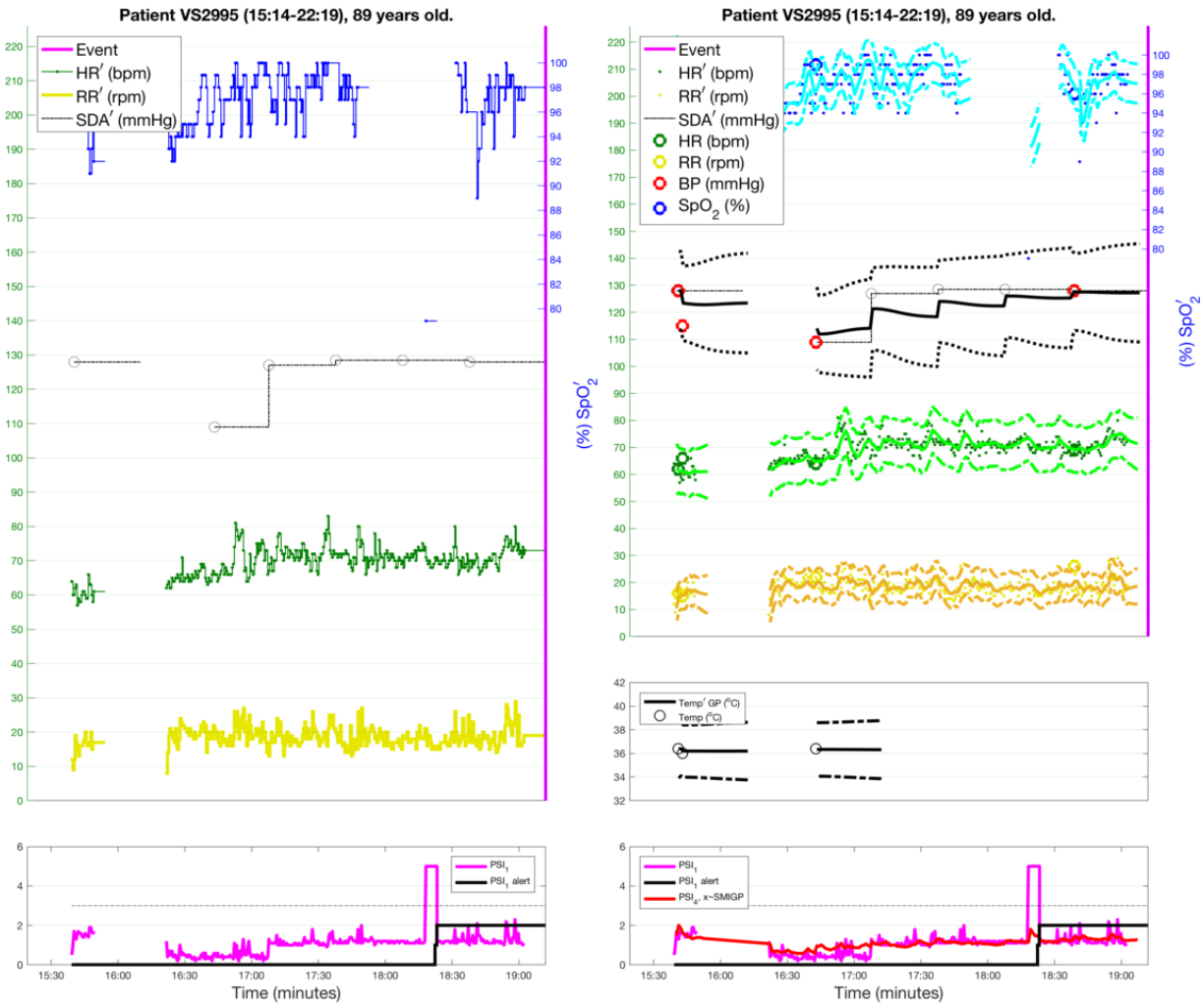
(a) (b)

Figure 8.8: *(a) The Visensia system alerts on both persistent periods of low RR′ (< 3 rpm). However the nurse's RR measurements, during the second continuous low RR′ period, represented in yellow circles, indicate a RR above 20 rpm. Therefore the low RR′ values are likely to be due to artefacts in the respiratory impedance signal. (b) The SMIGP model estimates higher $\hat{RR}$ values during the artefactual periods.*



(a) **(b)**

Figure 8.9: *(a) The Visensia system alerts on a physiologically infeasible value of DBP′ (10 mmHg), which is held for 5 minutes. (b) In the preprocessing step used in the retrospective analysis carried out in this thesis, non-physiological BP values (DBP and DBP′ < 20 mmHg) were removed, and therefore no alert is generated by the SMIGP model.*



(a) **(b)**

Figure 8.10: *(a) The Visensia system alerts on a low SpO_2' value (79%) caused by an artefactual value, which was held for 5 minutes. (b) The SMIGP model estimates much higher $\hat{SpO}_2$ values, and hence does not generate an alert.*

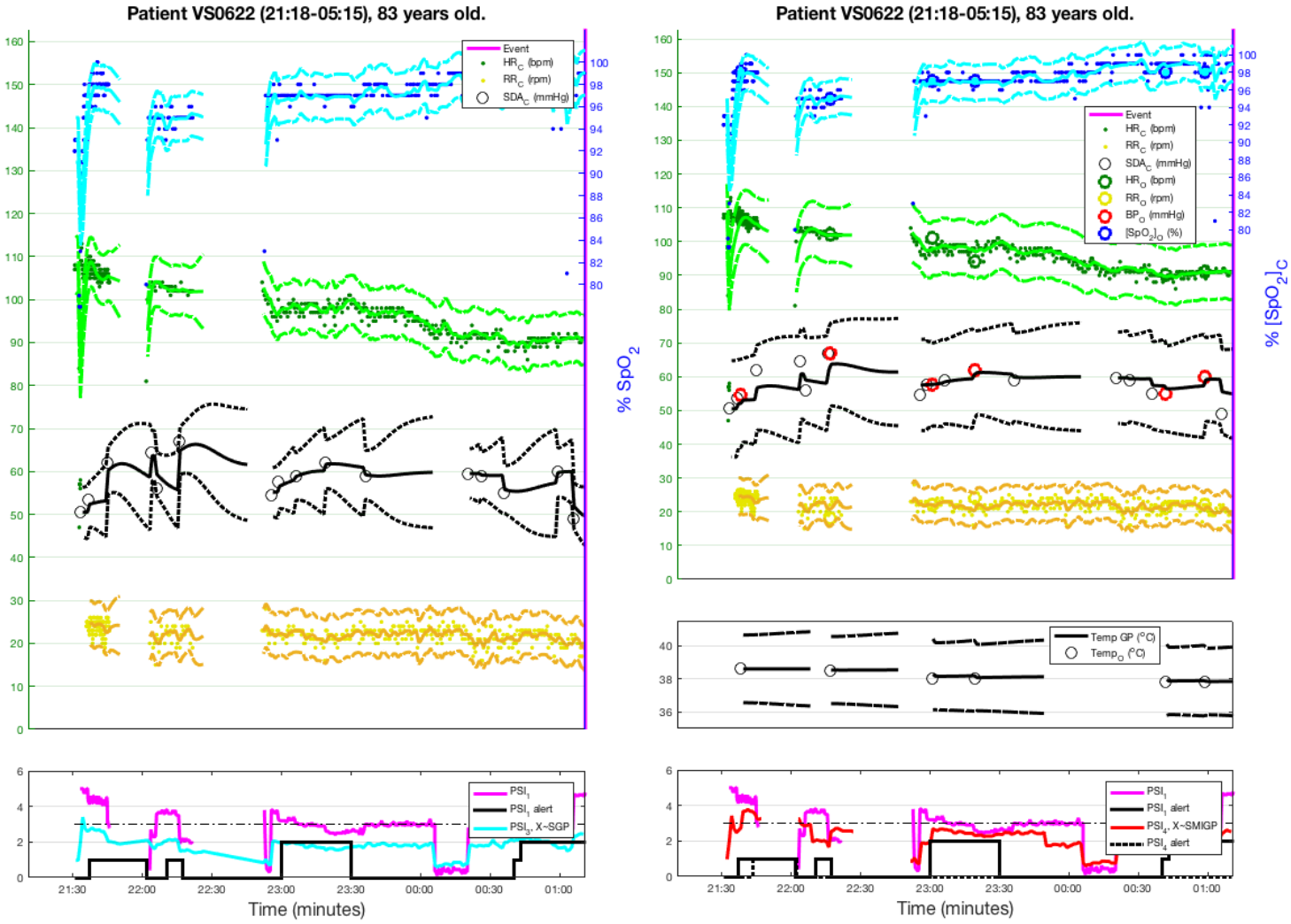
8.4.3.3 Multi-instance versus single-instance approach

We note that the PSI_4 (using the MIGP) identifies physiological deterioration before an escalation in one more event patient than the PSI_3 (single instance, GP model) due to the use of the temperature value, as show in Figure 8.11. The presence of a high temperature value (38.5^oC, at 21:40) in Patient VS0622, measured by the clinical staff and integrated in the continuous data by using the SMIGP model, increased the PSI_4 novelty score, generating the alert before the escalation event. This shows the advantages of combining the complementary information from the intermittent and continuous vital-sign data (the multi-instance model).

8.5 Discussion

In this chapter we have used Bayesian modelling to fuse the observational and continuous vital-sign data acquired by the ED clinical staff, and by the bedside monitors, respectively. A multi-instance GP model and Bayesian optimisation frameworks have been used to learn prior knowledge about the patients' short- and long-term dynamics for a multi-instance HR, RR, SpO_2, BP and temperature time-series.

This approach allowed: (i) synchronising observational data information that is not available in continuous data, such as the temperature, with the patient's continuous data, and consequently use a more complete feature set for the novelty detection model; (ii) better estimation of the intermittent beside monitor data, such as the BP, by fusing them with BP recorded by clinical staff. We note these data sources have a similar number of data points, which gives them the same importance (or "weight") in both estimating the MIGP hyperparameters at training time, and then when filtering the data at test time. In addition, when BP data from the monitor is not available, such as in the triage room, or when the patient is not connected to the monitor, BP data recorded by staff from the "Dinamap" spot-check monitors (see Figure 2.4) into e-T&T systems can be fused with that collected from the monitors (at training and test phases); (iii) for continuous



(a) *PSI applied to the SGP model (PSI_3).*

(b) *PSI applied to the SMIGP model (PSI_4).*

Figure 8.11: *(a) The SGP model smooths the signals, reducing the PSI values (hence no alerts before the escalation). (b) The SMIGP model also smooths the signals, however it considers the temperature signal, from the clinical observations, which is abnormally high ($> 38.0\ ^oC$, at 21:40), increasing the PSI, and triggering the alert.*

data channels, such as HR′, RR′ and SpO_2', the correspondent observational data instance would only be of help in the MIGP model when the continuous data is not available, as otherwise the MIGP estimates are mainly driven by the continuous data, in our current MIGP configuration. As an example, the noisy periods in the continuous data often have gaps and extreme values, and a nurse observation set taken during this period provides data that the MIGP model can use to generate (filter) more appropriate estimates. This approach is promising in estimating the patient physiology, specially in the aforementioned cases, and can be further extended, for example, to model the correlations between the different vital-sign streams, using the multi-task GP model.

A multi-response linear regression has been used to generate hyperparameters for the test data. The patient age and sex were used as independent variables, as significant correlations were observed with some of the patients' vital-sign time-series, examples being given for systolic blood pressure.

This time-series modelling approach was shown to be more robust than the zero-order hold procedure in coping with artefacts that caused technical alerts. The latter decreased from 331 (generated by Visensia during the study) to 159, with the use of SMIGP. This model was also shown to be better than the heuristics defined in Table 7.1, in mitigating the technical alerts.

The addition of the observational temperature data, also contributed to the baseline novelty detection model to achieve the highest F1-score, as, apart from having a lower number of FP alerts, it was also able to alert on one more patient than the PSI_3 model (applied to SGP, which was derived from the available continuous data only), by making use of the temperature information estimated by the SMIGP model. However, we observe that, as there are only a few temperature measurements to fit the MIGP model for each patient, our approach learned a high value for the Gaussian noise hyperparameter associated with each patient temperature covariance function. Its median value, shown in Table 8.1, was high, 1 ^oC. The high variance around the temperature estimates can be seen in the temperature plots in all the Figures used in section 8.4.3. Our current PSI

does not make use of the precision of the estimates, and therefore accepts the temperature SMIGP estimates with high variance, which worked in our current model performance evaluation methodology.

Finally, it was shown that TP alerts generated by the PSI_1 model, but missed by the PSI_4, were technical alerts, that coincided with the period before the patient escalation event; i.e. the time-series model filtered patterns that are extremely rare, in both normal and abnormal patients. However, this does not mean that these patterns are not connected with the physiological deterioration occurring in event patients. For example, in Figure 8.8, the HR of patient VS0095, varies considerably throughout the patient stay, and if the patient was in distress, or agitated as a result of the illness, then this would explain the noisy respiratory impedance signal, that may have caused the monitor to estimate a respiratory rate of zero. As this value is not physiological, our view is that the physiological risk model should have alerted on other patterns, such as the HR variability, which is likely to be an indicator of illness and precede the noisy period that created the technical alert. The challenge exists in differentiating between artefact caused by a random patient movement, and that caused by an underlying illness, correct it, and score the abnormal patterns. In summary, we must now take into account the fact that the time-series model will deviate from rare data artefacts, which can occur in both normal or abnormal patients, and we must re-train our physiological risk models to identify the underlying physiological deterioration pattern. As the raw waveform data is not available in our dataset, we rely on a Bayesian approach to estimate the expected values in periods of noisy data.

8.6 Conclusion

This chapter has presented evidence that modelling the vital-sign time-series correctly can remove a significant number of data artefacts that interfere with the performance of automated patient condition monitoring systems. Furthermore, fusing observational and

continuous data can augment the physiological information available to these systems and improve the detection of physiological deterioration in the ED setting. We now require that the physiological risk models are re-trained on the multi-instance data model, to better identify periods of abnormal physiology in ED patients. We discuss this in the next chapter.

Chapter 9

Model of normality for ED patients

9.1 Introduction

In this chapter we analyse the physiological trajectory of ED patients, from the point of view of both the observational and the continuous vital-sign data. In addition, we analyse the use of data from both of those data-sources (i.e. the multi-instance model, discussed in chapter 8), and information about the patients' clinical context (e.g. requiring oxygen therapy), and demographics (i.e. age and sex), to improve the performance of ML based physiological risk score models, that can be used for automatic patient condition monitoring in this setting (the second stage of the 2-stage ML approach reviewed in chapter 7, illustrated in Figure 9.1).

9.2 Physiological trajectory of ED patients

9.2.1 Univariate physiological trajectories

To analyse the physiological trajectory of patients from the majors area, we use dataset $E_{\{1,2,3\}}$, defined in section 5.2.1, i.e. 2,803 patients in total, from phases 1, 2 and 3. We analyse three main groups, the 658 patients discharged home after their ED stay, the 2,049 patients admitted from the ED to a hospital ward, both without events in the ED,

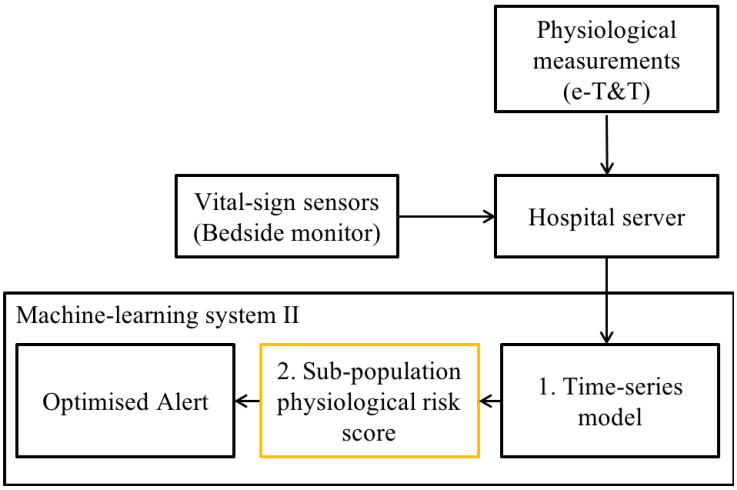


Figure 9.1: *This figure follows from Figure 8.1. A 2-stage ML architecture is proposed: the first stage learns the patient physiolodgical time-series dynamics, as described in chapter 8. The second layer learns a sub-population physiological risk score, i.e. that takes the patient context (current ward/condition/disease/treatment), demographic (age, sex, etc...) and physiology information, into account, the scope of this chapter.*

and the 96 patients with events in the ED. The following procedure is used to represent the average univariate physiological trajectory, from their continuous data collected in the majors area[1]:

1. Each vital-sign time-series (HR′, RR′, SpO_2', SBP′, and DBP′) is first normalised with respect to the patient arrival time in the ED.

2. Each set of patient vital-sign data is then averaged over 30-minute windows (in sequence, without overlap).

3. For each time-window of each vital sign, the average is computed from the windows of all patients with data available for that time-window, for each of the three patient groups.

This process is repeated for the clinical observations data, and the average physiological trajectory is computed using hourly windows in this case. In the case of the

[1]As in chapter 8, we use superscript prime to distinguish observational and continuous vital-sign data, and the hat symbol to represent variables estimated by a GP; e.g. HR, HR′ and $\hat{\text{HR}}$ represent the observational, the continuous (zero-order hold model), and the posterior mean (GP model) heart rate data, respectively; this nomenclature is applied to the remaining vital signs.

FiO_2 support, the percentage of patients with FiO_2 support in the same time windows is determined for each of the three groups. The trajectories for the discharged, admitted and event patient groups, are represented in blue, green, and red, respectively, in Figures 9.2, to 9.5, as described next.

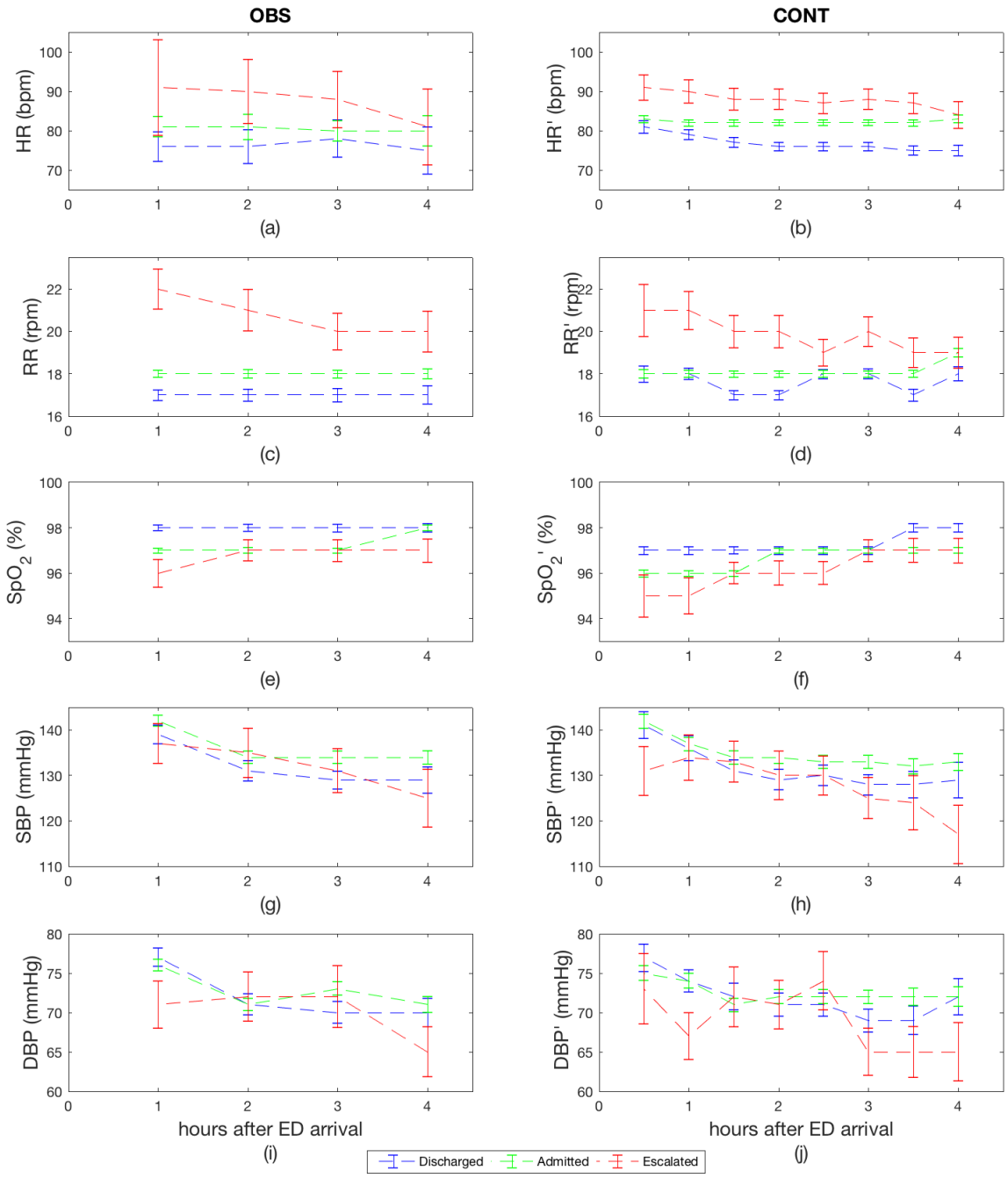


Figure 9.2: *Average trajectory of vital-sign data from the ED arrival time up to 4 hours afterwards for discharged, admitted and escalated ED patients, represented in blue, green and red, respectively. The left and right plots represent observational and continuous vital-sign data, respectively.*

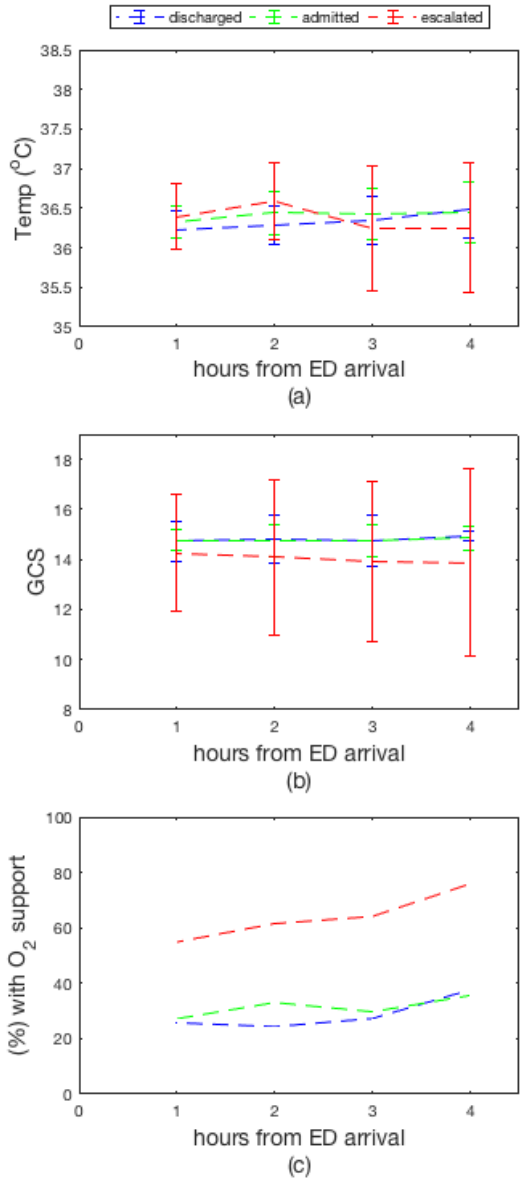


Figure 9.3: *Average trajectory, from arrival up to 4 hours later, of temperature, GCS and use of oxygen support, recorded in T&T charts of discharged, admitted and escalated ED patients, represented in blue, green and red, respectively.*

Physiological trajectory up to four hours of ED stay

Figure 9.2 represents the physiological trajectory, for 5 vital signs, from the patient arrival up to four hours afterwards. In this figure we observe that, at ED arrival escalated patients present higher mean HR and HR′ ($>$ 90 bpm versus $<$ 85 bpm); higher mean RR and RR′ ($\geq$ 21 bpm versus $\leq$ 18 bpm); and lower mean SpO_2 ($<$ 97% versus $\geq$97%) and mean

SpO_2' ($< 96\%$ versus $\geq 96\%$), than the corresponding averages for the no-event patients (both the discharged and admitted groups).

The plots in Figure 9.3 show the trajectories, from arrival up to four hours of ED stay, for those vital signs only present in the T&T charts, namely the average TEMP and GCS, and the percentage of patients under active oxygen therapy ($FiO_2 > 0.21$), for the same hourly time-windows. In Figure 9.3.c, it can be observed that oxygen therapy is available for most of the escalated patients, e.g. for approximately 80% of them, four hours into their ED stay, with the objective of improving their SpO_2.

In Figure 9.2, escalated patients are also observed to have lower mean BP$'$ than no-event patients (both groups) at ED arrival and after 4 hours in the ED. Finally, we note that admitted patients present marginally higher average HR and HR$'$ and average RR and RR$'$ values than discharged patients, i.e. patients discharged home present, on average, a more stable physiology than those admitted to the hospital.

Physiological trajectory up to the patient outcome time

Figure 9.4 shows the vital-sign trajectory from 4 hours before one of the three possible outcomes (discharge home, admission to the next ward, escalation to resus), to the time when that patient outcome occurs. In this case, the time-series time is normalised with respect to the outcome time ($t_{outcome} = 0$ hours), and then presented from 4 hours before the outcome.

The mean HR of escalated patients presents a different trend, with respect to Figure 9.2: it increases, as the patients get closer being escalated to the resus area. The HR$'$ of event patients remains higher than that of the no-event patients. As before, the presence of the oxygen mask treatment is the likely cause for the improvement in the SpO_2 values. Nevertheless, the mean SpO_2' for the event patients is still significantly lower before the escalation time than that of no-event patients discharged home, at ED discharge time. The mean BP and BP$'$ is indistinguishable between the groups for most of the trajectory. More specifically, the SBP and the DPB are significantly lower (< 130 mmHg, and < 75

mmHg, respectively) near the escalation time, when compared to the value for the non-event patients at the ED discharge time (> 130 mmHg, and > 75 mmHg, respectively). The SBP$'$ and DBP$'$, also shows lower values at the escalation time, but not significantly different[2] from those of the no-event patients, at their discharge time.

It is also interesting to observe in Figure 9.5 that the FiO_2 support seems to be gradually decreased during the course of treatment, up to the outcome time, for all groups, with a marginal increase, near the escalation event time, for the event patients. As with Figure 9.2, the average TEMP and GCS trajectories are not significantly different, between the groups.

[2]In this section we have considered significant differences between the means, at each point of the physiological trajectory plots, when their standard error bars did not overlap (and the difference between the means is not significant otherwise).

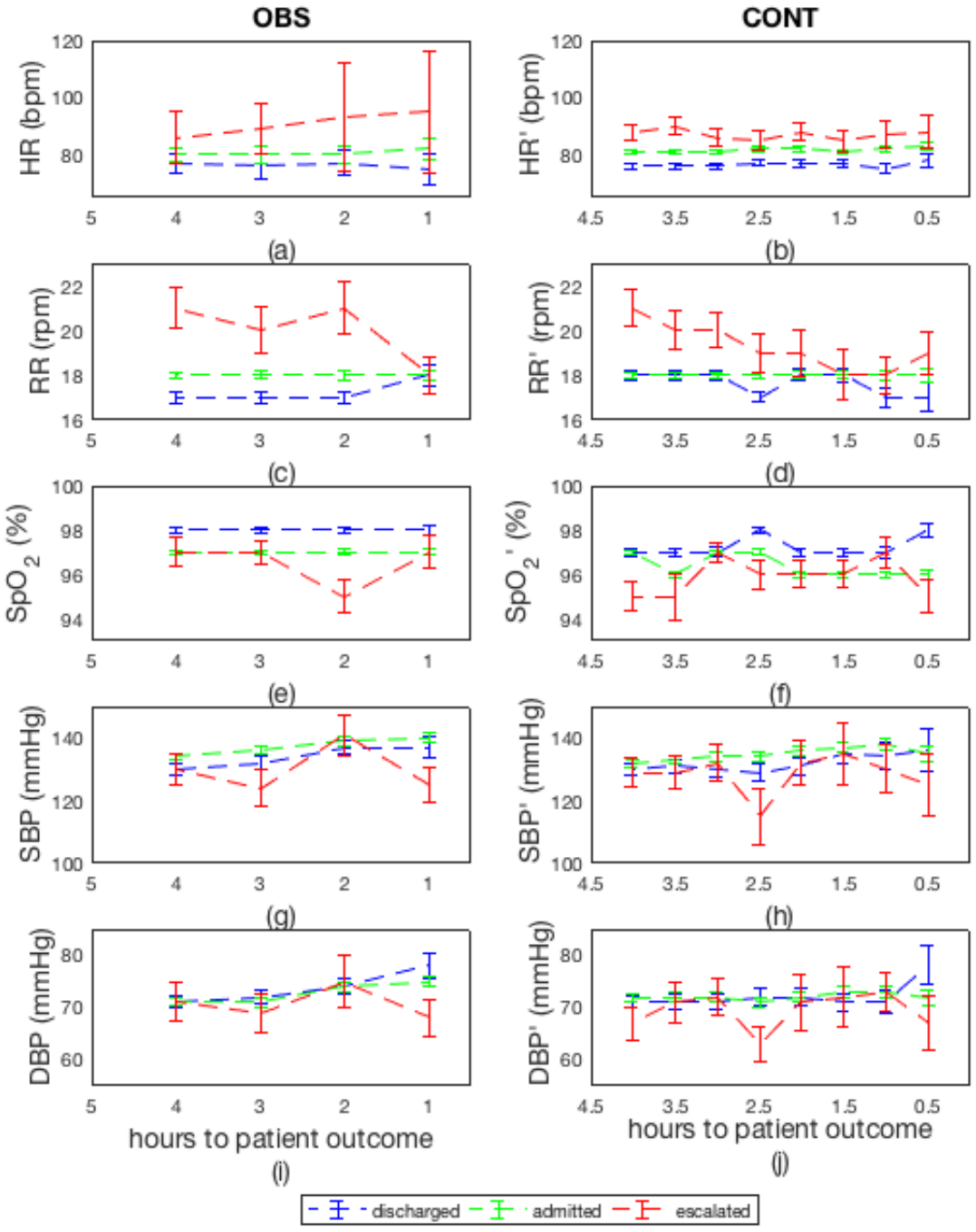


Figure 9.4: *Average trajectory of vital signs from 4 hours prior to the patient outcome time to the patient outcome time, for discharged, admitted and escalated ED patients, represented in blue, green and red, respectively. The left and right plots represent observational and continuous vital-sign data, respectively.*

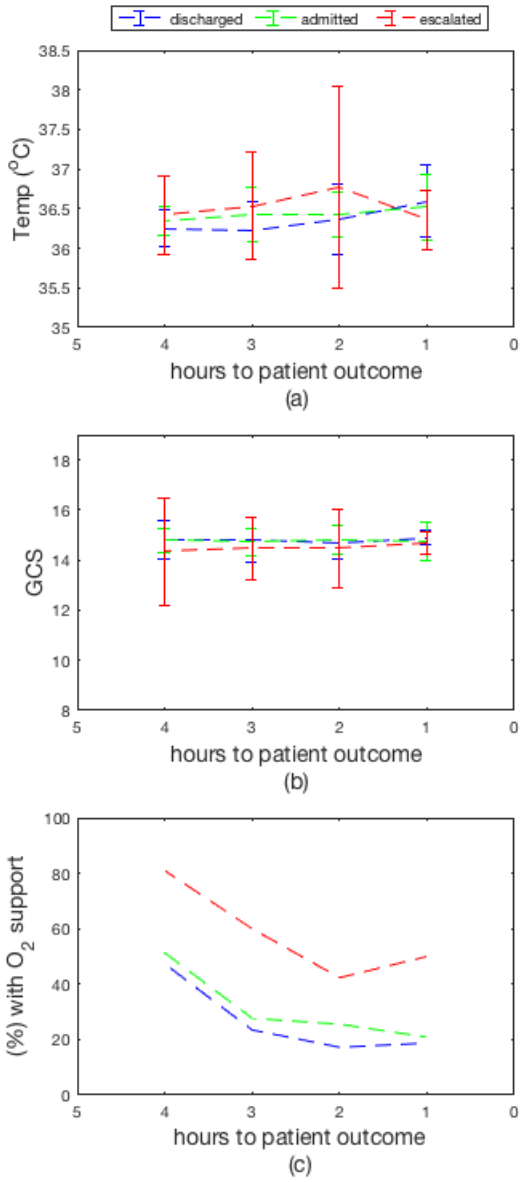


Figure 9.5: *Average trajectory, from 4 hours prior to the patient outcome until the patient outcome time, of temperature, GCS and use of oxygen support, recorded in T&T charts of discharged, admitted and escalated ED patients, represented in blue, green and red, respectively.*

Analysis of univariate distributions

We now compare the trajectories of the vital-sign distribution between the patients considered "stable" and those considered "unstable". To that end we compute the distance between the distribution of continuous vital-sign data from those patients considered "stable" and the vital-sign distributions of different periods of ED stay G_i for each of the three previously analysed patient groups. G_i is defined as the vital-sign data within the following ED stay intervals i: G_1 = [0, 25[%; G_2: [25, 50[%; G_3 = [50, 75[%; and G_4 = [75, 100]% of the patients' ED stay time. Using this approach, the entire vital-sign time-series can be visualised (unlike in the previous analysis). The distances between the vital-sign distributions are determined using the following metrics:

- the Kolmogorov-Smirnov distance (Massey, 1951):

$$\Delta KS(p(\mathbf{x}), q(\mathbf{x})) = sup\left(|P(\mathbf{x}) - Q(\mathbf{x})|\right), \tag{9.1}$$

  corresponds to the maximum distance between two PDFs, $p(\mathbf{x})$ and $q(\mathbf{x})$, of a random variable $\mathbf{x}$. $P(\mathbf{x})$ and $Q(\mathbf{x})$ are their CDFs, respectively. $sup(\mathbf{d})$ is the supremum of the set of distances $\mathbf{d}$;

- the Symmetrical Kullback-Leibler distance (Veldhuis, 2002):

$$\Delta KL(p(\mathbf{x}), q(\mathbf{x})) = \frac{1}{2}\sum_{\mathbf{x}\in\mathbf{X}}(p(\mathbf{x}) - q(\mathbf{x}))log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right), \tag{9.2}$$

  compares the entropy between two distributions over the same random variable, quantifying the additional information (bits) required when encoding a random variable with distribution $p(\mathbf{x})$ using the alternative distribution $q(\mathbf{x})$;

- the Bhattacharya distance (Bhattachayya, 1943):

$$\Delta Bhat(p(\mathbf{x}), q(\mathbf{x})) = -log\left(\sum_{\mathbf{x}\in\mathbf{X}}\sqrt{p(\mathbf{x})q(\mathbf{x})}\right), \tag{9.3}$$

measures the amount of overlap between two PDFs, $p(\mathbf{x})$ and $q(\mathbf{x})$, of the random variable $\mathbf{x}$.

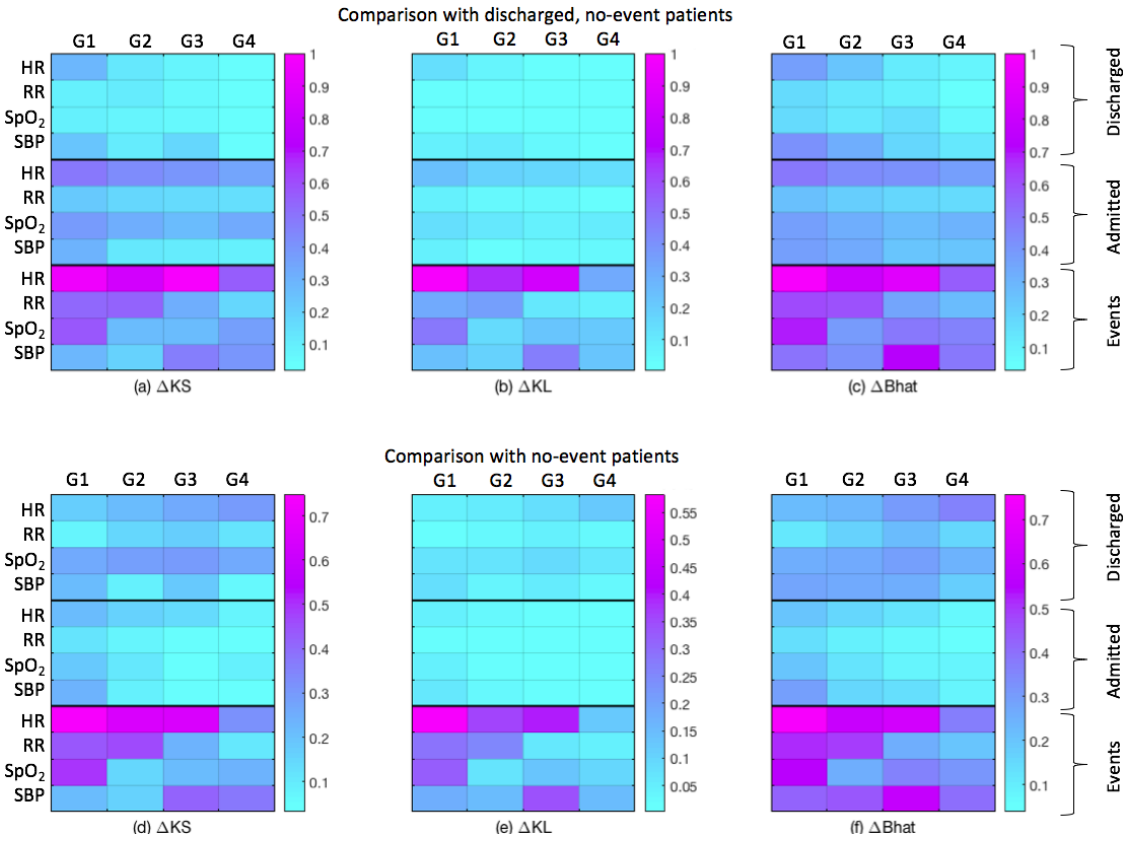


Figure 9.6: *Top row: Distance between the distributions of continuous vital-sign data (HR', RR', SpO_2' and SBP') from fractions of ED stay time G_i of patient groups E and all data from group $E_{\{1,2,3\},d}$ (stable patients discharged home). Bottom row: Same as above but the distances are computed against all data from groups $E_{\{1,2,3\},d}$ and $E_{\{1,2,3\},a}$ (all no-event patients). (a) and (d) represent the Kolmogorov-Smirnov distances ΔKS; (b) and (e) represent the symmetric Kullback-Leibler distances ΔKL; and (c) and (f) represent the Bhattacharya distances $\Delta Bhat$. Note that the plots with the same distance metric are normalised by the maximum value of that distance metric.*

Two data visualisation experiments are performed:

- Experiment 1: Each of the three distance metrics are computed between each continuous vital-sign data period G_i of each patient of the three outcome groups, and all data from no-event patients, discharged home.

- Experiment 2: Each of the three distance metrics are computed between each continuous vital-sign data period G_i of each the three outcome groups, and all data

from all no-event patients (discharged and admitted to a hospital ward).

The distances for each experiment are illustrated in the top and bottom rows of Figure 9.6, respectively, and are normalised by the maximum distance value computed by distance metric (i.e. the same normalisation is applied to both experiments (i) and (ii), by distance metric, for visual comparison).

Two conclusions can be derived from this figure:

- Escalated patients show higher distance from "stable" patients data at the beginning of the ED stay for the HR′, RR′ and SpO_2' distributions, for all the distance metrics. Towards the escalation time, these vital signs show some recovery, i.e. the distances between the distributions become smaller. However, the SBP of escalated patients seems to deteriorate near the escalation time, its distances being higher at 75% of the patients' ED stay.

- If the vital-sign distribution trajectories are compared against the group of patients discharged home (in experiment 1), then the distances with respect to the vital-sign distribution of the group of patients admitted to the hospital, are higher than those of experiment 2. The difference between the no-event discharged and admitted patients can be reduced by using all their data, as in experiment 2. This was done as well to model the ED-CEWS thresholds, in chapter 5, when we had the same objective of maximising the separation between the no-event ED patients admitted to the hospital and the escalated patients. This will make the model more likely to alert on the abnormal vital signs that precede an escalation to resus area.

9.2.2 Multivariate data visualisation

Multivariate trajectory of the novelty score

We now consider a multivariate representation of the trajectory of the patients' vital signs during their ED stay. To this end, we make use of the KDE novelty detection

model to evaluate the likelihood of the multivariate data with respect to a model of normal physiology, and its visualisation in one dimension. The following steps are used to produce this representation:

1. **Novelty score**: First the novelty score, $z(\mathbf{x})$, is computed for both observational and continuous data for the entire dataset, by applying the baseline KDE model described in section 6.3.1. In the case of the clinical observations, we also make use of the available temperature values. Missing values are set to the mean of the training data for this data-source, as done in section 7.4.1. The novelty score is then averaged hourly and in 30-minute windows, for observational and continuous data, respectively, for each patient. Finally, the average between all time-coinciding time-windows is computed to represent the average trajectories.

2. **Physiological Trajectories**: The novelty score trajectories are averaged, using the same methodology as for the univariate trajectory analysis, i.e. from the patients' arrival up to 4 hours of their ED stay, and from 4 hours before the patients' outcome time to the time of one of the three possible outcomes. The trajectories are shown in Figures 9.7 and 9.8, respectively, for both observational and continuous data, represented as $\hat{z}(\mathbf{x})$ and $\hat{z}(\mathbf{x})'$, respectively.

For comparison, the CEWS average trajectories are also computed using the same strategy, and presented in the same figures.

Multivariate trajectory up to four hours of the ED stay

As observed in Figure 9.7, the mean novelty score is consistently higher for event patients, and increases during the 4-hour ED stay, for both the observational and continuous data, represented on the top left and right plots, respectively. For example, for continuous data, $\hat{z}(\mathbf{x})' > 8.5$ for event patients, versus $\hat{z}(\mathbf{x})' < 8$ for no-event patients.

The $\hat{z}(\mathbf{x})'$ for those patients admitted to the next hospital ward is marginally higher

than that of discharged patients. There seems to be no distinction between the observational data trajectories of the no-event patient groups (in Figure 9.7a).

We note that, as shown in Figures 9.7e and 9.7f, the number of data-points contributing to estimate the average trajectories of any patient group beyond 5 hours of ED stay is low, for both observational and continuous data. This justifies the analysis of the interval from ED arrival up to 4 hours afterwards (the same as for the univariate cases, analysed in the previous section).

There is no distinction in the average CEWS trajectories between the groups, at patient arrival. The average CEWS increases marginally to 1 (when rounded to the nearest integer), near the end of the event patients' ED stay.

Multivariate physiological trajectory up to the patient outcome

As observed in the univariate case, for the multivariate case, the $\hat{z}(\mathbf{x})$ and $\hat{z}(\mathbf{x})'$ of event patients, shown in Figures 9.8a and 9.8b, are consistently higher than that of no-event patients. The difference between the means is significant for the entire trajectory of the continuous data case. A recovery trajectory can also be visualised for the latter, i.e. the novelty score decreases throughout the event patients ED stay (but the average score is still significantly higher than that of no-event patients, at the outcome time).

The no-event patients present stable physiological trajectories, $\hat{z}(\mathbf{x})'$ being marginally higher for those admitted to a hospital ward, also observed for some vital signs in the univariate analysis.

With the CEWS, there is no difference between the physiological trajectories for the observational data. In the continuous data case the average CEWS increases marginally, to 1(when rounded to the nearest integer) just before the escalation time, for event patients, as observed in the previous case. However it is not significantly different at the outcome time.

In summary, we have shown that on average the novelty score indicates a more unstable physiological trajectory for the escalated patients, when compared with no-event

patients, and much earlier than for the CEWS system.

Neuroscale Visualisation of ED vital-sign time-series data

In novelty detection applications, visualisation tools such as the Neuroscale algorithm Lowe and Tipping (1997) and the Sammon Mapping Sammon (1969), can be used to study how the high dimensional feature vectors are distributed over the space of normal data, in the 2-D or 3-D space, especially near the boundaries of normality.

We use the Neuroscale algorithm to visualise the multivariate ED time-series data in 2D. Given an high i-dimensional input space of N data points $\mathbf{x}_i$, a Gaussian RBF neural network is trained to predict a lower k-dimensional feature space $\mathbf{y}_k$, with $i >> k$. The output y_k is a linear function of the basis functions centred on the RBF units ϕ_j defined as:

$$\phi(\mathbf{x}_i)_j = \exp\left(\frac{||\mathbf{x}_i - \boldsymbol{\mu}_j||^2}{\sigma_j^2}\right), \tag{9.4}$$

where μ_j and σ_j are the centre and the width of the j^{th} unnormalised Gaussian RBF unit, respectively. The output layer results from the standard neural network weighted summation propagation formula:

$$\mathbf{y}_k(\mathbf{x}_i) = \sum_{j=1}^{J} w_{jk}\phi_j(\mathbf{x}_i) + w_{k_0}, \tag{9.5}$$

where w_{jk} and w_{k_0} are the weights and bias of the output layer k, and J is the total number of total centres. The network parameters (i.e. w_{jk} and w_{k_0}) are optimised using an iterative gradient descent algorithm, by minimising the error term (cost function):

$$E = \sum_{i=1}^{N}\sum_{k>i}^{N}\left(d_{ik}^* - d_{ik}\right)^2, \tag{9.6}$$

where d_{ik}^* and d_{ik} are the Euclidean distances between the points in the data space, and in the corresponding (lower dimensional) feature space, respectively.

The following steps are used to compute the Neuroscale representation from an i-

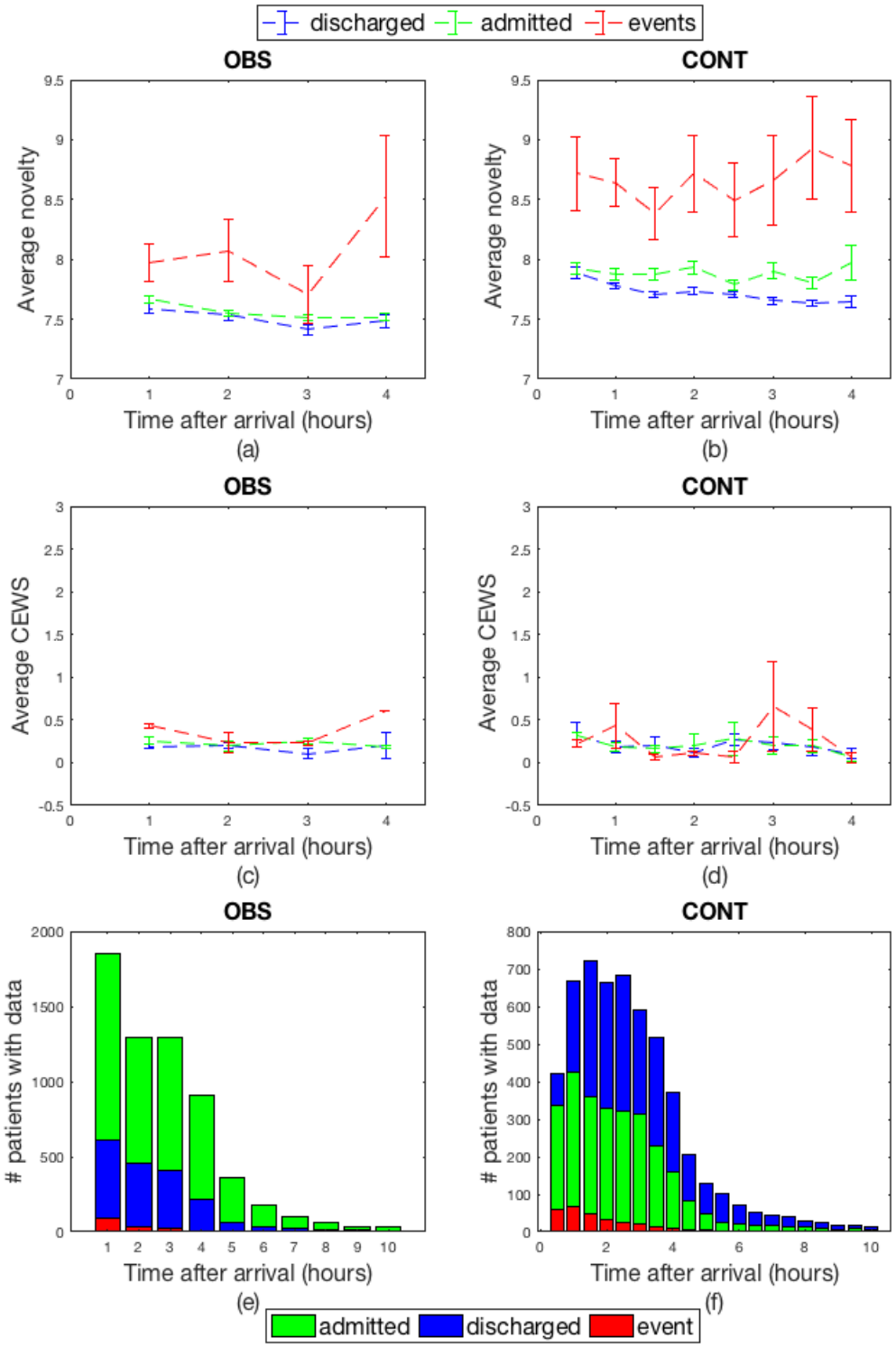


Figure 9.7: *Comparison of the multivariate vital-sign trajectories from arrival in the ED to the event or ED discharge, represented by the mean novelty score $\hat{z}(\mathbf{x})$ (the novelty score is $z(\mathbf{x}) = -log\,(p(\mathbf{x}))$), and mean CEWS (the mean CEWS values were rounded to the nearest integer). When using the clinical observations, the novelty score trajectory is higher for (i) event patients and cannot be distinguished between no-event patients (ii) admitted to another hospital ward or (iii) discharged from the hospital. In the continuous data, it is possible to distinguish each of these trajectories (i, ii and iii). The plots in the last row show the number of patients used to compute each hourly or 30-minute score average for the observational or continuous data, respectively.*

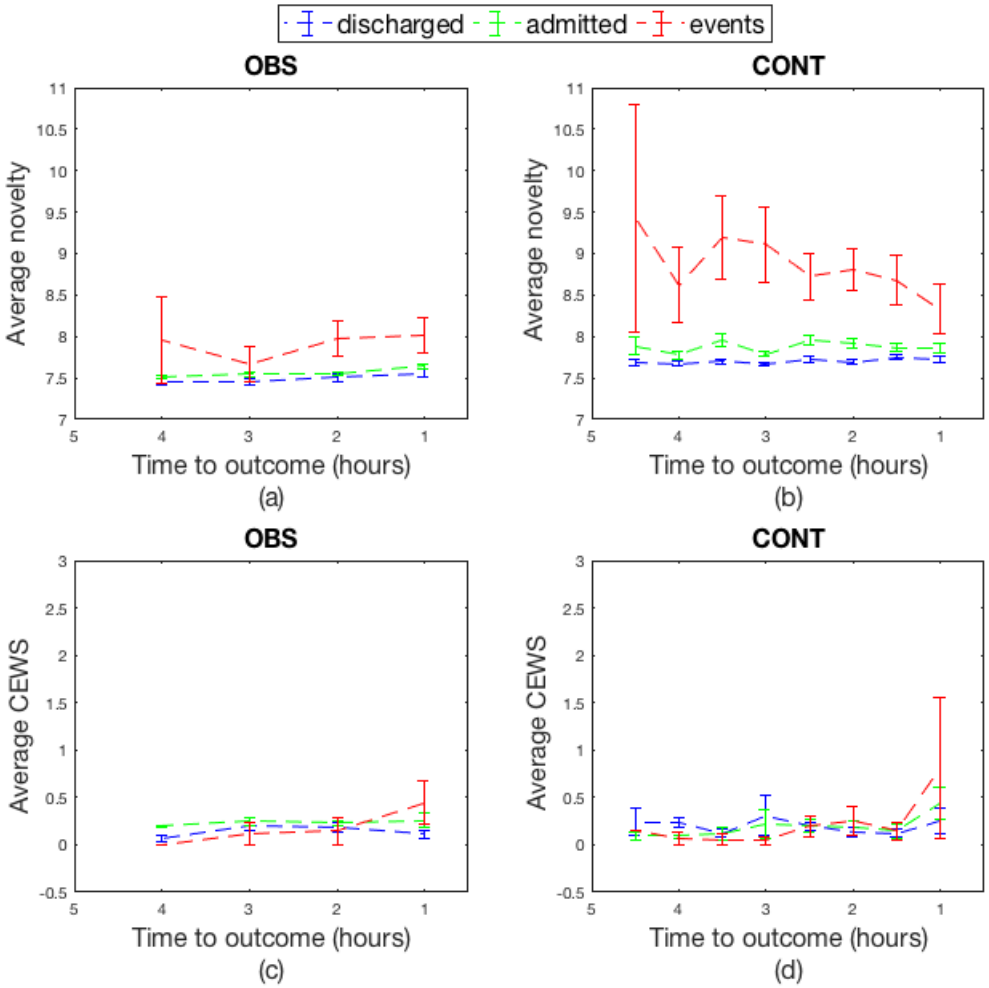


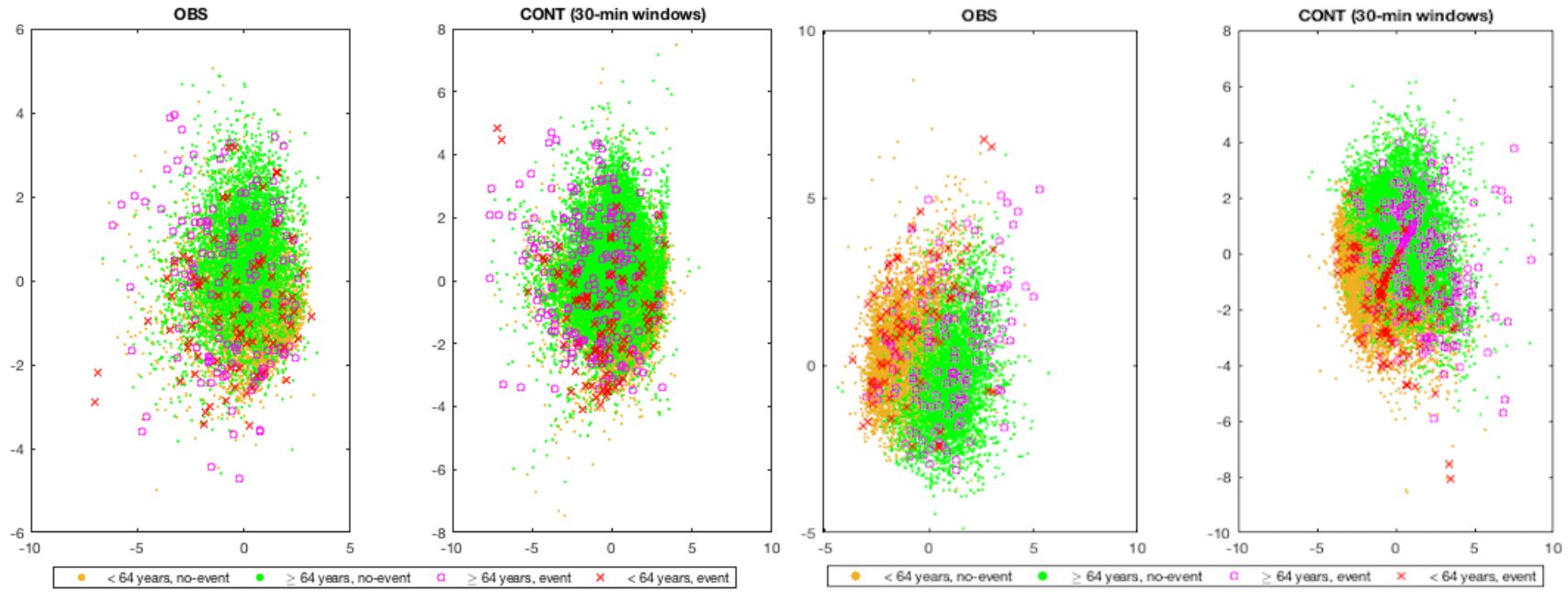
Figure 9.8: *Comparison of the multivariate vital-sign trajectories from 4 hours before the outcome to the outcome time, represented by the mean novelty score $\hat{z}(\mathbf{x})$ (novelty score is $z(\mathbf{x}) = -log(p(\mathbf{x}))$), and mean CEWS. The novelty score trajectory is higher for event patients. No-event patients present a stable trajectory, those admitted to the hospital presenting a marginally higher score than those discharged. For the continuous data, the $\hat{z}(\mathbf{x})$ values of the escalated patients show a recovery trend up to the event time, but the score is still higher for these patients.*

dimensional input space to a lower dimensional feature space $k = 2$ dimensions, using the implementation from Nabney (2002)[3]:

1. continuous data is first synchronised using the zero-order hold procedure;
2. all numerical variables from dataset $E_{\{1,2,3\}}$ are first normalised using the zero-mean unit-variance transform. All the remaining categorical variables are coded as binary variables in this chapter, i.e. the GCS is coded as 1 if GCS < 15 and 0 if GCS $= 15$, and the presence of FiO_2, coded as $\{0, 1\}$ as in chapter 5;
3. vectors with missing data were discarded;
4. only the unique data examples (vectors) are used (i.e. repetitions are discarded);
5. the number of RBF centres is allowed to change according to the expression $N_c = i \times 10 + g_c$, with $g_c \in \{0, 10, 20, 30, 40, 50\}$, where i is the number of features (from the i-dimensional input space);
6. the algorithm is then allowed to optimise the neural network parameters for each experiment (i.e. number of considered feature combinations, and number of centres) up to 60 iterations;
7. finally, the high dimensional data is mapped to the lower dimensional space by using equations 9.4 and 9.5; the final number of centres was selected by visual inspection.

As described in Tarassenko (1998), we found that $i \times 10$ Gaussian RBF centres, in the hidden layer, were enough to achieve a good visualisation of the data. For the continuous data, combinations of the feature group {HR, RR, SpO_2, SBP, DBP}, and age and sex, were used. For observational data, combinations with the temperature, the GCS, and the FiO_2, were also analysed additionally. The most relevant Neuroscale results are shown in Figure 9.9.

[3]The default settings are used, in which the RBF units centres μ_j are initialised using 5 iterations of the k-means algorithm, followed by 10 iterations of the EM algorithm, for isotropic Gaussian-mixture models; and the width parameter σ_j, which is equal for all units, is set to the largest distance between centres.



(a) *Features: HR, RR, SpO_2 and SBP.*

(b) *Features: HR, RR, SpO_2, SBP and age.*

Figure 9.9: *(a) Neuroscale models of HR, RR, SpO_2 and SBP, for observational (on the left plot) and continuous (30 minute window averages, on the right plot) vital-sign data from patients with and without events. The model is coloured according to patient outcome and according to age (with respect to the median, 64 years). (b) Neuroscale models that differ from (a) only in the addition of the age feature. To accommodate for the extra-feature 50 centres, instead of 40, are used. The use of age allows the vital signs of normal and abnormal patients to be clustered towards their age group. i.e. physiological abnormality might be better discriminated with the use of this feature.*

Four groups are coloured in this figure: young adult patients (i) with events (red) and (ii) without events (orange), and elderly patients (iii) with events (magenta) and (iv) without events (green). The median age of the patients in the dataset, 64 years old, was used to define the young (< 64) and elderly (≥ 64) groups in this figure. Figure 9.9a represents the mapping of HR, RR, SpO_2 and SBP to 2D, and Figure 9.9b uses age as an additional feature. We observe that in the Neuroscale representation, the data of patients with events overlapped with the data of those patients without events that were latter admitted to a ward for further investigation.

Furthermore, we found that the addition of age into the model (Figure 9.9b) can cluster the normal and abnormal physiology patterns according to age, i.e. the abnormal physiology of young patients might be better separated if evaluated against this group's normal patterns. Based on this analysis, we also make use of the demographics data in the physiological risk models, discussed next.

9.3 Physiological risk scoring models for multi-instance time-series data

9.3.1 Data preprocessing

Patient Labels: The dataset analysed in the previous section, $E_{\{1,2,3\}}$, is also used to develop the physiological risk scoring models. As in the previous chapters, the data from phases 1 and 2 of the large-scale study, are used for model training, with 1,670 no-event and 63 event patients (group $E_{\{1,2\}}$), and the data from phase 3 (group E_3) are used to test the performance of the models, with 1,037 no-event and 33 event patients.

Augmented MIGP time-series model: For the training data, each patients' vital-sign time-series (downsampled to one-minute window medians) are synchronised by instantiating a MIGP at coinciding 60-second timestamps. For the cases of missing data, the HR, RR, and the SpO_2 are estimated up to 5 minutes, and the SBP, DBP, and the Temp,

up to 30 minutes, as long as data from at least 3 other vital signs exist. The MIGP hyperparameters were obtained for each patient as described in step 1 in section 8.3.2. For both the validation and test data, we used the SMIGP derived in steps 2 and 3 in section 8.3.2, i.e. a filtering process (minute-to-minute step-ahead predictions).

We analyse the augmentation of the multi-instance time-series data, which include up to this point observational and continuous HR, RR, SpO_2, SBP, DBP and temperature data, with the level of consciousness (GCS), and the FiO_2 support, usually collected by the ED clinical staff. These parameters are coded as binary variables as described in the previous section. Their value is set to zero in the instances for which they are missing in the observation sets. For both parameters, the values are held to match the timestamps from the MIGP models, from the time of their respective observation set to the time of the next observation set, or to the end of the time-series. For both cases, their values are assumed to be equal to zero (i.e. the most normal value is assumed) when continuous data exists before the first observation set. The resulting data is denoted as GCS'_b and FiO'_2, respectively (the sub-index b denotes that the GCS was converted into a binary variable, and the superscript $'$ is used to indicate that the data were augmented to match the synchronised multi-instance time-series data).

Demographics: Demographics information such as the age, coded as a continuous-variable, and sex, coded as $\{0, 1\}$, were analysed as possible features of the physiological risk score model.

9.3.2 Novelty detection models

Each model considered for the multi-instance data, corresponds to a different combination of the feature group {HR, RR, SpO_2, SBP}, and the DBP, temperature, GCS, FiO_2, age, and sex. This results in 64 different combinations, for each of the models described below. Data examples with more than one feature missing from the group {HR, RR, SpO_2, SBP, TEMP}, are discarded, and the remaining missing data are set to the mean of the training data (or to the mode in the case of discrete variables). All continuous variables are then

normalised using the zero-mean unit-variance transform, of the training data.

Kernel Density Estimate (KDE)

The baseline KDE algorithm (described in section 6.3.1), is re-trained using the ED training data with the following adaptations: (i) the KDE centres are reduced to 500 centroids using the k-means[4] clustering model that results in the lowest cost from five runs (or the lowest inertia, which is defined as the sum of squared distances of the data samples to their closest cluster centre); (ii) then, the 400 centres with lowest Euclidean distance to the mean of the resulting data (i.e. the centroid of the data) are kept; (iii) finally, an isotropic Gaussian kernel is used to model the joint probability distribution of the continuous-variables, using equation 6.4. The kernel bandwidth σ, is found using the leave-one-out likelihood $J(\sigma)$ of the training data

$$J(\sigma) = \frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{N-1}\sum_{i=1, i\neq j} K(\mathbf{x}, \mathbf{x_i}|\sigma)\right) \tag{9.7}$$

where $K(\cdot)$ is a Gaussian kernel. A gradient descent method is employed to find the best $\sigma \in \mathbb{R}_+$, and equation 6.5 is used to initialise σ.

KDE for mixed continuous and discrete data

For models with mixed continuous and discrete variables, the k-prototypes algorithm, from Huang (1998), is used instead, to find the 500 prototype centres. This algorithm, like k-means, iteratively recomputes cluster prototypes and reassigns clusters. Clusters are assigned using the distance, $d(\mathbf{x}, \mathbf{y}) = d_1(\mathbf{x}, \mathbf{y}) + \lambda' d_2(\mathbf{x}, \mathbf{y})$, where d_1 is the Euclidean distance and d_2 is the number of matching discrete factors between $\mathbf{x}$ and $\mathbf{y}$ weighted by a λ' factor, initialised as the overall average of the standard deviations of the numerical attributes. Cluster prototypes are computed as cluster means for numerical variables and modes for the discrete variables. From five runs, that with lowest inertia is kept. d is used

[4]The k-means++ approach is used to initialise the centres.

to determine the inertia metric in this case; and it is also used to compute the 400 centres closest to the centroid, which results in the final compressed dataset. The generalised kernel product, from equation 6.10, is used to model the joint probability distribution of the mixed data features. As before, the kernel bandwidths σ, and $\boldsymbol{\lambda}$ (i.e. a different $\lambda \in \mathbb{R}_+$ for each discrete variable is used) are found using the leave-one-out likelihood $J(\sigma, \boldsymbol{\lambda})$ of the training data, defined by equation 6.11, using a gradient descent method. Equations 6.5 and 6.12 are used to initialise σ and each λ, respectively.

One-class support vector machine (OSVM)

This approach also allows mixed data models. The previous clustering approach is used to compress the normal data to 400 prototype centres. A grid search over (ν, σ) (the fraction of false positives in a novelty classification task, and the width parameter of the Gaussian kernel used in our OSVM configuration, respectively) is then performed, where $\sigma \in \{0.01, 0.1, 0.5, 1, 1.5, 2, 3, 4\}$ and ν is varied over values in the interval from 0.1 to 0.9, with increments of 0.05 (as suggested in Pimentel (2015)).

Model validation and test

Five-fold cross-validation is used to select the best model hyperparameters (e.g. the prototype centres, and the kernel bandwidths in the case of the KDE model, or the best (ν, σ) in the OSVM case). MIGP data are assigned to each fold patient-wise, i.e. patients are allocated to the folds and all their fused time-series data are also allocated to those folds. To avoid considering SpO_2 values above the mean as abnormal, those values are set to the mean value of the training data as in Wong (2011) (done also in the test phase). At model validation time, SMIGP data from all 63 event patients ($E_{\{1,2\},e}$) and the remaining 1/5 no-event patients in the fold, are used to determine the models' performance. The persistence criterion is used to generate alerts from the continuous data, as described in section 6.3.1. The AUC metric is determined through patient-wise performance analysis (section 8.3.4). The KDE and OSVM models with highest AUC in the validation dataset,

are selected for the testing phase.

Finally, as before, data from phase 3 (E_3) is used as the test set. Patient-wise performance analysis is also used to assess model performance, ranked by the AUC metric.

9.3.3 Logistic regression model

Logistic regression (LR) is a linear model for classification, widely used in healthcare problems because of its interpretability. In this model, the logit function is used to link the likelihood of the linear combination of data from exemplar $\mathbf{x}_i$ (independent variables) belonging to the outcome class, $y_i = 1$ (dependent variable). The LR model is of the form:

$$p(y = 1|\mathbf{x}; \mathbf{w}) = f(w^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})},$$

where $\mathbf{w}$ is the vector of model coefficients, and $f(\cdot)$ represents the sigmoid function, defined by the second equality. Predicted parameters above a pre-specified threshold are considered to be from the positive class, $y_i = 1$, and the negative class, $y_i = 0$, is assumed otherwise. In this section, the L2 regularisation procedure is used to optimise the LR hyperparameters, i.e. the following cost function is minimised:

$$J(w, c) = \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{x}_i^T\mathbf{w} + c))) + \frac{C}{2}\mathbf{w}^T\mathbf{w}, \tag{9.8}$$

where $\mathbf{w}$ and c are the LR model coefficients and bias parameters and C is the regularisation term (Hastie et al., 2009).

Spline Logistic Regression

We use spline regression to fit the LR model parameters to different quantiles of the continuous variables with symmetrical regions of abnormality, and normal physiology in-between. This modelling procedure requires the definition of two additional hyperparameters in our LR model:

1. The number of knots, k: $\mathcal{S} : [a, b] \Rightarrow \mathbb{R}$, is a piecewise spline function defined by k subintervals that partition $[a, b] = \cup[s_j, s_{j+1}], j = 0, ..., k-1$, where the s_j parameters are called knots. Equidistantly distributed knots (uniform spline) are used;

2. The degree of the polynomial p: In each subinterval $[s_j, s_{j+1}]$, S is defined by a polynomial function $\boldsymbol{\beta}_j$ of degree order p.

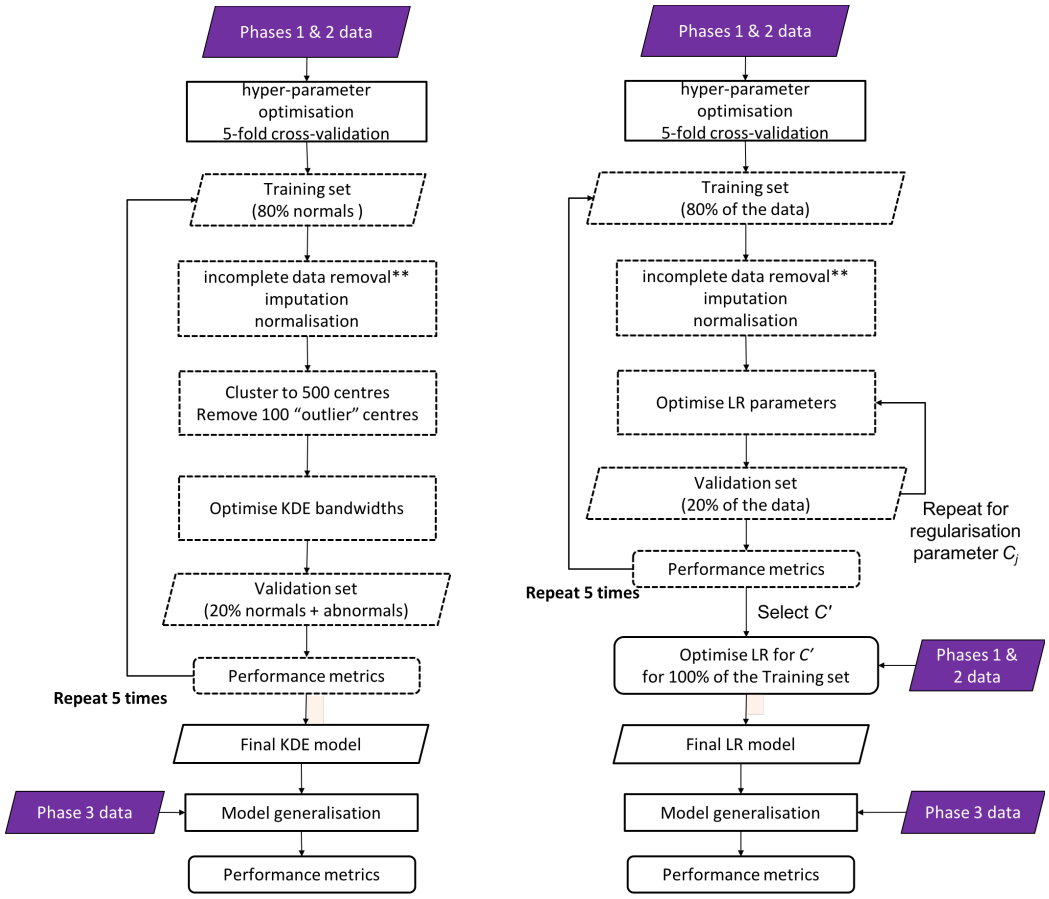
We vary the number of knots in the set $k \in \{3, 6, 9\}$, and the polynomial function degree is optimised within the set $p \in \{1, 2, 3\}$. Spline regression is applied to the same set of feature combinations as defined in the previous section. The sex, FiO_2' and GCS′ are coded as a binary variables, also as in the previous section. The patient-wise data selection, data imputation, and data normalisation approaches, used in the previous section, are kept for the LR models. Both the negative and the positive classes (no-event and event patients, respectively) are included in model training as follows.

LR model training occurs in two stages: (i) Five-fold cross-validation is used to select the optimal regularisation parameter C. In each fold C is changed from 10^{-6} to 50 (using 15 equidistant points on the log scale), for each spline regression configuration (k, p). The parameters that minimise the L2-based cost function are found using the training set and coordinate descent[5]. The training set includes time-series data of 80% of the normal and abnormal patients from phases 1 and 2 (from dataset $E_{\{1,2\}}$), and the remaining are used for model validation. The patient-wise performance analysis is used to compute the AUC on the validation dataset. The results for each regularisation term C are averaged over all five folds, and that with the highest average AUC is selected; (ii) The final spline LR-L2 model hyperparameters ($\mathbf{w}_j$, c_j, corresponding to each spline polynomial $\boldsymbol{\beta}_j$) are trained using gradient descent on the entire training dataset, by fixing the regularisation term to the pre-optimised value, and for a given spline configuration (k, p). We found that setting high SpO_2 values to the mean value of the training set,

[5]The Logistic regression implementation from (Pedregosa et al., 2011) is used, and the package from http://pypi.python.org/pypi/patsy/ is used to code the spline regression.

did not improve model performance, as in this case, it is likely that the LR model learns small coefficients for the polynomial of the knot encoding high SpO_2 (hence this heuristic is not used in these models).

Finally, the patient-wise performance analysis is used to assess the models' performance by the AUC metric, on test dataset E_3. The novelty detection and the LR models' validation and testing approaches are illustrated in Figure 9.10.



(a) *KDE - novelty detection.* **(b)** *LR-L2 classification.*

Figure 9.10: *Comparison of model optimisation and generalisation procedure between the KDE and the LR-L2 (Logistic Regression trained using L2 regularisation) model approaches. The OSVM procedure follows that of the KDE. No data compression was used when applying these models to observational data. Dashed lines represent steps that are repeated five times (i.e. used in the five-fold validation process). **Training data samples with one feature missing from the group {HR, RR, SpO_2, SBP, TEMP} are removed.*

9.3.4 Baseline models and performance analysis plan

Table 9.1 summarises the models compared in this chapter. The following baseline models are also implemented, for a more complete analysis of the benefits of the multi-instance data models:

Observational data models:

- EWS models (Table 9.1, a.1.i): The CEWS, NEWS and DT-EWS systems, are applied to the observational data as described in chapter 5.
- ML models (Table 9.1, b.1.i): The KDE, OSVM and spline LR models are also trained using observational data. These data are already synchronised, at the observations times. Missing values are set to the mean (or mode, in the case of discrete data) of the training data. The same feature combinations used in the previous section are used here. However, as the amount of data is much lower (by at least 1 order of magnitude), we do not apply the clustering step in the KDE and OSVM models. In addition, we use the baseline KDE model, which we code as PSI_5 (following from the notation used in chapter 8).

Continuous data models:

We consider "single-instance" continuous vital-sign data models, synchronised using the zero-order hold procedure (such as the baseline data-fusion system model, described in section 6.3.1). For these data, we train ML models resulting from the combination of the feature group {HR, RR, SpO_2, SBP} and DBP, age and sex, as the remaining vital-signs and clinical context, i.e. TEMP, GCS, and FiO_2 support, are not available.

- EWS models (Table 9.1, a.2.ii): The CEWS, NEWS and DT-EWS systems, are applied to the synchronised continuous data. The persistence criterion is used to generate alerts.
- ML models (Table 9.1, b.2.ii): The KDE, OSVM and spline LR models are also trained and applied to the synchronised continuous data, using the same method-

ology as in the previous section. In addition, we apply the baseline KDE model, which we code as PSI_6.

Multi-instance data models (i.e. Fusion of observational and continuous data):

- EWS models (Table 9.1, a.2.iii): The CEWS, NEWS and DT-EWS systems, are applied to the SMIGP time-series data. The original GCS values are held out in this case, being represented by GCS′, to denote that they are matched to the time-series as defined in the previous section. The persistence criterion (section 6.3.1), is also used to generate alerts in these data.

- ML models (Table 9.1, b.2.iii): In addition to the models described in the previous section, the baseline KDE model, PSI_4 (defined in section 8.3.4), is also applied in this section, for comparison.

The validation and test procedures from the previous section are applied to all these models.

Table 9.1: *Overview of the data-fusion approaches evaluated in this chapter. The red ✓, represents the multi-instance data based system developed in this chapter. The baseline systems, highlighted in blue, are also computed and presented in this chapter. Observational data generate alerts if their score is above a pre-specified threshold. Continuous data require a persistence criterion to generate clinically relevant alerts. CONT - continuous data; OBS - Observational data; ML - Machine Learning; EWS - Early Warning Score.*

Models	**Alert strategy**	**Data source model**		
		(i) OBS	**(ii) CONT**	**(iii) Multi-instance**
(a) EWS	(1) $\geq$ threshold	✓	-	-
	(2) persistence criterion	-	✓	✓
(b) ML	(1) $\geq$ threshold	✓	-	-
	(2) persistence criterion	-	✓	✓

9.4 Results

We first consider the difference in model performance when using observational and continuous data, separately, shown in Tables 9.2 and 9.3, respectively. The performance on

the training and the test datasets, are represented in each table by $\text{AUC}_{E_{\{1,2\}}}$ and AUC_{E_3}, respectively. In both cases, the ML models re-computed from the ED dataset present a higher AUC_{E_3} than that of the EWS or the baseline PSI models. The continuous data KDE model achieved a higher AUC_{E_3}, 0.699 (0.580, 0.792) than the best observational data model (also the KDE model, with an AUC_{E_3} of 0.682 (0.571, 0.772)). This occurred with the addition of the DBP and the age features, not used by the PSI or the EWS models. For these data cases, amongst the models with lower ranking, the results are as expected: the CEWS and PSI models perform better in the continuous data case (as they are modelled from continuous data), and the NEWS and DT-EWS models perform better for the observational data (as they are tuned from, and for, observational data).

Table 9.4 shows the performance for the multi-instance data based models. The results show the ideal scenario, in which, if the fusion of complementary observational and continuous data is modelled correctly, it maximises the physiological information available for any given patient, and provides a more accurate representation of the patient physiology. For all the re-computed ML models, a combination of multi-instance data features could be found (best AUC_{E_3} was 0.737 (0.623, 0.830)), that achieved a higher performance than that of the EWS (best AUC_{E_3} was 0.643 (0.522, 0.749), when applied to observational data, as in current practise) or the baseline PSI (best AUC_{E_3} was 0.678 (0.550, 0.778), when applied to continuous data, as in current practise) models, on any of the considered data regimes (observational, continuous or multi-instance data). The combination of multi-instance data features also achieved a higher performance than any version of the ML algorithms trained on each data source, i.e. separately. We note however that, as in chapter 5, due to the class imbalance the difference between the model's AUCs_{E_3} values is not statistically significant.

Finally, we note that the KDE model trained from the observational data presented a higher performance on the training set ($\text{AUC}_{E_{\{1,2\}}} = 0.694$ (0.613, 0.763)) than any model from the other two cases (best $\text{AUC}_{E_{\{1,2\}}}$ were 0.673 (0.612, 0.729) and 0.657 (0.591, 0.715), for the multi-instance and continuous data LR models, respectively).

Table 9.2: *Performance of the best ML algorithms (KDE, OSVM, and LR-L2) versus that of EWS systems, to detect patients escalated to the ED resus area, when using the clinical observation sets data.* [a] *LR spline model with three equidistant knots and quadratic polynomials.*

MODEL	CONTINUOUS	DISCRETE	AUC_{E_3} (95% CI)	$AUC_{E_{\{1,2\}}}$ (95% CI)
KDE	HR, RR, SPO_2, SBP, DBP, TEMP	GCS_b	0.682 (0.571, 0.772)	0.694 (0.613, 0.763)
OSVM	HR, RR, SPO_2, SBP, DBP	GCS_b, SEX	0.675 (0.571, 0.766)	0.660 (0.572, 0.738)
LR-$L2^a_{3,2}$	HR, RR, SPO_2, SBP, DBP, AGE	GCS_b, FiO_2	0.660 (0.571, 0.741)	0.651 (0.589, 0.705)
NEWS	HR, RR, SPO_2, SBP, TEMP	GCS, FiO_2	0.643 (0.522, 0.749)	0.658 (0.596, 0.718)
DT-EWS	HR, RR, SPO_2, SBP, TEMP	GCS, FiO_2	0.636 (0.531, 0.749)	0.652 (0.590, 0.715)
PSI_5	HR, RR, SPO_2, SDA, TEMP	-	0.626 (0.515, 0.729)	0.593 (0.512, 0.665)
CEWS	HR, RR, SPO_2, SBP, TEMP	GCS	0.604 (0.492, 0.707)	0.619 (0.560, 0.674)

Table 9.3: *Performance of best ML models (KDE, OSVM, and LR-L2) to detect patients escalated to the ED resus area, when using the bedside monitors continuous data.* [a] *LR spline model with three equidistant knots and cubic polynomials.*

MODEL	CONTINUOUS	DISCRETE	AUC_{E_3} (95% CI)	$AUC_{E_{\{1,2\}}}$ (95% CI)
KDE	HR′, RR′, SpO'_2, SBP′, DBP′, AGE	-	0.699 (0.580, 0.792)	0.608 (0.533, 0.676)
LR-$L2^a_{3,3}$	HR′, RR′, SpO'_2, SBP′, AGE	SEX	0.696 (0.602, 0.783)	0.657 (0.591, 0.715)
OSVM	HR′, RR′, SpO'_2, SBP′, DBP′	SEX	0.678 (0.536, 0.784)	0.600 (0.519, 0.681)
PSI_6	HR′, RR′, SpO'_2, SDA′	-	0.657 (0.521, 0.755)	0.597 (0.513, 0.676)
CEWS	HR′, RR′, SpO'_2, SBP′	-	0.582 (0.458, 0.703)	0.574 (0.494, 0.643)
NEWS	HR′, RR′, SpO'_2, SBP′	-	0.579 (0.449, 0.700)	0.586 (0.516, 0.653)
DT-EWS	HR′, RR′, SpO'_2, SBP′	-	0.577 (0.456, 0.687)	0.584 (0.518, 0.649)

Table 9.4: *Performance of best ML algorithms (KDE, OSVM, and LR-L2) versus that of EWS systems, to detect patients escalated to the ED resus area, using the multi-instance data (fused observational and continuous data).* [a] *LR spline model with three equidistant knots and cubic polynomials.*

MODEL	CONTINUOUS	DISCRETE	AUC_{E_3} (95% CI)	$\text{AUC}_{E_{\{1,2\}}}$ (95% CI)
KDE	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	SEX, GCS'_b	0.737 (0.623, 0.830)	0.620 (0.545, 0.690)
OSVM	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	SEX, GCS'_b	0.711 (0.593, 0.807)	0.647 (0.565, 0.717)
$\text{LR-L2}^a_{3,3}$	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	SEX, FiO'_2	0.698 (0.611, 0.783)	0.673 (0.612, 0.729)
PSI_4	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SDA}}$, $\hat{\text{TEMP}}$	-	0.678 (0.550, 0.778)	0.621 (0.544, 0.692)
CEWS	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$	GCS'	0.642 (0.520, 0.749)	0.627 (0.558, 0.694)
NEWS	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$	GCS', FiO'_2	0.625 (0.492, 0.747)	0.639 (0.569, 0.703)
DT-EWS	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$	GCS', FiO'_2	0.633 (0.511, 0.743)	0.642 (0.573, 0.704)

The amount of data available to generate alerts in observational data is one order of magnitude lower than that available for continuous data (even when considering the persistence criterion). It is therefore possible that a marginal higher number of false positive alerts, might have contributed for the lower $\text{AUC}_{E_{\{1,2\}}}$ in those models using continuous data. We note that, as in chapter 6, due to the class imbalance the difference in the AUCs was not significantly different. The 39 ML models with highest performance are shown in appendix B, for all the ML approaches and data regimes.

9.5 Discussion

In this chapter the vital-sign patterns of "stable" and "unstable" patients undergoing treatment in the ED were analysed. In our analyses the "unstable" patients comprised those requiring escalation to the resus area during their ED stay. The other patients were deemed to have "stable" physiology (without events). Three main physiological trajectories were discussed (Figure 9.7): (a) patients that arrived with abnormal physiology, and continued to deteriorate requiring escalation to the resus area (event patients); (b) patients that arrived with unstable physiology, did not deteriorate - i.e. were not escalated during their ED stay -, but required more observation beyond the 4-hour ED LOS (no-event ED patients admitted to the next ward); and (c) patients that arrived with stable physiology (or low risk of deteriorating), and remained stable during their ED stay (no-event ED patients discharged home).

We first observed that the average physiological trajectory of no-event patients differs from that of event patients from the first hour of their arrival in the ED, being much more abnormal during that time. Our hypothesis is that patients are escalated to the resus area, after ED arrival, if (i) they are being treated but continue to be abnormal, (ii) their risk of deterioration is masked at first on arrival, or (iii) there are no beds available in the resus area (at the time of their arrival to the ED). Option (i) seems more likely, as, in our dataset, on average, event patients present a small recovery trend in their

vital signs during their ED stay, but still require a higher level of care. However, we note that if resus beds are unavailable (either due to over-crowding in the ED or inadequate staffing levels), the patient's treatment may be partially administered in the majors area, and therefore some of the escalation labels may be imprecise.

All data from both of the no-event patients groups (discharged and patients admitted to other wards) were used to build the ML models. This was done so that it would be possible to better separate the abnormal patterns presented by the no-event patients admitted to another ward, and event patients. Using the patient-wise performance analysis, the models trained using the ED data were compared to baseline EWS and PSI models, currently in use in the hospital setting. Our hypothesis was that fusing the continuous and intermittent vital-sign data, from the ED clinical context, would further improve their performance.

Multi-instance data based novelty detection (KDE and OSVM) and LR models, outperformed the baseline EWS or PSI models or any model trained on each data-source independently. There was always a combination of relevant continuous and observational clinical features, that when fused correctly, provided more accurate physiological information for the patients in our test dataset. The mixed data KDE model achieved the highest AUC_{E_3}, 0.737 (0.623, 0.830), using $\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, age, sex and GCS'_b. The incorporation of the age information was expected to improve the model performance, as visualised in the Neuroscale plots: the vital signs of young and elderly patients lie in separate clusters, i.e. the physiological deterioration was better identified by taking this information into account. As discussed in chapter 7, there were three event patients in our test data with low GCS, and without Visensia alerts before the escalation event. The inclusion of GCS information in the multi-instance ML models contributed to their higher AUC, as, at some threshold, it was able to help identify these patients.

We note that the multi-instance data OSVM model was also able to achieve a reasonable AUC_{E_3} of 0.711 (0.593, 0.807), by using the demographics, and temperature data, provided through the clinical observations. The MIGP model not only improved

model performance by reducing the effect of noisy periods of data, as seen in the previous chapter, but also by allowing the fusion of observational and continuous data.

The LR models were ranked below the novelty detection approaches. We note that, as expected, the best LR models make use of the FiO_2, as it is a predominant feature in the positive class (the event patients).

Finally, one limitation of our analysis, is that we did not correct for the class imbalance, and hence we did not discuss the optimal threshold for each model, as high thresholds would be preferred by our methodology, and moreover the difference in the AUCs$_{E_3}$ were not statistically significant. However, as seen in chapter 6, the AUC is still an indication of which model presents higher probability of classifying the different labels correctly. We leave the threshold optimisation problem for future work.

9.6 Conclusion

Our preliminary results, on an unbalanced dataset, provide an indication that (i) the correct modelling of the vital-sign time-series, (ii) and the re-training of the ML models with multi-instance data from the ED clinical context and with demographics information, should allow for a better detection of physiological deterioration in those patients requiring escalation to the resus area during their ED stay.

Chapter 10

Conclusion & Future Work

10.1 Thesis overview

This thesis has evaluated the use of an electronic T&T system and a data-fusion system, to help clinical staff identify physiological deterioration in patients attending the majors area of the JR ED. These technologies were used prospectively in phases 2 and 3 of the large-scale ED study described in chapter 3, respectively, which collected clinical data from a total of 10,488 patients.

The e-T&T system (VitalPAC) intervention was analysed in chapter 4. By moving from paper to electronic T&T charts, we observed that automating the calculation of the EWS and the suggestion of the corresponding time to the next observation set and clinical action, guided staff to conduct more observations on the more acutely-ill patients, i.e. those with $\text{EWS}_{max} \geq 1$ (about 9% more observation sets, than in phase 1, p-value < 0.05).

In chapter 5 we discussed the design and optimisation of EWS systems for the ED setting. The Centile-based EWS modelling approach was extended to allow for a non-quantised EWS, and the inclusion of age and oxygen support (FiO_2) data. Our analysis was focused on those patients with at least three observation sets (dataset E). An observation-wise performance analysis was used to evaluate the EWS systems' ability

to alert on abnormal vital-sign sets that preceded escalations to the resuscitation (resus) area of the ED. Those systems using the FiO_2 parameter and derived from an emergency clinical context, achieved the best results, namely, the DT-EWS (AUC = 0.720), and the Subbe(1) EWS (AUC = 0.698). However, the class imbalance in our dataset means that both a balanced data experiment and the use of efficiency curves were considered, to obtain the optimal EWS system configuration. Using the efficiency curves approach on the entire dataset, we concluded that a centile-based approach (such as in CEWS), re-computed from the ED training set, without quantisation in the interval [1-3] (and decreasing the score to zero throughout the normal range), conditioned on two age groups and taking FiO_2 into account, would provide the best PPV for the least amount of staff workload.

In chapter 7 we analysed the response of clinical staff to a novelty detection based data-fusion system, Visensia, able to alert automatically as a result of abnormal vital-sign data collected from bedside monitors. It was observed that 85% of the physiological alerts were followed by a clinical staff response. However, only 49% of the escalations to the resus area were preceded by a Visensia alert. One of the reasons for the failure to detect abnormal patterns in these cases, was the fact that the PSI score does not include other important physiological parameters, such as temperature (for which the bedside monitor sensor was deemed unreliable) and the level of consciousness, which was not taken into account in this model. Also, we were able to identify a high number of technical alerts, generated by data artefacts, which may have influenced the performance of this system, in the ED ward (as 15% of the physiological alerts did not have a response, and 36% of the total number of alerts were deemed to be technical alerts).

In line with recent literature in this area, reviewed in chapter 6, a 2-stage ML approach was proposed to improve the identification of patient physiological deterioration in the ED.

In chapter 8, a multi-instance Gaussian Process model (MIGP), that fuses the multiple sources (or instances) of a vital sign, was used to better estimate each patient's

vital-sign data stream. In this model most of the MIGP covariance hyperparameters are shared between the instances of the same vital sign, except for the noise term, which is instance-specific. The Bayesian optimisation procedure was used to find the optimal hyperparameters for each vital sign of each patient in the training set. Then, a multi-output linear regression transfer-learning function was used to estimate the hyperparameters for each patient's time-series in the test data E_3, based on the patient's age and sex.

This model was able to decrease the number of FP alerts, caused by artefacts in the continuous data, while allowing the fusion of complementary information from the observational and the continuous data into a synchronised multivariate time-series. Patient-wise performance analysis was used to show that the MIGP-based novelty detection could out-perform the baseline data-fusion approach, on the basis of the F1-score metric (used due to the data imbalance).

Finally, in chapter 9, we analysed ML based physiological risk models (one-class and two-class classification approaches), trained from the multi-instance data model (augmented with features available only in observational data, such as the use of oxygen support and the level of consciousness), and the patient demographics information (i.e. age and sex). A mixed data KDE approach, applied over the SMIGP time-series model, achieved the highest AUC on the test set, 0.737 (0.623, 0.830), amongst all the ML and EWS models applied to any of the data regimes (observational, continuous or multi-instance data).

10.2 Conclusion

A recent statement from The Royal College of Emergency Medicine (2016) concluded that, "common themes contributing to clinical incidents (in the ED) were, failure to recognise acute clinical deterioration, failure to escalate physiologically unstable patients to a sufficiently senior practitioner, and unsafe transfer of unstable patients owing to inadequate handover to ward-based teams", and recommended the use of NEWS to manage

the patient condition in the ED, specially for "boarded" patients. "Boarding" is used by some hospitals to avoid ED overcrowding (Boyle et al., 2015). In this process a small number of ED patients, with the lowest risk of deterioration, are immediately admitted to the hospital's acute units, while those at higher risk remain to be treated in the ED. Although this may improve patient flow (note that the consequences of ED crowding can go as far as having ambulances waiting to unload patients, and not attending to other emergencies), opponents claim that it just shifts the problem to other wards, and puts the patients' life at risk if they are not sent to the correct wards.

In our own study there was a change in the ED patient triage protocol at the end of phase 2, when nurse assessment was introduced before the start of phase 3, replacing the Manchester Triage System. Underlying any of the ED patient care protocols, that are required to cope with the ED's busy environment by meeting specific patient flow targets, is the need for the accurate assessment of patient physiology, and this is where our work can help. The question is: what is the best methodology to build models able to evaluate the risk of patient deterioration in the ED?

The Royal College of Physicians (2017) recommends the use of NEWS in the ED based on studies that correlate increased NEWS scores during the ED stay, with the need for hospital or ICU admission, in-hospital survival, 30- and 90-day mortality, and the identification of sepsis. Our study also provides these statistics (except for the identification of sepsis), but our approach was based on the development of EWS and ML models capable of distinguishing the vital-sign patterns of patients requiring escalation, from those of stable patients, during their ED stay; i.e. we optimised and evaluated our and other published models against no-escalation/escalation labels relevant to the ED setting. As discussed in Tarassenko et al. (2011), these labels are hard to define precisely (in contrast to mortality), depend on clinical judgement, and only take account of recognised deteriorations, making it difficult to set the EWS threshold for alerting on "early deterioration". In addition, although most of the data collected in our study could be used to assess the technology interventions, only 27% of the ED admissions (2,803 ad-

missions, from a total of 10,488) presented complete data to develop our models against ED-related escalation outcomes, according to our inclusion criteria.

Our work has revealed clear differences in the physiological trajectories of patients with and without escalations during their ED stay, in both observational and continuous data. Continuous data are therefore important in the ED setting, since staff may not be immediately available to observe ED patients, who are at risk of deteriorating rapidly. Our analysis showed that fusing both data-sources, together with the use of patient demographics information, provided the highest performance in identifying patient deterioration prior to escalations in the ED setting.

To conclude, we have presented a methodology that shows that e-T&T charts and continuous bedside monitors information can be integrated to provide a technological solution that more accurately assesses the risk of patient deterioration in the ED. More broadly, this approach could be extended to improve patient condition monitoring in other hospital wards, using wearable sensors to track patient condition.

10.3 Future Work

10.3.1 Extensions to the vital-sign time-series model

The potential limitations of our time-series modelling methodology include making the following assumptions: (i) there are no correlations between the different vital-sign streams; (ii) the GP prior dynamics is stationary, and does not change during the patients' time in the ED; and (iii) the multiple data instances of each vital sign have the same mean (i.e. the same bias). To tackle these limitations, our time-series modelling approach could be extended by using:

- Multi-task GP regression: as in Dürichen et al. (2015), the correlations between the different vital-sign channels could be learnt with different length-scales for each channel, as we observe that high-rate (e.g. HR), and intermittent data (e.g. BP)

have very different dynamics, but are physiologically correlated. Also the prior hyperparameters for single-instance data in our dataset, such as the temperature time-series, could be learnt by using information from the other vital-sign channels, by using a multi-instance and multi-task model. Furthermore, the temperature values (and BP values, in our case) estimated for the 30-minute forecasting window, would be inferred using observations and prior information from the remainder of the multi-instance vital-sign channels.

- Heteroscedastic regression: the multi-instance noise hyper-parameters from our model could be input-dependent (non-stationary). As an initial approach each $\theta_{WN}^{i,j,k}$ noise parameter could be modelled as an independent GP as in Goldberg et al. (1998); Kersting et al. (2007). These approaches use MCMC and Expectation-Maximisation, respectively, to find the noise process GP hyperparameters since its addition results in an intractable integral.

- Modelling the GP mean function: i.e. modelling the bias related to each data instance (for example, biases that staff may have in recording observation sets, or biases in bedside monitor sensors), and correct it, to allow a more accurate estimation of patient physiology.

With the previous extensions to the time-series modelling approach, we would have a greater number of patient-specific GP hyperparameters to estimate for test data. Non-linear transfer-learning functions, such as neural networks, could be used to improve the estimation of these variables. In addition, other relevant patient-specific inputs, such as the existence of a comorbidities (e.g. chronic obstructive pulmonary disease, or hypertension), could be used to improve the hyperparameter estimation, as these conditions are known to alter the expected values and the dynamics of the patients' vital signs.

10.3.2 Extensions to the physiological risk score models

A major assumption in our scoring models, is that the ED clinical data are i.i.d.. As we are dealing with time-series data, which are not i.i.d., sequence-based models such as a Hidden Markov Models (HMM) or Recurrent Neural Networks (RNN) should be considered.

Recently, versions of the Hidden Semi-Markov Models (HSMM) have been used to predict physiological deterioration states, combining intermittent and high-rate data from hospitalised patients (from various wards) who were later admitted to the ICU (Alaa and Van der Schaar, 2016; Alaa et al., 2017). Their HSMM version, i.e. the Hidden Absorbing Markov Model, achieved higher performance (measured using the AUC calculated from the TPR versus PPV curve, in their case) than other ML approaches including RNNs, and the baseline MEWS system, in identifying patients at risk of escalation to the ICU.

The use of this approach could allow: (i) physiological deterioration to be learnt as the latent dependency structure in sequence data, i.e. abnormal patterns are represented as latent variables Z_i such that, only when conditioned on Z_i, the relevant clinical data, $\mathbf{X}_i$ (such as the vital signs in our ED dataset), become conditionally independent; (ii) "the disease progression" to be learnt, i.e. the HSMM topology, leading to a specific ED outcome; (iii) use of the HSMM forward filtering algorithm to be used to infer the patients' clinical state, i.e. patient diagnosis and prognostic risk scoring (Alaa and Van der Schaar, 2016); and (iv) probabilistic criterion to be used to generate clinical alerts, i.e. specific HSMM states could represent the clinically relevant alerts that precede the escalation, learnt from the training data, rather than set heuristically, such as the persistence criterion used in our baseline data-fusion system.

The GP regression model could be used to generate the observations at specific times, to support any specific data sampling assumptions, such as those in the HSMM that requires a uniform sampling-rate (Lawrance and Rabiner, 1989). In this category of models, Expectation-Maximisation is usually used to learn the model parameters (Murphy, 2002), and the combination of the Bayesian Information criteria (as an exemplar

criterion amongst others) and domain knowledge are used to learn the appropriate model topology.

Other latent variable models that could provide similar advantages include the work of Saria et al. (2010), which combined a hierarchical Dirichlet process with autoregressive models to infer latent disease "topic" in the heart rate signals of premature babies; the work of Quinn et al. (2009), which used linear dynamical systems with latent switching variables to model physiological decompensating events like bradycardia, also in premature babies; and the work of (Stanculescu et al., 2014), which used autoregressive HMM to produce real-time predictions about the onset of neonatal sepsis, while handling missing data periods.

However, all these models achieved their performance by requiring some feature selection engineering. We conclude by making reference to the recent RNN architectures, which are able to learn the relevant features, from sequenced-based data, achieving comparable or better performance than previous models (i.e. those using human-engineered features). From the various examples of recent work on RNN architectures to classify clinical time-series data (Lipton et al., 2015; McCarthy and Williams, 2016), we note those using "multi-task learning", in which correlations between related tasks can improve prediction performance, especially with limited training data. Harutyunyan et al. (2017) used a multi-task Long Short-Term Memory (LSTM) RNN and a customised cost function, consisting of a weighted combination of the individual task losses, for the prediction of (i) in-hospital mortality, (ii) physiological decompensation[1], (iii) LOS, and (iv) acute care phenotype classification. They used time-series clinical data from 42,276 hospital admissions and ICU stays, from the MIMIC-III critical care database, 15% of which were used as the test set. Their model out-performed the baseline single-task Logistic Regression and LSTMs, on both the mortality and decompensation tasks (and presented similar results for the LOS prediction and phenotyping tasks). The AUROCs for the de-

[1]This outcome is related with that discussed in this thesis, although in their work it was defined as that occurring in periods in the 24 hours before death, similar to that done in the EWS systems literature.

compensation task were 0.9119, 0.8704, and 0.8946, respectively. This work was extended in Song et al. (2017), in which they used a multi-task attention-model RNN, that was able to out-perform the previous model and all the baseline models, for every task.

In summary, future work should investigate ML models that can learn from the sequence-based structure of the time-series data, and jointly train and predict the multiple ED patient outcomes analysed in this thesis, namely, patient discharge, admission to a hospital ward and escalation to the resus area, in addition to those ED-related events removed from our analysis criteria, namely unscheduled admission to ICU, and 48-hour in-hospital mortality and cardio-pulmonary resuscitation. Such models can provide useful information to help, not only in patient condition monitoring, but also in the fundamental clinical service offered by the ED, patient triage.

Appendix A

Large-scale study supporting material

A.1 Demographics of the large-scale ED study

Table A.1 shows the demographics information for the ED attendances in each phase of the large-scale ED study. As defined in chapter 3, the total number of patients included in the study is represented by the groups A_1, A_2 and A_3, respectively.

Table A.1: *Large-scale ED study demographics.* [a] *Wilcoxon test;* [b] χ^2*-test; *p-value* $\leq$ *0.05.*

Parameter	A_1	A_2	A_3	p-value A_1 vs A_2	p-value A_2 vs A_3
Age (years)[a]	54 (33-76)	55 (35-77)	60 (37-79)	0.006*	0.012*
Sex (% male)[b]	49.2	47.1	46.1	0.403	0.095
ED Admission (%)[b]					
Dawn (0 - 5am)	18.6	18.7	19.4	0.804	0.695
Morning (6 - 11am)	22.7	22.0	22.5		
Afternoon (12 - 17pm)	29.1	30.4	29.8		
Night (18 - 11pm)	29.5	29.0	28.4		
Triage (%)[b]					
Blue	0.2	0.1	0.1	0.001*	0.001*
Green	21.6	14.7	2.7		
Yellow	45.3	54.7	11.7		
Orange	9.9	16.3	2.4		
Red	0.3	0.1	0.0		
Not available	22.7	14.0	83.0		
Presenting Complaint (%)[b]					
Unwell Adult	13.1	12.9	2.8	0.303	0.039*
Chest Pain	12.2	12.7	2.9		
Abdominal Pain	11.7	14.1	3.0		
Collapsed Adult	7.3	9.1	1.7		
Shortness of Breath	4.5	4.7	1.1		
Overdose and Poisoning	3.9	3.9	0.5		
Falls	3.6	5.4	1.3		
Other	20.8	23.8	5.6		
Not available	22.9	13.5	81.1		
Referral (%)[b]					
111/NHS direct	0.0	6.2	5.2	0.012*	0.001*
Educational establish.	0.8	0.6	1.2		
Emergency services	65.9	69.6	71.6		
General practitioner	3.8	3.5	3.7		
Health care provider	3.1	1.6	1.6		
Local auth. social serv.	0.0	0.0	0.0		
Police	0.3	0.5	0.1		
Self referral	23.3	16.9	15.4		
Other	2.6	0.7	0.6		
Not available	0.2	0.3	0.6		
Arrival Mode (%)[b]					
Air ambulance	0.5	0.4	0.7	0.160	0.001*
Ambulance	67.5	74.8	76.3		
Police Vehicle	0.5	0.8	0.4		
Private Transport	26.6	21.4	19.7		
Public Transport	2.5	1.2	1.2		
Walk	1.5	0.6	0.6		
Other	0.7	0.4	0.5		
Not available	0.2	0.3	0.6		
Total Attendances	3,219	3,352	3,445	-	-

Appendix B

Performance of Machine Learning models

The results for the Kernel Density Estimate (KDE), One-class Support Vector Machines (OSVM), and Logistic Regression (LR) models, with highest AUC_{E_3} in the patient-wise performance analysis, are shown below (39 models are shown, from the 64 experiments conducted for the observational and the multi-instance data cases; and eight models are shown for the continuous data case). In each section, a different data-source is used, i.e. the models are evaluated when using only observational, continuous or the fused multi-instance data, are available to assess the patient physiology. Training data comprised the 1,733 patients from phases 1 and 2, 63 of which had an escalation event (group $E_{\{1,2\}}$). Test data comprised the 1,070 patients from phase 3, 33 of which had an escalation event (group E_3). The results are ranked by AUC_{E_3} (i.e. the AUC for the test dataset E_3).

B.1 Observational data

Table B.1: *Patient-wise performance of the KDE models applied to observational data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
23	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b	0.682	0.694
44	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b	0.669	0.699
43	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, FiO_2	0.663	0.710
60	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, SEX	0.663	0.726
62	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, FiO_2, SEX	0.663	0.720
33	HR, RR, SpO_2, SBP, TEMP	GCS_b, FiO_2	0.661	0.728
25	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	-	0.660	0.746
59	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, FiO_2, SEX	0.657	0.720
50	HR, RR, SpO_2, SBP, DBP	GCS_b, FiO_2, SEX	0.656	0.699
46	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	FiO_2	0.655	0.753
32	HR, RR, SpO_2, SBP, DBP, AGE	SEX	0.654	0.789
58	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, FiO_2	0.653	0.725
45	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, SEX	0.652	0.714
48	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	SEX	0.652	0.751
27	HR, RR, SpO_2, SBP, DBP	GCS_b, FiO_2	0.651	0.672
51	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, SEX	0.651	0.747
49	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, FiO_2	0.647	0.710
52	HR, RR, SpO_2, SBP, DBP, AGE	FiO_2, SEX	0.645	0.812
64	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, FiO_2, SEX	0.644	0.727
53	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b, FiO_2	0.643	0.712
35	HR, RR, SpO_2, SBP, TEMP	GCS_b, SEX	0.641	0.695
2	HR, RR, SpO_2, SBP, DBP	-	0.640	0.708
55	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b, SEX	0.640	0.719
61	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	FiO_2, SEX	0.640	0.763
12	HR, RR, SpO_2, SBP, DBP	SEX	0.639	0.715
28	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b	0.638	0.715
10	HR, RR, SpO_2, SBP, DBP	FiO_2	0.637	0.694
54	HR, RR, SpO_2, SBP, TEMP	GCS_b, FiO_2, SEX	0.637	0.727
63	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b, FiO_2, SEX	0.637	0.727
11	HR, RR, SpO_2, SBP, DBP, AGE	-	0.633	0.745
31	HR, RR, SpO_2, SBP, DBP	FiO_2, SEX	0.631	0.716
29	HR, RR, SpO_2, SBP, DBP	GCS_b, SEX	0.630	0.681
57	HR, RR, SpO_2, SBP, AGE	GCS_b, FiO_2, SEX	0.630	0.766
13	HR, RR, SpO_2, SBP, TEMP	GCS_b	0.629	0.689
24	HR, RR, SpO_2, SBP, DBP, TEMP	FiO_2	0.628	0.741
47	HR, RR, SpO_2, SBP, DBP, TEMP	FiO_2, SEX	0.628	0.738
18	HR, RR, SpO_2, SBP, AGE	GCS_b	0.627	0.704
30	HR, RR, SpO_2, SBP, DBP, AGE	FiO_2	0.626	0.751
34	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b	0.625	0.704

Table B.2: *Patient-wise performance of the OSVM models applied to observational data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
29	HR, RR, SpO_2, SBP, DBP	GCS_b, SEX	0.675	0.660
51	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, SEX	0.671	0.796
45	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, SEX	0.670	0.744
62	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, FiO_2, SEX	0.669	0.769
50	HR, RR, SpO_2, SBP, DBP	GCS_b, FiO_2, SEX	0.667	0.666
52	HR, RR, SpO_2, SBP, DBP, AGE	FiO_2, SEX	0.665	0.800
64	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, FiO_2, SEX	0.663	0.707
60	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, SEX	0.659	0.742
31	HR, RR, SpO_2, SBP, DBP	FiO_2, SEX	0.655	0.735
44	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b	0.654	0.669
59	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, FiO_2, SEX	0.652	0.713
58	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	GCS_b, FiO_2	0.651	0.742
47	HR, RR, SpO_2, SBP, DBP, TEMP	FiO_2, SEX	0.650	0.776
32	HR, RR, SpO_2, SBP, DBP, AGE	SEX	0.647	0.814
43	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b, FiO_2	0.647	0.705
61	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	FiO_2, SEX	0.647	0.735
23	HR, RR, SpO_2, SBP, DBP, TEMP	GCS_b	0.645	0.690
46	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	FiO_2	0.645	0.772
24	HR, RR, SpO_2, SBP, DBP, TEMP	FiO_2	0.644	0.757
12	HR, RR, SpO_2, SBP, DBP	SEX	0.643	0.778
26	HR, RR, SpO_2, SBP, DBP, TEMP	SEX	0.643	0.777
49	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b, FiO_2	0.639	0.774
25	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	-	0.636	0.699
48	HR, RR, SpO_2, SBP, DBP, TEMP, AGE	SEX	0.634	0.713
57	HR, RR, SpO_2, SBP, AGE	GCS_b, FiO_2, SEX	0.633	0.724
9	HR, RR, SpO_2, SBP, DBP	GCS_b	0.632	0.670
30	HR, RR, SpO_2, SBP, DBP, AGE	FiO_2	0.631	0.814
63	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b, FiO_2, SEX	0.630	0.700
27	HR, RR, SpO_2, SBP, DBP	GCS_b, FiO_2	0.629	0.685
11	HR, RR, SpO_2, SBP, DBP, AGE	-	0.627	0.850
18	HR, RR, SpO_2, SBP, AGE	GCS_b	0.626	0.654
41	HR, RR, SpO_2, SBP, AGE	GCS_b, SEX	0.626	0.655
2	HR, RR, SpO_2, SBP, DBP	-	0.623	0.702
54	HR, RR, SpO_2, SBP, TEMP	GCS_b, FiO_2, SEX	0.623	0.726
28	HR, RR, SpO_2, SBP, DBP, AGE	GCS_b	0.621	0.752
35	HR, RR, SpO_2, SBP, TEMP	GCS_b, SEX	0.621	0.707
53	HR, RR, SpO_2, SBP, TEMP, AGE	GCS_b, FiO_2	0.619	0.676
42	HR, RR, SpO_2, SBP, AGE	FiO_2, SEX	0.618	0.756
8	HR, RR, SpO_2, SBP, DBP, TEMP	-	0.617	0.702

Table B.3: *Patient-wise performance of the LR-L2$_{3,2}$ models applied to observational data. As indicated by the sub-indices, results are shown only for LR splines models with three equidistant knots and quadratic polynomials.*

CONTINUOUS	DISCRETE	AUC$_{E_3}$	AUC$_{E_{\{1,2\}}}$
HR, RR, SPO2, SBP, DBP, AGE	FiO$_2$, GCS$_b$	0.660	0.651
HR, RR, SPO2, SBP, DBP, TEMP, AGE	-	0.658	0.601
HR, RR, SPO2, SBP, TEMP, AGE	FiO$_2$, GCS$_b$	0.652	0.652
HR, RR, SPO2, SBP, AGE	FiO$_2$, GCS$_b$	0.652	0.661
HR, RR, SPO2, SBP, DBP	-	0.637	0.667
HR, RR, SPO2, SBP, TEMP	FiO$_2$	0.636	0.624
HR, RR, SPO2, SBP, DBP, AGE	-	0.634	0.601
HR, RR, SPO2, SBP, DBP	GCS$_b$	0.633	0.657
HR, RR, SPO2, SBP, DBP, TEMP, AGE	FiO$_2$	0.633	0.649
HR, RR, SPO2, SBP, DBP, AGE	GCS$_b$	0.628	0.603
HR, RR, SPO2, SBP, DBP, TEMP	GCS$_b$	0.627	0.611
HR, RR, SPO2, SBP, DBP, TEMP	-	0.627	0.651
HR, RR, SPO2, SBP, TEMP	-	0.626	0.668
HR, RR, SPO2, SBP, DBP	SEX, GCS$_b$	0.623	0.662
HR, RR, SPO2, SBP, TEMP	FiO$_2$, GCS$_b$	0.622	0.635
HR, RR, SPO2, SBP	FiO$_2$, GCS$_b$	0.616	0.663
HR, RR, SPO2, SBP, DBP	FiO$_2$, GCS$_b$	0.616	0.671
HR, RR, SPO2, SBP, DBP	SEX	0.615	0.669
HR, RR, SPO2, SBP, TEMP, AGE	-	0.614	0.624
HR, RR, SPO2, SBP, DBP, TEMP, AGE	GCS$_b$	0.613	0.599
HR, RR, SPO2, SBP, AGE	GCS$_b$	0.611	0.612
HR, RR, SPO2, SBP, DBP, TEMP	FiO$_2$, GCS$_b$	0.608	0.637
HR, RR, SPO2, SBP, DBP, TEMP	SEX	0.607	0.661
HR, RR, SPO2, SBP, TEMP	GCS$_b$	0.604	0.618
HR, RR, SPO2, SBP, DBP, AGE	SEX, FiO$_2$, GCS$_b$	0.603	0.673
HR, RR, SPO2, SBP	-	0.603	0.608
HR, RR, SPO2, SBP, AGE	-	0.603	0.641
HR, RR, SPO2, SBP, DBP	SEX, FiO$_2$	0.602	0.672
HR, RR, SPO2, SBP, DBP, TEMP	FiO$_2$	0.600	0.681
HR, RR, SPO2, SBP, TEMP, AGE	GCS$_b$	0.600	0.611
HR, RR, SPO2, SBP	SEX, FiO$_2$, GCS$_b$	0.599	0.662
HR, RR, SPO2, SBP, DBP	FiO$_2$	0.598	0.680
HR, RR, SPO2, SBP	GCS$_b$	0.598	0.596
HR, RR, SPO2, SBP, DBP	SEX, FiO$_2$, GCS$_b$	0.597	0.667
HR, RR, SPO2, SBP, DBP, TEMP	SEX, FiO$_2$, GCS$_b$	0.596	0.679
HR, RR, SPO2, SBP	FiO$_2$	0.595	0.664
HR, RR, SPO2, SBP, TEMP	SEX, FiO$_2$, GCS$_b$	0.595	0.670
HR, RR, SPO2, SBP	SEX, FiO$_2$	0.593	0.652
HR, RR, SPO2, SBP, TEMP	SEX, FiO$_2$	0.592	0.678

B.2 Continuous data

Table B.4: *Patient-wise performance of KDE models applied to the continuous time-series data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
5	HR′, RR′, SpO_2', SBP′, DBP′, AGE	-	0.699	0.608
7	HR′, RR′, SpO_2', SBP′, AGE	SEX	0.697	0.630
1	HR′, RR′, SpO_2', SBP′	-	0.692	0.637
4	HR′, RR′, SpO_2', SBP′	SEX	0.692	0.622
2	HR′, RR′, SpO_2', SBP′, DBP′	-	0.691	0.605
8	HR′, RR′, SpO_2', SBP′, DBP′, AGE	SEX	0.687	0.592
6	HR′, RR′, SpO_2', SBP′, DBP′	SEX	0.685	0.585
3	HR′, RR′, SpO_2', SBP′, AGE	-	0.653	0.625

Table B.5: *Patient-wise performance of OSVM models applied to continuous time-series data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
6	HR′, RR′, SpO_2', SBP′, DBP′	SEX	0.678	0.600
2	HR′, RR′, SpO_2', SBP′, DBP′	-	0.676	0.614
3	HR′, RR′, SpO_2', SBP′, AGE	-	0.667	0.613
1	HR′, RR′, SpO_2', SBP′	-	0.666	0.621
7	HR′, RR′, SpO_2', SBP′, AGE	SEX	0.655	0.617
4	HR′, RR′, SpO_2', SBP′	SEX	0.649	0.616
8	HR′, RR′, SpO_2', SBP′, DBP′, AGE	SEX	0.644	0.606
5	HR′, RR′, SpO_2', SBP′, DBP′, AGE	-	0.640	0.624

Table B.6: *Patient-wise performance of the LR-$L2_{3,3}$ models applied to continuous data. As indicated by the sub-indices, results are shown only for LR splines models with three equidistant knots and cubic polynomials.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
7	HR′, RR′, SPO2′, SBP′, AGE	SEX	0.696	0.657
4	HR′, RR′, SPO2′, SBP′	SEX	0.693	0.641
1	HR′, RR′, SPO2′, SBP′	-	0.693	0.627
3	HR′, RR′, SPO2′, SBP′, AGE	-	0.656	0.639
5	HR′, RR′, SPO2′, SBP′, DBP′, AGE	-	0.649	0.654
6	HR′, RR′, SPO2′, SBP′, DBP′	SEX	0.641	0.670
8	HR′, RR′, SPO2′, SBP′, DBP′, AGE	SEX	0.639	0.672
2	HR′, RR′, SPO2′, SBP′, DBP′	-	0.636	0.670

B.3 Multi-instance data

Table B.7: *Patient-wise performance of KDE models applied to multi-instance (MIGP) time-series data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$AUC_{E_{\{1,2\}}}$
51	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	GCS'_b, SEX	0.737	0.620
29	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	GCS'_b, SEX	0.728	0.590
23	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	GCS'_b	0.725	0.629
45	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	GCS'_b, SEX	0.720	0.648
50	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	GCS'_b, FiO'_2, SEX	0.720	0.613
9	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	GCS'_b	0.719	0.626
8	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	-	0.715	0.644
52	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	FiO'_2, SEX	0.715	0.489
49	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	GCS'_b, FiO'_2	0.713	0.666
12	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	SEX	0.710	0.643
10	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	FiO'_2	0.709	0.661
43	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	GCS'_b, FiO'_2	0.707	0.634
48	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	SEX	0.707	0.650
59	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	GCS'_b, FiO'_2, SEX	0.707	0.649
63	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{TEMP}$, AGE	GCS'_b, FiO'_2, SEX	0.706	0.682
27	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$	GCS'_b, FiO'_2	0.704	0.641
19	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$	GCS'_b, SEX	0.702	0.626
24	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	FiO'_2	0.702	0.659
47	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	FiO'_2, SEX	0.702	0.676
60	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	GCS'_b, SEX	0.702	0.646
58	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	GCS'_b, FiO'_2	0.701	0.659
62	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	GCS'_b, FiO'_2, SEX	0.701	0.632
41	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, AGE	GCS'_b, SEX	0.700	0.606
26	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$	SEX	0.699	0.642
61	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	FiO'_2, SEX	0.698	0.659
4	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$	GCS'_b	0.696	0.642
28	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	GCS'_b	0.696	0.636
40	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$	GCS'_b, FiO'_2, SEX	0.696	0.656
64	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	GCS'_b, FiO'_2, SEX	0.696	0.640
30	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, AGE	FiO'_2	0.691	0.675
34	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{TEMP}$, AGE	GCS'_b	0.691	0.658
5	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$	FiO'_2	0.689	0.656
44	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	GCS'_b	0.689	0.624
18	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, AGE	GCS'_b	0.688	0.631
17	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$	GCS'_b, FiO'_2	0.686	0.641
25	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{DBP}$, $\hat{TEMP}$, AGE	-	0.686	0.654
56	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{TEMP}$, AGE	FiO'_2, SEX	0.685	0.678
20	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, AGE	FiO'_2	0.684	0.662
33	$\hat{HR}$, $\hat{RR}$, $\hat{SpO}_2$, $\hat{SBP}$, $\hat{TEMP}$	GCS'_b, FiO'_2	0.684	0.665

Table B.8: *Patient-wise performance of OSVM models applied to multi-instance (MIGP) time-series data.*

ID	CONTINUOUS	DISCRETE	AUC_{E_3}	$\text{AUC}_{E_{\{1,2\}}}$
60	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b, SEX	0.711	0.647
51	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	GCS'_b, SEX	0.708	0.649
58	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b, FiO'_2	0.708	0.666
29	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	GCS'_b, SEX	0.705	0.629
64	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b, FiO'_2, SEX	0.704	0.663
28	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	GCS'_b	0.702	0.646
49	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	GCS'_b, FiO'_2	0.701	0.652
9	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	GCS'_b	0.700	0.638
62	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	GCS'_b, FiO'_2, SEX	0.700	0.652
61	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	FiO'_2, SEX	0.699	0.654
2	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	-	0.698	0.621
52	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	FiO'_2, SEX	0.698	0.650
45	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	GCS'_b, SEX	0.696	0.644
59	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	GCS'_b, FiO'_2, SEX	0.695	0.656
26	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	SEX	0.690	0.655
31	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	FiO'_2, SEX	0.689	0.652
32	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	SEX	0.689	0.638
50	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	GCS'_b, FiO'_2, SEX	0.689	0.654
12	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	SEX	0.688	0.630
23	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	GCS'_b	0.686	0.651
48	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	SEX	0.685	0.655
34	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b	0.683	0.652
41	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	GCS'_b, SEX	0.683	0.643
44	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b	0.682	0.651
27	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$	GCS'_b, FiO'_2	0.681	0.654
47	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	FiO'_2, SEX	0.681	0.655
30	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	FiO'_2	0.680	0.662
22	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	SEX	0.678	0.646
33	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$	GCS'_b, FiO'_2	0.678	0.669
18	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	GCS'_b	0.677	0.649
55	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$, AGE	GCS'_b, SEX	0.677	0.652
19	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$	GCS'_b, SEX	0.676	0.635
39	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	GCS'_b, FiO'_2	0.676	0.665
11	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, AGE	-	0.675	0.647
43	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$	GCS'_b, FiO'_2	0.674	0.668
57	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	GCS'_b, FiO'_2, SEX	0.673	0.654
6	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, AGE	-	0.670	0.648
46	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{DBP}}$, $\hat{\text{TEMP}}$, AGE	FiO'_2	0.670	0.677
38	$\hat{\text{HR}}$, $\hat{\text{RR}}$, $\hat{\text{SpO}}_2$, $\hat{\text{SBP}}$, $\hat{\text{TEMP}}$, AGE	SEX	0.669	0.656

Table B.9: *Patient-wise performance of the LR-L2$_{3,3}$ models applied to multi-instance data. As indicated by the sub-indices, results are shown only for LR splines models with three equidistant knots and cubic polynomials.*

CONTINUOUS	CATEGORICAL	AUC$_{E_3}$	AUC$_{E_{\{1,2\}}}$
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	FiO$_2'$, SEX	0.698	0.673
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	GCS$_b'$, SEX	0.697	0.659
ĤR, R̂R, SP̂O2, SB̂P, AGE	GCS$_b'$, FiO$_2'$, SEX	0.697	0.667
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	GCS$_b'$, FiO$_2'$, SEX	0.694	0.678
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	FiO$_2'$, SEX	0.687	0.664
ĤR, R̂R, SP̂O2, SB̂P, AGE	FiO$_2'$,	0.686	0.662
ĤR, R̂R, SP̂O2, SB̂P, AGE	GCS$_b'$,	0.686	0.649
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	GCS$_b'$,	0.686	0.636
ĤR, R̂R, SP̂O2, SB̂P, AGE	GCS$_b'$, SEX	0.684	0.656
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	GCS$_b'$, FiO$_2'$,	0.681	0.679
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	-	0.679	0.635
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP	GCS$_b'$	0.678	0.654
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	SEX	0.677	0.651
ĤR, R̂R, SP̂O2, SB̂P, AGE	FiO$_2'$, SEX	0.676	0.668
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	FiO$_2'$	0.674	0.666
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP, AGE	GCS$_b'$, FiO$_2'$,	0.673	0.687
ĤR, R̂R, SP̂O2, SB̂P	FiO$_2'$, SEX	0.671	0.664
ĤR, R̂R, SP̂O2, SB̂P	GCS$_b'$, SEX	0.670	0.645
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	GCS$_b'$, FiO$_2'$	0.670	0.675
ĤR, R̂R, SP̂O2, SB̂P	FiO$_2'$	0.670	0.664
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	GCS$_b'$, FiO$_2'$, SEX	0.669	0.686
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP	GCS$_b'$, FiO$_2'$	0.669	0.671
ĤR, R̂R, SP̂O2, SB̂P, AGE	GCS$_b'$, FiO$_2'$,	0.669	0.672
ĤR, R̂R, SP̂O2, SB̂P, AGE	-	0.668	0.662
ĤR, R̂R, SP̂O2, SB̂P	GCS$_b'$, FiO$_2'$, SEX	0.668	0.675
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	SEX	0.667	0.675
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP	FiO$_2'$	0.666	0.681
ĤR, R̂R, SP̂O2, SB̂P, TÊMP	FiO$_2'$	0.663	0.669
ĤR, R̂R, SP̂O2, SB̂P, TÊMP, AGE	GCS$_b'$, FiO$_2'$, SEX	0.661	0.661
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	GCS$_b'$, SEX	0.659	0.677
ĤR, R̂R, SP̂O2, SB̂P, AGE	SEX	0.658	0.663
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, AGE	FiO$_2'$,	0.658	0.665
ĤR, R̂R, SP̂O2, SB̂P, TÊMP, AGE	FiO$_2'$, SEX	0.657	0.657
ĤR, R̂R, SP̂O2, SB̂P, TÊMP, AGE	GCS$_b'$, FiO$_2'$,	0.656	0.669
ĤR, R̂R, SP̂O2, SB̂P, TÊMP, AGE	GCS$_b'$,	0.656	0.651
ĤR, R̂R, SP̂O2, SB̂P	GCS$_b'$, FiO$_2'$	0.655	0.678
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP	FiO$_2'$, SEX	0.654	0.658
ĤR, R̂R, SP̂O2, SB̂P, DB̂P	GCS$_b'$	0.654	0.663
ĤR, R̂R, SP̂O2, SB̂P, DB̂P, TÊMP, AGE	FiO$_2'$, SEX	0.653	0.658

Bibliography

Agyemang, M., Barker, K., and Alhajj, R. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538, 2006.

Aitchison, J. and Aitken, C. G. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976.

Alaa, A. M. and Van der Schaar, M. A hidden absorbing semi-markov model for informatively censored temporal data: Learning and inference. *arXiv:1612.06007*, 2016.

Alaa, A. M., Yoon, J., Hu, S., and Van der Schaar, M. Personalized risk scoring for critical care patients using mixtures of gaussian process experts. *arXiv:1605.00959*, 2016.

Alaa, A. M., Hu, S., and Van der Schaar, M. Learning from clinical judgments: Semi-markov-modulated marked hawkes processes for risk prognosis. *arXiv:1705.05267*, 2017.

Alcock, K., Clancy, M., and Crouch, R. Physiological observations of patients admitted from A&E. *Nursing Standard*, 16(34):33–37, May 2002.

Allen, K. Recognising and managing adult patients who are critically sick. *Medicine*, 22: 244–247, 2004.

Altman, D. G. and Bland, J. M. Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317, 1983.

Amoateng-Adjepong, Y., Del Mundo, J., and Manthous, C. A. Accuracy of an infrared tympanic thermometer. *Chest Journal*, 115(4):1002–1005, 1999.

Andrews, T. and Waterman, H. Packaging: a grounded theory of how to report physio-

logical deterioration effectively. *Journal of advanced nursing*, 52(5):473–481, 2005.

Armagan, E., Yilmaz, Y., Olmez, O. F., Simsek, G., and Gul, C. B. Predictive value of the modified early warning score in a turkish emergency department. *European Journal of Emergency Medicine*, 15(6):338–340, 2008.

Armstrong, B., Walthall, H., Clancy, M., Mullee, M., and Simpson, H. Recording of vital signs in a district general hospital emergency department. *Emergency Medicine Journal*, 25(12):799–802, 2008.

Badriyah, T., Briggs, J. S., Meredith, P., Jarvis, S. W., Schmidt, P. E., Featherstone, P. I., Prytherch, D. R., and Smith, G. B. Decision-tree early warning score (DTEWS) validates the design of the national early warning score (NEWS). *Resuscitation*, 85(3): 418–423, March 2014.

Bakar, Z. A., Mohemad, R., Ahmad, A., and Deris, M. M. A comparative study for outlier detection techniques in data mining. In *IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6. IEEE, 2006.

Becker, D., Miller, J., Ward, J., Greenberg, R., Young, H., and Sakalas, R. The outcome from severe head injury with early diagnosis and intensive management. *Journal of Neurosurgery*, 47(4):491–502, 1977.

Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., John McConnell, K., Pines, J. M., Rathlev, N., et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10, 2009.

Bhattachayya, A. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

Bishop, C. M. Novelty detection and neural network validation. *Vision, Image and Signal Processing, IEE Proceedings*, 141(4):217–222, August 1994. doi: 10.1049/ip-vis: 19941330.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, February 2007.

Bitterman, H. Bench-to-bedside review: oxygen as a drug. *Crit Care*, 13(1):205, 2009.

Bonnici, T., Gerry, S., Wong, D., Knight, J., and Watkinson, P. Evaluation of the effects of implementing an electronic early warning score system: protocol for a stepped wedge study. *BMC medical informatics and decision making*, 16(1):1, 2016.

Boyle, A., Viccellio, P., and Whale, C. Is "boarding" appropriate to help reduce crowding in emergency departments? *BMJ: British Medical Journal (Online)*, 350, 2015.

Buist, M. D., Moore, G. E., Bernard, S. A., Waxman, B. P., Anderson, J. N., and Nguyen, T. V. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *BMJ: British Medical Journal*, 324(7334):387–390, February 2002.

Burch, V. C., Tarr, G., and Morroni, C. Modified early warning score predicts the need for hospital admission and in-hospital mortality. *Emerg Med J*, 25(10):674–678, Oct 2008.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.

Charbonnier, S. and Gentil, S. A trend-based alarm system to improve patient monitoring in intensive care units. *Control Engineering Practice*, 15(9):1039–1050, September 2007.

Chatterjee, M., Moon, J., Murphy, R., and McCrea, D. 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.

Choukalas, C. G., Galvan, E. M., and Wallace, A. W. Aggregate Vital-Sign Monitoring Prior to Cardiac Arrest. *American Society of Anesthesiologists Annual Meeting*, A651, October 2011.

Choukalas, C. G., Galvan, T. S., and Stotts, J. Identifying ICU patients at high risk for cardiac arrest: a retrospective analysis of the visensia algorithm. *International Anesthesia Research Society 2015 Annual Meeting and International Science Symposium*, A651, March 2015.

Chu, C.-Y., Henderson, D. J., and Parmeter, C. F. Plug-in bandwidth selection for kernel density estimation with discrete data. *Econometrics*, 3(2):199–214, 2015.

Clifton, D. A., Wong, D., Fleming, S., Way, R., Wilson, S., Pullinger, R., and Tarassenko, L. Identifying patient deterioration in the emergency department using data fusion systems. In *British Medical Journal International Conference of Quality and Safety in Healthcare*, April 2011.

Clifton, D. A., Clifton, L., Sandu, D.-M., Smith, G., Tarassenko, L., Vollam, S. A., and Watkinson, P. J. 'Errors' and omissions in paper-based early warning scores: the association with changes in vital signs - a database analysis. *BMJ open*, 5(7):e007376, 2015.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. Gaussian process regression in vital-sign early warning systems. In *Engineering in Medicine and Biology Society*, pages 6161–6164. IEEE, 2012.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., Tarassenko, L., et al. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics*, 18(3): 722–730, 2014.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Colopy, G. W., Pimentel, M. A., Roberts, S. J., and Clifton, D. A. Bayesian gaussian processes for identifying the deteriorating patient. In *Engineering in Medicine and Biology Society*, pages 5311–5314. IEEE, 2016.

Colopy, G. W., Pimentel, M. A., Roberts, S. J., and Clifton, D. A. Bayesian optimisation of gaussian processes for identifying the deteriorating patient. In *Biomedical & Health Informatics*, pages 85–88. IEEE, 2017.

Considine, J., Lucas, E., and Wunderlich, B. The uptake of an early warning system in an australian emergency department: a pilot study. *Critical care and resuscitation*, 14 (2):135, 2012.

Cooper, N. Patient at risk! *Clinical Medicine*, 1(4):309–311, 2001.

Cretikos, M., Chen, J., Hillman, K., Bellomo, R., Finfer, S., and Flabouris, A. The objective medical emergency team activation criteria: A case–control study. *Resuscitation*, 73(1):62–72, April 2007.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

DeVita, M., Braithwaite, R., Mahidhara, R., Stuart, S., Foraida, M., and Simmons, R. Use of medical emergency team responses to reduce hospital cardiopulmonary arrests. *Quality and safety in health care*, 13(4):251–254, Aug 2004.

DeVita, M. A., Smith, G. B., Adam, S. K., Adams-Pizarro, I., Buist, M., Bellomo, R., Bonello, R., Cerchiari, E., Farlow, B., Goldsmith, D., et al. Identifying the hospitalised patient in crisis: A consensus conference on the afferent limb of rapid response systems. *Resuscitation*, 81(4):375–382, 2010.

Donaldson, L. J., Panesar, S. S., and Darzi, A. Patient-safety-related hospital deaths in England: thematic analysis of incidents reported to a national database, 2010-2012. *PLoS Med*, 11(6), 2014.

Duckitt, R. W., Buxton-Thomas, R., Walker, J., Cheek, E., Bewick, V., Venn, R., and Forni, L. G. 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.

Dürichen, R., Pimentel, M. A., Clifton, L., Schweikard, A., and Clifton, D. A. Multitask gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

Elliott, M. and Coventry, A. Critical care: the eight vital signs of patient monitoring. *Br J Nurs*, 21(10):621–625, 2012.

Etter, R., Ludwig, R., Lersch, F., Takala, J., and Merz, T. M. Early prognostic value of the medical emergency team calling criteria in patients admitted to intensive care

from the emergency department. *Critical care medicine*, 36(3):775–781, 2008.

Frey, C. B. and Osborne, M. A. The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.

Frigola-Alcade, R. *Bayesian Time Series Learning with Gaussian Processes*. PhD thesis, University of Cambridge, 2015.

Gao, H., McDonnell, A., Harrison, D. A., Moore, T., Adam, S., Daly, K., Esmonde, L., Goldhill, D. R., Parry, G. J., Rashidian, A., Subbe, C. P., and Harvey, S. Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine*, 33(4):667–679, 2007.

Gardner, J. R., Kusner, M. J., Xu, Z. E., Weinberger, K. Q., and Cunningham, J. P. Bayesian optimization with inequality constraints. In *ICML*, pages 937–945, 2014a.

Gardner, J. R., Kusner, M. J., Xu, Z. E., Weinberger, K. Q., and Cunningham, J. P. Bayesian optimization with inequality constraints. In *ICML*, pages 937–945, 2014b.

Gardner-Thorpe, J., Love, N., Wrightson, J., Walsh, S., and Keeling, N. The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. *Annals of the Royal College of Surgeons of England*, 88(6):571, 2006.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. Regression with input-dependent noise: A gaussian process treatment. In *Advances in neural information processing systems*, pages 493–499, 1998.

Goldhill, D. and McNarry, A. Physiological abnormalities in early warning scores are related to mortality in adult inpatients. *British Journal of Anaesthesia*, 92(6):882–884, 2004.

Goldhill, D., McNarry, A., Mandersloot, G., and McGinley, A. A physiologically-based early warning score for ward patients: the association between score and outcome. *Anaesthesia*, 60(6):547–553, 2005.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Graham, K. C. and Cvach, M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. *American Journal of Critical Care*, 19(1):

28–34, 2010.

Griffiths, J. R. Current use of early warning scores in UK emergency departments. *Emergency Medicine Journal*, 29(1):65–66, 2012.

Gupta, A. K. Respiration rate measurement based on impedance pneumography. *Texas Instruments application report SBAA181*, 2011.

Hancock, H. C. and Durham, L. Critical care outreach: the need for effective decision-making in clinical practice (part 2). *Intensive and Critical Care Nursing*, 23(2):104–114, 2007.

Hands, C., Reid, E., Meredith, P., Smith, G. B., Prytherch, D. R., Schmidt, P. E., and Featherstone, P. I. Patterns in the recording of vital signs and early warning scores: compliance with a clinical escalation protocol. *BMJ quality & safety*, 22(9):719–726, 2013.

Hann, A. *Multi-parameter Monitoring for Early Warning of Patient Deterioration*. PhD thesis, 2008.

Harutyunyan, H., Khachatrian, H., Kale, D. C., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *arXiv:1703.07771*, 2017.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning 2nd edition*. New York: Springer, 2009.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Health and Social Care Information Centre, Hospital Episode Statistics. Accident and Emergency Attendances in England (Experimental Statistics) 2010-11. 26 January 2012.

Heitz, C. R., Gaillard, J. P., Blumstein, H., Case, D., Messick, C., and Miller, C. D. Performance of the maximum modified early warning score to predict the need for higher care utilization among admitted emergency department patients. *Journal of Hospital Medicine*, 5(1):E46–E52, 2010.

Higginson, I. Emergency department crowding. *Emergency Medicine Journal*, 2012.

Hodge, V. J. and Austin, J. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

Hogan, H., Healey, F., Neale, G., Thomson, R., Vincent, C., and Black, N. Preventable deaths due to problems in care in english acute hospitals: a retrospective case record review study. *BMJ quality & safety*, 2012.

Hosking, J., Considine, J., and Sands, N. Recognising clinical deterioration in emergency department patients. *Australasian Emergency Nursing Journal*, 17(2):59 – 67, 2014.

Hravnak, M., DeVita, M. A., Clontz, A., Edwards, L., Valenta, C., and Pinsky, M. R. Cardiorespiratory instability before and after implementing an integrated monitoring system. *Critical care medicine*, 39(1):65, 2011.

Hravnak, M., Chen, L., Dubrawski, A., Bose, E., Clermont, G., and Pinsky, M. R. Real alerts and artifact classification in archived multi-signal vital sign monitoring data: implications for mining big data. *Journal of clinical monitoring and computing*, pages 1–14, 2015.

Hsiang, S. M. Visually-weighted regression. 2013.

Huang, D. T. Clinical review: Impact of emergency department care on intensive care unit costs. *Critical Care*, 8(6):498–502, 2004.

Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304, 1998.

Hugueny, S. *Novelty Detection with Extreme Value Theory in Vital-Sign Monitoring*. PhD thesis, University of Oxford, 2013.

Johnson, A. E., Burgess, J., Pimentel, M. A., Clifton, D. A., Young, J. D., Watkinson, P. J., Tarassenko, L., et al. Physiological trajectory of patients pre and post ICU discharge. In *Engineering in Medicine and Biology Society*, pages 3160–3163. IEEE, 2014.

Johnson, K. D. and Winkelman, C. The effect of emergency department crowding on patient outcomes: A literature review. *Advanced Emergency Nursing Journal*, 33(1), 2011.

Johnson, K. D., Winkelman, C., Burant, C. J., Dolansky, M., and Totten, V. The factors that affect the frequency of vital sign monitoring in the emergency department. *Journal of Emergency Nursing*, 40(1):27–35, 2012.

Kendall, M. G. Rank correlation methods. *Griffin, London*, 1948.

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, page 393–400, New York, NY, USA, 2007. ACM.

Khan, S. S. and Madden, M. G. A survey of recent trends in one class classification. In *Artificial Intelligence and Cognitive Science*, pages 188–197. Springer, 2009.

Lam, T. S., Mak, P. S. K., Siu, W. S., Lam, M. Y., Cheung, T. F., and Rainer, T. H. Validation of a modified early warning score (MEWS) in emergency department observation ward patients. *Hong Kong J Emerg Med*, 13(1):24–30, 2006.

Lawrance, R. and Rabiner, A. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Le Onn Ho, H. L., Shahidah, N., Koh, Z. X., Sultana, P., and Ong, M. E. H. Poor performance of the modified early warning score for predicting mortality in critically ill patients presenting to an emergency department. *World*, 4(4):273–278, 2013.

Lilienfeld-Toal, M., Midgley, K., Lieberbach, S., Barnard, L., Glasmacher, A., Gilleece, M., and Cook, G. Observation-based early warning scores to detect impending critical illness predict in-hospital and overall survival in patients undergoing allogeneic stem cell transplantation. *Biology of Blood and Marrow Transplantation*, 13(5):568–576, 2007.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv:1511.03677*, 2015.

Lowe, D. and Tipping, M. E. Neuroscale: Novel topographic feature extraction using RBF networks. *Adv Neur In*, 9:543–549, 1997.

Maas, A. and Franke, H. Women's health in menopause with a focus on hypertension. *Netherlands Heart Journal*, 17(2):68–72, 2009.

Marieb, E. N. and Hoehn, K. *Human anatomy & physiology*. Pearson Education, 7th edition, 2006.

Markou, M. and Singh, S. Novelty detection: a review–part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, December 2003a. doi: 16/j.sigpro.2003.07.018.

Markou, M. and Singh, S. Novelty detection: A review - part 2: Neural network based approaches. *Signal Processing*, 83, 2003b.

Marsland, S. Novelty detection in learning systems. *Neural computing surveys*, 3(2): 157–195, 2003.

Massey, F. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

McCarthy, A. and Williams, C. K. Predicting patient state-of-health using sliding window and recurrent classifiers. *arXiv:1612.00662*, 2016.

Miltner, R. S., Johnson, K. D., and Deierhoi, R. Exploring the frequency of blood pressure documentation in emergency departments. *Journal of Nursing Scholarship*, 46(2):98–105, 2014.

Mitchell, I. and Van Leuvan, C. Missed opportunities? An observational study of vital sign measurements. *Critical Care and Resuscitation*, 10(2):111, 2008.

Mitchell, T. M. Machine learning. *Computer Science Series. Singapure: McGraw-Hill Companies, Inc*, 1997.

Murphy, K. P. Hidden semi-markov models (HSMMs). *unpublished notes*, 2002.

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Nabney, I. *NETLAB: algorithms for pattern recognition*. Springer Science & Business Media, 2002.

Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

National Institute of Clinical Excellence. Acutely ill patients in hospital - recognition of and response to acute illness in adults in hospital, July 2007.

NHS. Quality and safety programme emergency departments: Case for change. February

2013.

Oberli, C., Urzua, J., Saez, C., Guarini, M., Cipriano, A., Garayar, B., Lema, G., Canessa, R., Sacco, C., and Irarrazaval, M. An expert system for monitor alarm integration. *Journal of Clinical Monitoring and Computing*, 15(1):29–35, 1999.

Olsson, T. Comparison of the rapid emergency medicine score and APACHE II in non-surgical emergency department patients. *Academic Emergency Medicine*, 10(10):1040–1048, October 2003.

Panday, R. N., Minderhoud, T., Alam, N., and Nanayakkara, P. Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): A narrative review. *European journal of internal medicine*, 45:20–31, 2017.

Paterson, R., MacLeod, D., Thetford, D., Beattie, A., Graham, C., Lam, S., and Bell, D. Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clinical Medicine*, 6(3):281–284, 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Phillips, J. and Barnsteiner, J. H. Clinical alarms: improving efficiency and effectiveness. *Critical care nursing quarterly*, 28(4):317–323, 2005.

Pimentel, M. A. *Modelling of Vital-Sign Data from Post-operative Patients*. PhD thesis, University of Oxford, 2015.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. A review of novelty detection. *Signal Processing*, 99:215–249, June 2014.

Prytherch, D. R., Smith, G. B., Schmidt, P., Featherstone, P. I., Stewart, K., Knight, D., and Higgins, B. Calculating early warning scores - a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation*, 70(2):173–178, August 2006.

Prytherch, D. R., Smith, G. B., Schmidt, P. E., and Featherstone, P. I. Views–towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*,

81(8):932–7, August 2010.

Pullinger, R., Wilson, S., Way, R., Santos, M., Wong, D., Clifton, D., Birks, J., and Tarassenko, L. Implementing an electronic observation and early warning score chart in the emergency department: A feasibility study. *European Journal of Emergency Medicine*, 2015.

Quinn, J. A., Williams, C. K., and McIntosh, N. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009.

Racine, J. S. Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, 3(1):1–88, 2008.

Ramsey III, M. Blood pressure monitoring: automated oscillometric devices. *Journal of clinical monitoring*, 7(1):56–67, 1991.

Rasmussen, C. E. and Nickisch, H. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11(Nov):3011–3015, 2010.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, November 2006.

Rees, J. and Mann, C. Use of the patient at risk scores in the emergency department: a preliminary study. *Emergency medicine journal*, 21(6):698–699, 2004.

Rhee, K. J., Fisher, C. J., and Willitis, N. H. The rapid acute physiology score. *The American journal of emergency medicine*, 5(4):278–282, 1987.

Riley, B. and Faleiro, R. Critical care outreach: rationale and development. *BJA CEPD Reviews*, 1(5):146–149, 2001.

Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401– 409, May 1969.

Santos, M. D., Clifton, D. A., and Tarassenko, L. Performance of early warning scoring systems to detect patient deterioration in the emergency department. In *International Symposium on Foundations of Health Informatics Engineering and Systems*, pages 159–169. Springer, 2013.

Saria, S., Koller, D., and Penn, A. Learning individual and population level traits from clinical temporal data. In *NIPS, Predictive Models in Personalized Medicine workshop*. Citeseer, 2010.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., Platt, J. C., et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588, 1999.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Sen, A., Rubinfeld, I., Azuh, O., Coba, V., Brady, P., Khouradji, I., Horst, M., and Patton, J. Visensia index predicts life-saving interventions in pre-hospital trauma patients. In *Critical Care Medicine*, volume 37, pages A71–A71, 2009.

Shelley, K. H. Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate. *Anesthesia & Analgesia*, 105(6):S31–S36, 2007.

Siebig, S., Kuhls, S., Imhoff, M., Gather, U., Schölmerich, J., and Wrede, C. E. Intensive care unit alarms - how many do we need? *Critical care medicine*, 38(2):451–456, 2010.

Slideplayer. Homeostasis. 2018. URL http://slideplayer.com/slide/8114550/.

Smith, G. B., Prytherch, D. R., Schmidt, P., Featherstone, P. I., Knight, D., Clements, G., and Mohammed, M. A. Hospital-wide physiological surveillance - a new approach to the early identification and management of the sick patient. *Resuscitation*, 71(1): 19–28, 2006.

Smith, G. B., Prytherch, D. R., Schmidt, P. E., and Featherstone, P. I. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170 – 179, 2008a.

Smith, G. B., Prytherch, D. R., Schmidt, P. E., Featherstone, P. I., and Higgins, B. A review, and performance evaluation, of single-parameter track and trigger systems. *Resuscitation*, 79(1):11–21, 2008b.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., and Featherstone, P. I.

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, April 2013.

Song, H., Rajan, D., Thiagarajan, J. J., and Spanias, A. Attend and diagnose: Clinical time series analysis using attention models. *arXiv:1711.03905*, 2017.

Stanculescu, I., Williams, C. K., and Freer, Y. Autoregressive hidden markov models for the early detection of neonatal sepsis. *IEEE journal of biomedical and health informatics*, 18(5):1560–1570, 2014.

Subbe, C. P. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):969–970, August 2011.

Subbe, C. P., Kruger, M., Rutherford, P., and Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM: An International Journal of Medicine*, 94 (10):521–526, 2001.

Subbe, C. P., Davies, R., Williams, E., Rutherford, P., and Gemmell, L. Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions. *Anaesthesia*, 58(8): 797–802, 2003.

Subbe, C. P., Slater, A., Menon, D., and Gemmell, L. Validation of physiological scoring systems in the accident and emergency department. *Emergency Medicine Journal*, 23 (11):841–845, 2006.

Subbe, C. P., Gao, H., and Harrison, D. A. Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward. *Intensive care medicine*, 33(4):619–624, 2007.

Taenzer, A. H., Pyke, J. B., and McGrath, S. P. A review of current and emerging approaches to address failure-to-rescue. *Anesthesiology*, 115(2):421–431, 2011.

Taenzer, A. H., Pyke, J., Herrick, M. D., Dodds, T. M., and McGrath, S. P. A comparison of oxygen saturation data in inpatients with low oxygen saturation using automated continuous monitoring and intermittent manual data charting. *Anesthesia & Analgesia*,

118(2):326–331, 2014.

Tarassenko, L., Hann, A., and Young, D. Integrated monitoring and analysis for early warning of patient deterioration. *British journal of anaesthesia*, 97(1):64–8, 2006.

Tarassenko, L. *Guide to neural computing applications*. Butterworth-Heinemann, 1998.

Tarassenko, L., Hann, A., Patterson, A., Braithwaite, E., Davidson, K., Barber, V., and Young, D. Biosign: Multi-parameter monitoring for early warning of patient deterioration. In *Medical Applications of Signal Processing, 2005. The 3rd IEE International Seminar*, pages 71–76, 2005.

Tarassenko, L., Clifton, D. A., Pinsky, M. R., Hravnak, M. T., Woods, J. R., and Watkinson, P. J. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–8, 2011.

Tax, D. M. and Duin, R. P. Support vector domain description. *Pattern recognition letters*, 20(11):1191–1199, 1999.

The Royal College of Emergency Medicine. *Position statement: National Early Warning Score (NEWS) for adult patients attending emergency departments*. 2016. URL www.rcem.ac.uk/docs/News/CEM10125-Position202016.pdf.

The Royal College of Physicians. *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*. 2017.

Tsien, C. L. and Fackler, J. C. Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine*, 25(4):614–619, 1997.

Turner, R. D. *Gaussian Processes for State Space Models and Change Point Detection*. PhD thesis, University of Cambridge, July 2011.

Veldhuis, R. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Processing Letters*, 9(3):96–99, March 2002.

Viera, A. J., Garrett, J. M., et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

Watkinson, P., Barber, V., Price, J., Hann, A., Tarassenko, L., and Young, J. D. 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, 2006.

Way, R. B., Beer, S. A., and Wilson, S. J. Whats that noise? Bedside monitoring in the Emergency Department. *International emergency nursing*, 22(4):197–201, 2014.

Williams, B., Alberti, G., Ball, C., Bell, D., Binks, R., Durham, L., et al. National early warning score (NEWS): Standardising the assessment of acute-illness severity in the NHS. *The Royal College of Physicians, London*, 2012.

Wilson, S. J., Wong, D., Clifton, D. A., Fleming, S., Way, R., Pullinger, R., and Tarassenko, L. Track and trigger in an emergency department: an observational evaluation study. *Emergency Medicine Journal*, 2012.

Wilson, S. J., Wong, D., Pullinger, R. M., Way, R., Clifton, D. A., and Tarassenko, L. Analysis of a data-fusion system for continuous vital sign monitoring in an emergency department. *European Journal of Emergency Medicine: Official Journal of the European Society for Emergency Medicine*, July 2014.

Winters, B. D., Weaver, S. J., Pfoh, E. R., Yang, T., Pham, J. C., and Dy, S. M. Rapid-response systems as a patient safety strategy: a systematic review. *Annals of internal medicine*, 158(5):417–425, 2013.

Wong, D. *Identifying Vital-Sign Abnormality in Acutely-Ill Patients*. PhD thesis, University of Oxford, 2011.

Wong, D., Bonnici, T., Knight, J., Morgan, L., Coombes, P., and Watkinson, P. SEND: a system for electronic notification and documentation of vital sign observations. *BMC medical informatics and decision making*, 15(1):68, 2015.

Wright, M., Stenhouse, C., and Morgan, R. Early detection of patients at risk (PART). *Anaesthesia*, 55(4):391–392, 2000.