Identifying Vital Sign Abnormality in Acutely-Ill Patients

David Wong

Brasenose College

Supervised by Professor Lionel Tarassenko

Submitted: Trinity Term, 2011

Identifying Vital Sign Abnormality in Acutely-Ill Patients

David Wong
Brasenose College
Doctor of Philosophy
Trinity Term 2011

The Emergency Department (ED) provides the first line of care for anyone seeking treatment for an urgent problem caused by an accident or illness. Physiological observations in the ED are a required part of patient care, and are used to monitor a patient's condition. Manual observations are recorded regularly by nursing staff, using a "Track and Trigger" (T&T) system, in which higher scores indicate greater physiological abnormality.

An observational study at the John Radcliffe Hospital, Oxford, was conducted to assess the effectiveness of T&T in the ED. Retrospective analysis showed that the effectiveness of T&T was limited by poor completion, and incorrect calculation of T&T scores. In response, we computed a retrospective, fully completed, scoring system which showed very clear improvements in both sensitivity and specificity.

In addition to nurse observations, higher acuity ED patients have their vital signs continuously monitored by bedside monitors. However, the alerts generated by the monitors are routinely ignored due to their high false alert rate. We investigated whether a baseline data fusion model and two alternative techniques, weighted Parzen windows and Support Vector Machines, could identify events relating to vital sign abnormality while keeping the number of false alerts to a minimum. The performance of each model was assessed by calculating its sensitivity and specificity. However, it was not possible to select an optimal model, due to the difficulty in assessing the relative importance of maximising true alerts and minimising false alerts.

In the final part of this thesis, two limitations of the data fusion models are highlighted. Firstly, missing data is not handled coherently within the current models, and secondly the models do not make use of temporal information. One method of addressing both of these issues, Gaussian processes, was considered. Using this method, a novel framework was derived that allowed for alerts to be generated even when there is uncertainty in the vital sign values.

Acknowledgements

This thesis would not have been started, let alone completed, without the help of many friends, family and colleagues. Firstly, I am especially grateful to my supervisor, Professor Lionel Tarassenko. I was once told to "learn everything" that I can from him, and it has taken a number of years to fully appreciate how studying under his supervision has not only resulted in a thesis, but it has also enriched my experience of medical engineering in a way that goes beyond the purely academic.

I am indebted to my clinical colleagues who have made this thesis possible. Dr. Rick Pullinger, Dr. Sarah Wilson, Mr. Rob Way - thank you for managing the ED research project, for your medical insight, and for the many hours spent trawling through reams of data and writing reports. Thank you also to the incredible team of research nurses, Ms. Sally Beer, Ms. Soubera Yousefi, and Ms. Karen Warnes, who were utterly instrumental in the collection of data in addition to their day jobs of saving lives.

Thank you to my friends and colleagues in SPANNER, and its biomedical offspring, BSP. Your chat, banter and cake have made the last few years a joy. Particular thanks go to Dr. David Clifton, for many a fruitful discussion, direction, and all manner of crazy Bayesian ideas. To Mr. Samuel Hugueny, for reading through this thesis, and introducing me to the Tour de France and the game of Go. To Dr. Susannah Fleming, for her inspiring work ethic, linux help, and assistance during the clinical project. To Miss Busi Vilakazi, for her solidarity in working late into the night and to Dr. Alistair Hann, who was kind enough to take me under his wing and provide guidance when I first joined the group. Thanks also to Dr Mark Larsen, Dr Christina Orphanidou, Dr Oliver Gibson, Dr Lei Clifton, Dr Alex Darrell, and Dr Tim Bonnici, who have made the lab a vibrant place (and have succeeded in making me feel under-qualified). I'd also like to thank Ms. Val Mitchell, who has always been helpful and friendly, and an unerring source of advice in all university and horticultural matters.

My friends and family have been a bedrock of support over the last few years in Oxford. Mum and Dad, thank you for your support over all of these years, particularly in my decision to study. Fi and Joel, and Sarah, you are wonderful. Thank you to my family in Oxford, Oxford Community Church. There are too many of you to name, but you have been brothers and sisters to me in both the easy and the tough times and I look forward to growing older with you. Thank you to my fiancée, Miss Claire Fitzgibbon, who has somehow dealt with the unpleasant, thesis-writing, version of me and has nevertheless found the grace to both love me and read the thesis.

Finally, to my saviour and friend, Jesus. Thank you for your friendship, for your relentless grace, and for inviting me into adventures with you.

Contents

1. Review of Vital Sign Monitoring Systems in the Hospital Emergency Department

1.1. Introduction

The Emergency Department (ED), which is often known as "Accident and Emergency" or "Casualty" in the UK, is a hospital department that provides the first line of care for anyone seeking treatment for an urgent problem caused by an accident or illness. Its primary goals are to diagnose the illness, provide initial treatment, and to escalate the patient to other hospital wards as necessary.

These goals are especially difficult to achieve for three main reasons. Firstly, ED patients have diverse reasons for presentation and the severity of a patient's condition is unlikely to be known prior to arrival at the ED. Unlike other hospital wards, where care can be focussed towards specific types of injury, the ED must have the capability to diagnose and treat many types of ailments. The wider range of decisions that are made in the ED means that there is a greater opportunity for mistakes to be made.

Secondly, the ED is extremely busy in comparison to other wards in the hospital, and often suffers from overcrowding [2]. The overcrowding is exacerbated when other wards are close to full capacity, as patients may be held in the ED until beds become available [51]. Furthermore, unlike with other wards, all of the patients who arrive at the ED are unscheduled, and may arrive at any hour of the day or night with little prior warning. In terms of patient throughput, a typical medium-sized hospital such as the John Radcliffe

Hospital, Oxford, may expect approximately 14,000 adult patients to be admitted to the the 28 acute care beds in the ED over a period of 6 months, in addition to those seen for minor injuries. The high workload may be the reason for a relatively high number of mistakes in patient care, as shown by Fordyce et al. [33] who identified 18 errors for every 100 patients in a busy ED.

The final reason why it is difficult to care for an ED patient is that there is substantial pressure to diagnose a patient as quickly as possible. From a medical viewpoint, a fast diagnosis is highly desirable as it is strongly correlated with improved patient outcome. For instance, it has been demonstrated that adverse outcomes in patients with conditions such as severe head trauma [7], cardiac arrests [16], or gangrene [72] can be reduced by earlier detection and intervention.

In addition, a prompt and efficient diagnosis reduces the patient's length of stay in the ward, thereby allowing a greater number of patients to be treated per day. Maximising a ward's efficiency has the added effect of reducing the financial cost of treatment per patient. A bed-day on a typical ward costs £225 for standard beds [32], and up to £1800 for an Intensive Care Unit-style bed equipped for acute patients [10], so an increased patient throughput may have a substantial financial benefit.

In light of these pressures, the UK Department of Health has imposed targets on acute hospitals in England, aiming for at least 98% of patients presenting to an ED to be seen, treated, and either admitted or discharged in less than four hours. While the benefits of making a fast diagnosis have been noted, it has been shown that time pressure often decreases an individual's ability to make accurate decisions [108] and in this context, may mean that there is an increased chance of making an incorrect diagnosis.

Despite the difficulties of providing a high standard of care in the ED, successful departments possess clinical expertise and systems that are well-suited to the ED's unique demands. This has led to a largely positive perception of the ED by patients [42], and a universal recognition of the importance of the ED in the patient's care pathway.

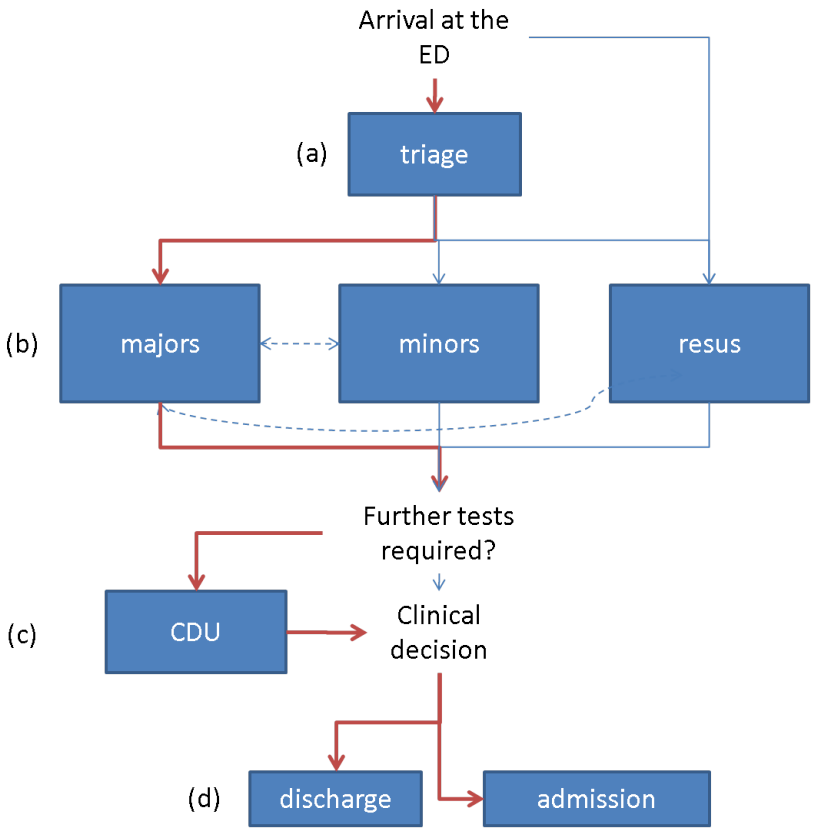

Figure 1.2.1.: Flow Chart showing the typical progression of a patient through the ED

1.2. Current Standards for Patient Care

We will now consider how an ED attempts to achieve its goals in practice, by highlighting the typical procedures that are currently in use. We will do this by first providing an overview of the care given to a typical patient using the flowchart in Figure 1.2.1, which depicts some of the most common stages of care for a patient in the ED. After this, we will examine one of the aspects of care, vital sign observations, in more detail using evidence from the literature.

When a patient first arrives in the ED, they will usually be triaged, so that patients are categorised by the severity of their condition (Figure 1.2.1(a)). This ensures that treatment is based on the order of clinical urgency, and that patients are sent to the correct treatment area within the ED. The triage involves combining information about the presenting problem as described by the patient, the patient's general appearance, and the recording of the values of the following physiological vital signs: heart rate (HR), respiratory rate (RR), peripheral oxygen saturation (SpO_2) levels, blood pressure (BP), core temperature, and some measure of consciousness. These are collectively known as vital

sign observations. The combined information is then presented in an easy-to-understand manner. For instance, one popular system, the Manchester Triage System, grades the patient on a five-level scale from red, for the most urgent conditions, to blue, which indicates the lowest severity. Triage is a very quick process, taking between two and five minutes. In some instances, it may not be appropriate to triage a patient. Most commonly, this will occur when the severity of a patient's condition has been assessed prior to arrival at the hospital by information provided by a paramedic team and immediate treatment is deemed necessary.

According to the result of the triage, the patient will be sent into one of three areas of the ED, which is physically divided in order to manage the patient population as effectively as possible (Figure 1.2.1(b)). The first of the ED areas is known as Minors, which admits patients who do not require immediate treatment and often includes patients with superficial injuries such as sprained ankles and wrists. The Majors area accommodates adult patients with a wide variety of illnesses and injuries that have a high likelihood of needing admission to hospital. The problems that are typically encountered are predominantly assigned a yellow triage category (the third highest acuity). The Resuscitation Room (Resus) is a clinical area in which patients with actual or potentially life-threatening illnesses or injury are assessed and treated. Resus patients attend the ED with the same types of problems as those in Majors, but are classified as more urgently in need of medical intervention. In addition to this, there may also be a Clinical Decision Unit (CDU) which is under the jurisdiction of the ED, and is used when a patient needs to be observed over a long period of time, or for patients who have taken an overdose of drugs or alcohol and do not need immediate assessment and treatment, but require regular observation.

Once assigned to an ED area, the patient receives a level of care specific to their needs. However, in all cases, vital sign observations will be made and recorded in order to monitor the patient's condition. The frequency of these observations is dependent on both the ED area and the condition of the patient, so Resuscitation Room patients may be observed at 5 minute intervals, while a patient in Majors may only be observed once per hour. The rate of observations may also increase if deterioration in the patient is identified. Patients in the Majors or Resus areas will also be continuously monitored using bedside monitors.

In addition to the vital sign observations, other physiological markers, such as pupil dilation and urine output, may also be recorded. How these observations can best be used to monitor a patient's condition is an open problem that we will consider in more detail in this thesis.

At some stage during the patient's stay, they will be attended to by a doctor who will attempt to diagnose the ailment. If this is not possible, the patient may be observed for longer and further tests may be conducted. During this stage, the patient may be moved to CDU, particularly if further observations are required beyond the 4-hour ED target (Figure 1.2.1(c)). If a diagnosis is made, the patient will be treated and discharged if appropriate, or else admitted to one of the other hospital wards for further treatment (Figure 1.2.1(d)). The decision to admit or discharge a patient may also be made for cases when a diagnosis has not yet been established, particularly in cases for which even the extended observations have been unable to provide enough information to form a detailed judgement.

1.3. Track-and-Trigger

The procedures whereby vital sign observations are used to assess patient status vary from hospital to hospital. However, the National Institute for Health and Clinical Excellence (NICE) has recommended implementing a standard method for analysing vital sign observations, known as the "Track-and-Trigger" system, for assessing all adult patients in acute hospital settings. This system is being increasingly adopted across the UK [31]. Track-and-Trigger is a methodology that was developed to facilitate clinical decision-making. The first step is the recording of a patient's vital signs at regular intervals ("tracking"). Nursing staff can then request the patient to be reviewed by a senior clinician if the patient's vital signs meet certain criteria or if they are concerned about a patient's condition ("triggering")[48]. The use of Track-and-Trigger is based on the premise that adverse events are often preceded by physiological derangement. For instance, cardiac arrests in hospital are often preceded by vital sign abnormalities up to 24 hours before the event [57, 94].

The introduction of the Track-and-Trigger system was designed to eliminate two prob-

lems with ward-based vital sign observations. Firstly, it had been noticed that poor communication between medical staff, and in particular, between nurses and doctors, could lead to mistakes in care. Hillman et al. discovered that, in many instances, nursing staff noted vital sign deterioration and would request further assistance by calling a junior doctor. However, in many instances, this was ineffective as the doctor did not possess the experience to deal appropriately with the situation [46]. Secondly, subtle deterioration in vital signs which includes long-term trends may not have previously been identified by nursing staff, an effect that may be exacerbated when information is miscommunicated during patient handover from one nursing shift to the next (see for example Patterson et al. [83]). Goldhill gives an example of such miscommunication, highlighting an instance when both doctors and nurses had been made aware of a patient who had abnormal physiology but had taken no action to prevent further deterioration for over 5 days [39]. The Track-and-Trigger system deals with these problems by providing clear guidelines for when assistance should be sought, thus empowering staff to call for help when vital signs are deteriorating.

In theory, Track-and-Trigger provides a way for nursing staff to identify changes in physiology and escalate the level of care if necessary, so that any unexpected deterioration can be dealt with at the earliest possible opportunity. The effect of this should be to help to reduce the number of preventable adverse events such as cardiac arrests or unplanned ICU admissions.

1.3.1. Trigger - Escalation of Care

The medical interventions that are triggered as a result of using a Track-and-Trigger system in the ED depend on the severity of the patient's deterioration. For moderate deterioration, the ED coordinator will be informed, and a doctor's review will be requested. In many cases, this will result in no immediate action, except that vital sign observations will then be made more frequently. If the patient continues to deteriorate, a senior consultant's review may be requested.

In more severe cases of deterioration, the patient may be immediately moved to the Resus area to be stabilised. In conjunction with this, the Critical Care Outreach (CCO)

or Intensive Care Unit (ICU) Outreach team may be called. The full intervention chart for the ED at the John Radcliffe Hospital from 2008 to 2011 is shown in Figure 1.3.1.

The CCO and ICU Outreach teams are specialised teams who are trained to deal with patients at immediate risk of cardiopulmonary arrest or other adverse events such as unplanned admission to the ICU or emergency surgical procedures, and aid ED staff in providing immediate care to the patients. They may also prepare patients to be transferred to other high-dependency areas of the hospital. These multidisciplinary teams typically include both a physician and a nurse. Outside of the UK, a very similar concept is often used, for example in the US, where the outreach team is known as the Medical Emergency Team (MET).

A number of studies have reported on the benefits of having such a team within the hospital. The benefits include marked reductions in mortality and morbidity associated with the seriously ill and those at risk from cardiac arrest. For instance, Bellomo et al. conducted a before-and-after trial that investigated the effect of introducing a MET at a tertiary referral hospital. Their results showed a 65% reduction in cardiac arrests and a 26% reduction in overall in-hospital mortality, while the survivors of cardiac arrest in the "after" period of the trial had a reduced hospital length-of-stay [9].

Offner et al. [79] also report that the MET equivalent in their hospital, a Rapid Response Team, was deployed successfully in a trauma center, resulting in early interventions that were believed to have been a factor in preventing patients from progressing to cardiac arrest, with a 50% reduction in the number of cardiac arrests that occurred.

1.3.2. Tracking - Single Parameter Early Warning Scores

Deviations from normal vital signs are tracked using Early Warning Score (EWS) criteria. In the earliest and most simple instance of an EWS, the criteria shown in Table 1.1 were adopted, and a response was triggered when any one of the criteria was met [48]. In this case, no numerical score is calculated; triggering only occurs when there are gross changes in a single vital sign parameter. While such a methodology has the advantage of being easy for a nursing team to follow, the criteria may be criticised as being over-simplistic, as they do not account for the fact that deterioration in patient condition may occasionally

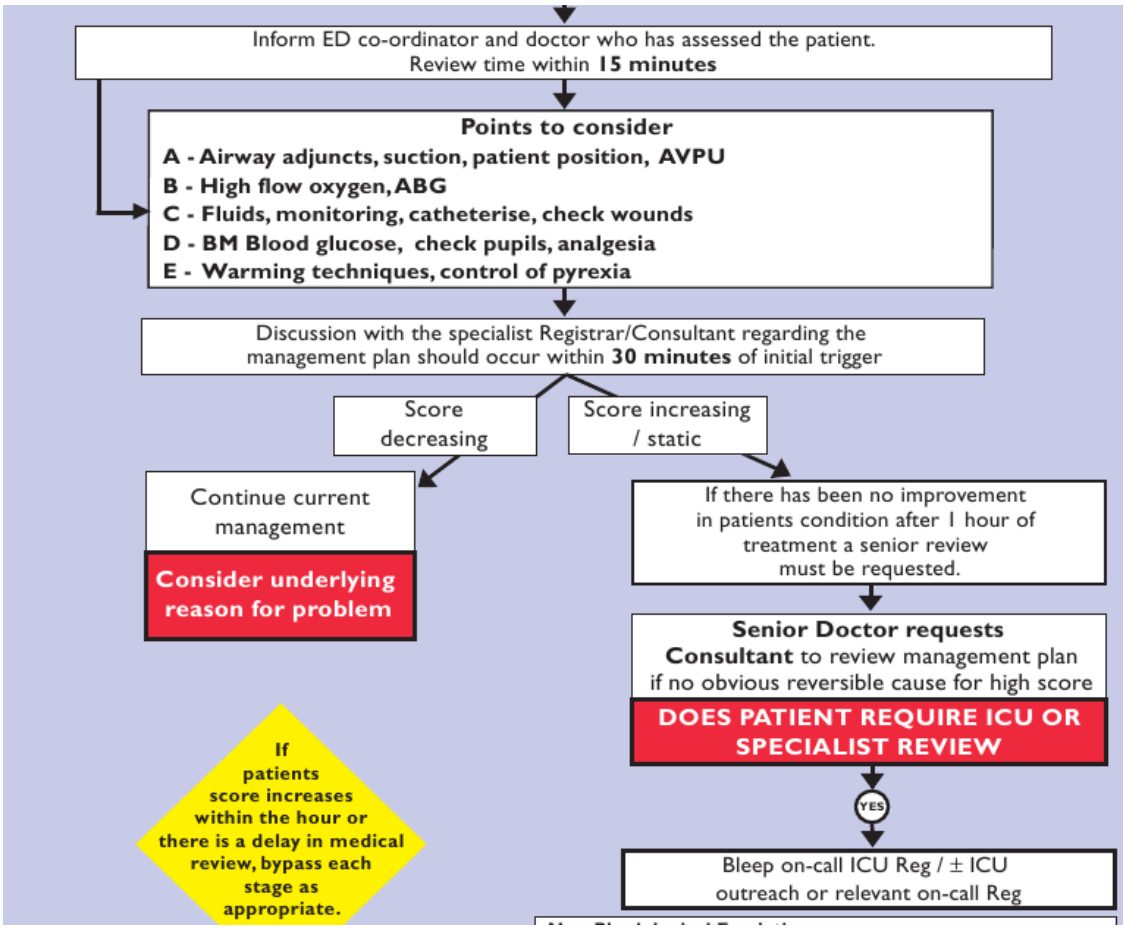

Figure 1.3.1.: Escalation flowchart chart for the ED at the John Radcliffe Hospital Oxford (2009). The chart describes actions that should be taken when the Track-and-Trigger criteria have been met.

ACUTE CHANGES IN	VITAL SIGNS
AIRWAY	
BREATHING	All Respiratory Arrests, Respiratory rate $<$ 5 breaths/min, Respiratory rate $>$ 36 breaths/min
CIRCULATION	All Cardiac arrests Pulse rate $<$ 40 beats/min Pulse rate $>$ 140 beats/min Systolic blood pressure $<$ 90mmHg
NEUROLOGY	Sudden fall in level of consciousness Repeated or prolonged seizures
Other	Any patient who does not fit the criteria above whom you are seriously worried about.

Table 1.1.: Early Warning criteria for the Track and Trigger system used at Liverpool Hospital, Sydney, Australia [48]

be heralded by moderate changes in several vital signs.

1.3.3. Multi-Parameter Early Warning Scores

In recent practice, it has become more common to use multi-parameter criteria, which attempt to assess a patient's condition based on the combination of all the vital sign measurements to produce an aggregated score. The simplest way to do this is to convert each vital sign into a sub-score according to a chart such as that in Table 1.2, which was used at the John Radcliffe Hospital, Oxford from 2008 to 2011. The EWS includes the Glasgow Coma Score as a measure of consciousness, in which a higher score indicates a greater degree of alertness. Vital sign observations that are outside the normal range for a typical adult are converted into higher sub-scores. The sum of the individual vital sign sub-scores then provides a total score, which is often known as the modified early warning score (MEWS). In this scheme, a response is triggered when the total score exceeds a given threshold.

For instance, the triggering threshold is 3 for a single vital sign, or 4 for multiple vital signs, using the criteria in Table 1.2. This means that a patient with a respiratory rate of 8 respirations/min or below would score a 3, causing a trigger regardless of the other vital signs. Similarly, a respiratory rate of 30 respirations/min or above would also cause

Score	3	2	1	0	1	2	3
Resp Rate	8 or below			9-18	19-24	25-29	30 or above
Heart Rate	40 or below		41-50	51-100	101-110	111-129	130 or above
O_2 Sats	91% or below on air						92% or below on 60% or more oxygen
Systolic BP	90 or below	91-99		100-179			180 or above
GCS	12 or below	13	14	15			
Urine Out-put	less than 10mls/hr	less than 20mls/hr	less than 0.5ml/kg/hr	more than 0.5ml/kg/hr			
Temp		35.0 or below					38.0 or above

Table 1.2.: The Track-and-Trigger scores for physiological variables as used in the ED at the John Radcliffe Hospital, Oxford between 2008 and 2011. A total score of 4 or more, or a subscore of 3 in any category, was defined as a 'critical' score that warranted a Trigger.

a trigger. A respiratory rate of 25-29 corresponds to a MEWS score of 2, which is not enough to cause a trigger, unless another vital sign also contributes a score of 2. For instance, this would occur if the heart rate was also between 111 and 129 beats/min.

Although such an aggregate scoring system was first described by Morgan [75], the concept of combining information from different parameters to evaluate a patient's condition was already prevalent in intensive care scoring systems, for example the APACHE III and SAPS II scores which are used to calculate the probability of in-ICU mortality for populations of adult ICU patients. The APACHE III score [59] divides the range of 20 physiological variables, including the six vital sign observations, into sub-ranges using "clinical judgement". Each sub-range is then treated as a separate variable for the purposes of logistic regression, which assigns a weight to each sub-range for each variable. The SAPS II score [65], uses similar methods in an attempt to predict mortality in surgical and medical patients.

The main advantage of the MEWS system is its ability to detect deterioration before any single vital sign may be considered abnormal. This advantage was demonstrated by

	Systolic BP	HR	RR
JR, Oxford - ED (2011) [120]	180	130	30
ViEWS (2010) [90]	250	131	25
Hourihan [48]	-	140	36
Offner [79]	-	131	25

Table 1.3.: A comparison of T&T criteria at different hospitals, showing the maximum values that Systolic Blood Pressure, Heart Rate, and Respiratory Rate are allowed to take before a single-parameter Trigger is generated.

Goldhill [39], who showed that a MEWS system would have highlighted deterioration in an example patient up to three days before an intervention took place. The patient was admitted to ICU and died a short time after.

One of the weaknesses of the MEWS systems, as highlighted by Cuthbertson [25], is that the thresholds used to convert the vital sign observations into sub-scores are solely based on expert opinion. Table 1.3 shows the upper threshold required to generate a critical score (score of 3) for Systolic BP, HR, and RR for various MEWS schemes.

1.3.4. Methods for Evaluating Early Warning Scores

Retrospective Analysis

Early Warning Scores can be evaluated retrospectively by relating a clinical outcome to the EWS score. This can be achieved by recognising that the goal of the EWS system is twofold: firstly, it must alert nursing staff to the need for clinical intervention, that is, it should correctly identify all events when patients are in a serious condition. Such events in this context are labelled as "true positives". Conversely, the EWS system must also correctly recognise when a patient's condition does not give rise to concern, and the number of "true negatives" should be maximised.

On the other hand, the number of "false positives", instances when the system detects patient deterioration, but the independent test shows this not to be the case, must be minimised. These false alerts will waste clinicians' time and can leave nursing staff desensitised to the occurrence of real deterioration. Similarly, instances when a patient is falsely classified as having no physiological problem could lead to delayed intervention, allowing more serious deterioration to occur, and should therefore also be minimised.

The classification information can be summarised in a 2x2 confusion matrix (Figure

	actual value = positive	actual value = negative
predicted value = positive	true positive	false positive
predicted value = negative	false negative	true negative

Table 1.4.: Confusion matrix for ROC analysis

1.4) [58]. We can determine the sensitivity, which is the ratio:

$$sensitivity = \frac{true\,positives}{true\,positives + false\,negatives} \tag{1.3.1}$$

and represents how good the system is at detecting deteriorating patients, irrespective of the number of false positives; and the specificity, which is the ratio:

$$specificity = \frac{true\,negatives}{true\,negatives + false\,positives} \tag{1.3.2}$$

For a system with high specificity, we can have high confidence that any event labelled positive will be a true event.

We can also determine other statistics such as the positive predictive value (PPV) which is the probability that a positive test result will be correct, and the negative predictive value (NPV), which is the probability that a negative test is correct.

The calculated sensitivity and specificity can only be understood with respect to the outcome markers for the test, that is, the criterion for a "true" event. In a clinical setting, collecting event markers may be difficult when the aim is to identify an outcome as broad and diffuse as "deterioration". For this reason, many authors may use a clearly defined measure such as mortality at 30 days after admission to the hospital. One drawback of this approach is that different choices of outcome markers may lead to vastly different sensitivities and specificities. Furthermore, the different outcome measures mean that it is often not possible to compare the sensitivities and specificities from different studies.

Receiver-Operator Characteristic (ROC) analysis [73] allows us to investigate the effect of changing the value of one or more of the parameter thresholds; for example, we may wish to investigate the effect of varying the threshold value for the calling criterion for heart rate on the frequency of escalation of care and the eventual patient outcome. This is known as a change in operating point in ROC analysis, and will cause a change in the sensitivity and specificity. The effect of changing the operating point can be shown on an

ROC curve, which plots sensitivity against (1-specificity) as the operating point is varied. The overall performance of a system is typically measured as the area under the ROC curve (AUROC), where an AUROC of 1 is optimal, and an AUROC of 0.5 is no better than a random guess. The optimal operating point, if false positives and false negatives are both valued equally, can be found by minimising the cost function:

$$C = (1 - sensitivity) + (specificity)$$

This is equivalent to finding the point on the ROC curve that is tangential to the line with a gradient of one closest to the point [0,1], as shown in Figure 1.3.2. If the relative importance of false positives and false negatives is known *a priori*, then the optimal operating point is found by minimising the more general cost function:

$$C = \nu(1 - sensitivity) + (1 - \nu)(specificity)$$

where ν is the cost of a false negative. Alternatively, an optimal operating point may be defined by setting a minimum allowable sensitivity or specificity. These are also shown in Figure 1.3.2.

Before-and-After Study

The effectiveness of the Early Warning Scores may be evaluated prospectively by documenting improvements in patient outcomes in a before-and-after study. Typical outcome markers include hospital length of stay, in-hospital mortality rate, and hospital readmission rates. These give a broad indication of whether the introduction of the EWS system has a positive effect on the management of the patient within the hospital. In the, "before", phase, standard care is administered, and the data collected were used as the baseline. During the second phase, the new EWS system is implemented and results are compared to the "before" data. The major disadvantage of this type of trial is that it is very difficult to ensure that no other factors contribute to the observed changes. For instance, mortality rate shows clear seasonal variations. In addition, before-and-after trials are prone to the Hawthorne effect, whereby the improvements can be attributed to changes in staff behaviour (such as increased alertness) as a result of their awareness that they are part

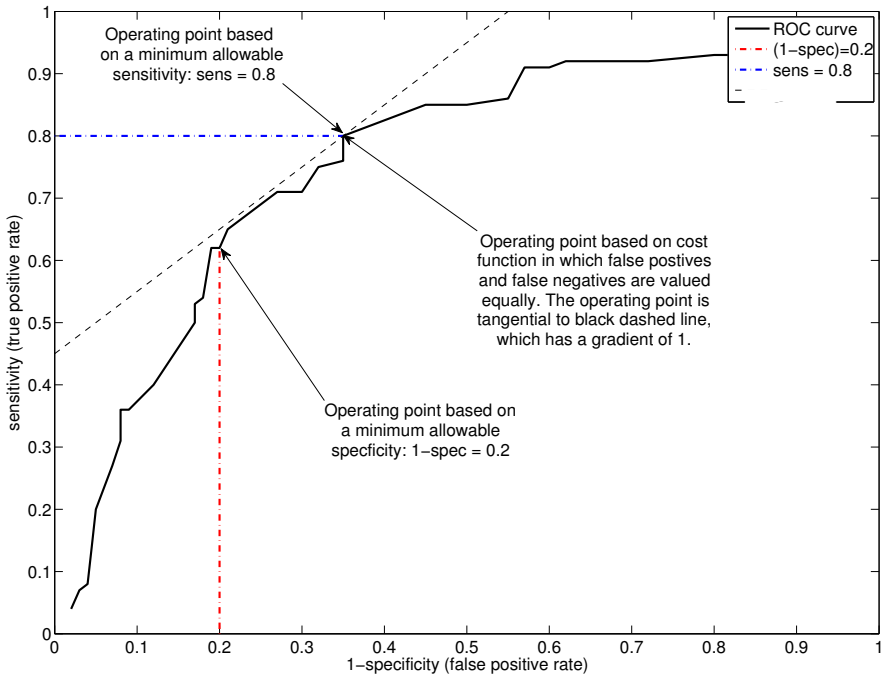

Figure 1.3.2.: An example of an ROC curve, showing how the operating point can be selected according to a minimum sensitivity, minimum specificity, or user-defined cost function.

of a trial.

Randomised Controlled Trials

The shortcomings of the before-and-after methodology are addressed by a Randomised Controlled Trial (RCT) design, which is a prospective study design that is commonly considered to be the most reliable form of scientific evidence. In a randomised controlled trial, patients are randomly allocated into one of two groups. In the "intervention" group the patients are managed using the EWS system, whilst the "control" group patients are given standard care. In some situations, such as drug testing, it may be appropriate to provide a placebo to the control group. Every effort is made to ensure that all other treatment of the two groups is as similar as possible.

1.3.5. Evidence-Based Early Warning Scores

To deal with Cuthbertson's criticism that EWS systems are generated heuristically, Prytherch et al. [90] have produced an evidence-based MEWS score called ViEWS. Their

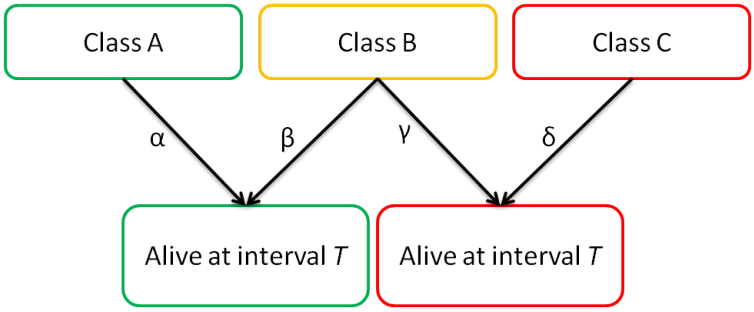

Figure 1.3.3.: Acutely ill patients can broadly be split into three classes: Class A - those who cannot be saved regardless of intervention, Class B - those who may be saved with appropriate intervention. Class C - those who require no intervention. We can consider both Class A and Class B as those containing the "abnormal" physiology, which we would like to detect. However, the division between Classes A-C is unknown at the time of the vital sign observation. In conventional supervised methods, some proxy, such as mortality at interval T is used to estimate the "abnormal" physiology. Such a system would detect all of those in class A, and the proportion β from Class B, but would misclassify the proportion γ, who had been "abnormal", but had later been redeemed by clinical intervention.

scoring system was assessed on vital sign observations collected from a hospital over a two-year period, and the best scoring system was chosen through "an iterative, pragmatic 'trial and error' process", in which each potential system was ranked using the AUROC based on patient mortality at 24 hours. The number of graduations for each vital sign was selected empirically, in keeping with current practice of assigning an integer score for each parameter between zero and three.

A hard outcome such as the 24-hour mortality was chosen due to the difficulty in defining "deterioration". However, there are also problems with this approach.

In practice, we may consider there to be three classes of acutely-ill patient on a ward: A.) those who have "abnormal physiology" and will die within 24 hours regardless of intervention, B.) those who have "abnormal physiology" but whose outcome is dependent on intervention, C.) those who have "normal physiology" and will be alive after 24 hours regardless of intervention (Figure 1.3.3). The critical objective of clinical care is to identify and provide clinical treatment for patients in class B so that their eventual outcome is improved. The ideal vital sign monitoring system will therefore detect patients from class B as well as class A. However, at the time of observation, the division of patients between the three classes is not known, so no ideal threshold can be determined easily.

The key problem of using the proxy of "alive at N days" to represent all patients with

"normal physiology" (and similarly, "dead at N days" as those with "abnormal physiology"), is that there is a significant time delay between the time of the vital sign observations and the time of the outcome. During this period, patients are cared for by the nursing staff, so that a patient who initially was deteriorating is "alive at N days" as a result of clinical intervention. The delayed outcome approach wrongly implies that clinical intervention during the 24 hours after vital sign observation makes no difference to the final outcome. The consequence of this is that all patients who had abnormal physiology at the time of the observations, but did not die, will be incorrectly classified.

We can demonstrate this more clearly using the criterion of 30-day mortality. Consider a patient that arrives at hospital with seriously abnormal physiology such that, without timely clinical intervention, we assume that the patient will continue to deteriorate which will lead to death. However, in this instance, the abnormal physiology is detected by clinical staff, and the patient leaves the hospital in a fit state and alive at 30 days from admission. In this highly plausible scenario, the abnormal physiology of the patient on arrival would be incorrectly associated with "normal physiology" as a result of the positive patient outcome (group labelled γ in Figure 1.3.3).

In response to this, Tarassenko et al. [109] have developed an alternative evidence-based MEWS scoring system using the vital sign data previously recorded by clinical trials of monitoring adult high-risk in-hospital patients. The distributions of each vital sign are first plotted, and Track and Trigger subscores are then assigned according to the 10th (90th), 5th (95th), or 1st (99th) centiles of the cumulative distribution functions of the corresponding vital signs. The system has since been introduced throughout the John Radcliffe hospital, Oxford.

1.3.6. Evaluation of Single Parameter EWS Systems

Retrospective Analysis

A number of studies use an ROC analysis approach to link the use of the EWS score system to patient outcome. Cretikos et al. [23] conducted a study in seven Australian public hospitals, which used 11 modified sets of EWS criteria, and unexpected deaths, unplanned ICU admissions and unexpected cardiac arrests as the outcomes. The different

EWS criteria had sensitivities between 49.1 and 71%, and specificities between 86.0 and 96.0%. Sensitivity decreased as the threshold for individual vital signs were set to more extreme values. Using this analysis, Cretikos et al. selected an "optimal EWS system", with a PPV of 15.7%, at a sensitivity of 53.6%, which indicates that an extremely high proportion, 84%, of patients who would trigger using this system would nevertheless not go on to have an adverse event.

Jacques et al. [53] assessed the association of 26 "early signs" and 21 "late signs" of severe adverse events (death, cardiac arrest, severe respiratory problem and transfer to critical care area) for 4617 patients. The "late signs" included 8 vital sign thresholds that comprise the single-parameter EWS system shown in Table 1.1, whereas the "early signs" included the same vital signs at less-acute values. The results showed that only 0.5% of patients had late signs alone, indicating that the EWS thresholds may not be sensitive enough for detecting early signs of deterioration [53]. Instead, they recommend using broader thresholds in their "early signs" so that extra attention is given to less severe derangements of physiology.

Before and After Studies

Bell describes a study at a Scandinavian teaching hospital in which the early-warning scores were calculated by a team of researchers and reported to the senior nurse if a trigger criterion was met [8]. In the study, three scores were used, a locally-standard score, an extended score that triggered at less extreme values, and a restricted score, which triggered at more extreme values. The study assessed these scores using patient mortality at 30 days, and it was discovered that the restricted score reduced the sensitivity significantly, thereby missing some of the deteriorating patients. This suggests that the locally-standard EWS thresholds were set at approximately the correct values.

Ball [6] reports on the effect of introducing an EWS and corresponding CCO team at the Royal Free Hampstead NHS trust, which led to a significant increase in patient survival rates and a decrease in hospital readmission rates from 7.4% to 4.8% for patients who had been admitted to the critical care floor. The author notes that some of the improvement in this before-and-after study may be attributed to the introduction of other

initiatives during the study period, including improving follow-up services after patients were discharged from hospital. An increase in survival rate was also demonstrated by Jones et al., who conducted a study that showed a 5% decrease in patient mortality at 1500 days, even after confounding factors had been taken into account [55].

Positive results are also reported by Rothschild et al. [93] and Buist et al. [16]. Rothschild et al. used a single-parameter based EWS system to detect deterioration in medical patients at an academic medical centre. In their study, the alerting threshold was exceeded in 60.8% of cases for patients who were later transferred to ICU, and triggers caused by high respiratory rates (tachypnoea) were strongly associated with future adverse events. Buist et al. reported a decrease in the incidence of unexpected cardiac arrest at a tertiary referral teaching hospital from 3.77 to 2.05 per 1000 admissions during a before-and-after study with two 12-month phases.

1.3.7. Evaluation of Multi-Parameter Early Warning Scores

Retrospective Analysis

In clinical practice there have been mixed results when Track-and-Trigger systems have been implemented, as reported in a 2007 Cochrane Review [71].

Gardner-Thorpe et al. [37] conducted an observational study to test the value of MEWS in identifying deterioration in surgical unit in-patients by observing how many patients were transferred to a critical care facility. The specificity and sensitivity at the optimum MEWS threshold were 83% and 75% respectively, and the researchers concluded that the MEWS system should be implemented for all surgical in-patients.

The EWS system proposed by Duckitt [27], known as the Worthing physiological scoring system (WPSS), used multivariate logistic regression to identify the strength of association between each vital sign and an increase in mortality rate, and then designed the system using the most important vital signs. In testing, the WPSS had a sensitivity and specificity of 0.63 and 0.72 respectively, compared to 0.60 and 0.67 for the original MEWS.

Before and After Studies

A number of studies reported a marked reduction in mortality and morbidity associated with the seriously ill and those at risk from cardiac arrest when using MEWS systems. For example, Garcea et al. [35] conducted a retrospective review of discharge data before and after the introduction of a Track and Trigger system with MEWS criteria. In the post-outreach period, in-hospital mortality and 30-day mortality were reduced for patients who had been readmitted to critical care, allowing them to conclude that the system had a positive overall impact. Priestley et al. have also shown, for 16 adult wards in a ward-randomised trial, that there was a statistically significant reduction of in-hospital mortality in wards for which a Track-and-Trigger was in use, when compared with those for which there was no such system in use [88]. Pittard assesses the use of MEWS in practice, and showed a reduction in unexpected ICU admissions, with a better outcome for the emergency patients [86].

Furthermore, Stenhouse et al. conclude that the introduction of a Track-and-Trigger system with MEWS scores in normal wards appears to lead to earlier detection of deterioration, as determined by the lower APACHE score of patients subsequently entering ICU [103].

Subbe [106] assesses the ability of MEWS to identify catastrophic deterioration during a patient's stay in an acute Medical Admissions Unit by calculating the maximum score during the stay, and showing that scores of 5 or more are associated with increased risk of death and Intensive Care Unit (ICU) admission. A second before-and-after study by Subbe [104] showed no discernible difference in outcome when using a MEWS system, and in particular that there were no changes in clinical outcomes for patients who had MEWS scores greater than four. Similarly, Leary et al. examined readmissions to critical care before and after the introduction of a Track-and-trigger system [66]. They did not detect any significant changes in the number of readmissions or the cause of readmissions between the two phases of the study.

Randomised Controlled Trials

The design of the previous trials has been questioned by Smith and Nolan [99], who noted that they were all before-and-after studies in which staff education may explain the observed improvement in care. In the largest example of a multi-centre study, a randomised control trial conducted by the Medical Early Response Intervention and Therapy (MERIT) investigators showed no improvement when tracking using MEWS was used in conjunction with the triggering of specialist Medical Emergency Teams in a standard ward setting [45].

ED studies

Very few Track-and-Trigger system evaluation studies have been conducted in the ED. It appears only one study has been conducted that analyses its use in monitoring and detecting deteriorating patients. In a study by Lam et al. [61], the researchers demonstrated that a MEWS scoring system was able to identify patients at risk of deterioration. They showed that high EWS scores were associated with increased risk of death, ICU admission and hospital admission on a 16-bed emergency department observation ward.

Burch et al. [17] also used the MEWS criteria in an ED setting, but instead used the scoring system as a proxy for the Manchester Triage score, showing that the proportion of patients admitted to hospital increased significantly as the MEWS score increased. They conclude that MEWS may be used as a rapid, simple triage method which agrees with similar work by Vorwerk et al. [116].

1.3.8. Overall Review

A recent evaluation of single-parameter EWS systems was conducted by Smith et al. to accurately compare the different types of EWS in the literature [101]. In their study, they used a data set of 10,051 vital sign observations recorded from May to December 2006 on adult patients at the time of patient admission to a hospital medical assessment unit (MAU). Using the data, they tried to determine how effective each EWS would be in predicting patient mortality at discharge by calculating the sensitivity and specificity of each EWS score. The results showed that all single-parameter EWS systems had low

sensitivities and positive predictive values, which indicates that the scores were unable to identify patients at risk of imminent in-hospital death using a single set of vital signs observations recorded on admission to a ward.

De Pennington et al. compared two multi-parameter EWS systems to a single parameter EWS system, using vital sign data from a Medical Admissions Unit, and assessed each scoring system's ability to classify patient death or admission to ICU [84]. They showed that both of the multi-parameter MEWS had better sensitivity than the single parameter EWS system, concluding that multi-parameter systems should be used to identify deteriorating patients. Subbe et al. [107] also conclude that multi-parameter EWS systems are more effective than single-parameter systems at detecting ED patients who are later transferred to ICU. Of the 100 patients admitted to the ICU, the single parameter criteria was found to have no additional value in comparison to the Manchester Triage Score preliminary assessment, while the Modified Early Warning Score identified an additional seven patients.

Many of the different MEWS systems were reviewed by Smith et al. [100], who evaluated 33 such systems on the Portsmouth data set, comparing their ability to discriminate between survivors and non-survivors of hospital admission. The most effective systems at predicting mortality at discharge, as determined by the AUROC curve, were found to be those that included the patient's age and temperature.

A systematic review of early warning scores conducted by Gao et al. [34] concluded that all the different MEWS criteria had "little evidence of reliability, validity and utility". The sensitivity of the systems was poor, and it was suggested that this might be due to the recording of the incorrect physiological parameters, or else due to a poor choice of trigger threshold.

The reason for such contrasting results may be explained by how well the Track-and-Trigger systems are implemented in practice. The MERIT study hinted at this by concluding that "monitoring, documentation and response to changes in vital sign were not adequate" and Buist [15] explicitly notes that retrospective inspections of the MERIT data highlighted many instances where the MEWS criteria had been fulfilled, but a response had not been triggered. This suggestion is given further credence by the results of

a study by Prytherch et al. [89], who suggest that the MEWS score is often being calculated inaccurately by ward staff. In a classroom exercise, MEWS scores were calculated for a data set of 2607 vital sign observations recorded at the Medical Assessment Unit of Queen Alexandra Hospital, Portsmouth. Of these, 2.5% contained scores that should have triggered an alert, but had been miscalculated.

Smith and Oakey [98] have also shown that in a clinical ward setting in a medium-sized general hospital, MEWS scores were incorrectly calculated 21.9% of the time. Of these, patients with abnormal vital signs were most likely to be mis-scored, and the scores were commonly underestimated, so a large number of patients who should have exceeded the alerting threshold did not. Edwards et al's [29] recent four ward prospective study concurs with this assessment, showing again that MEWS scores are often underestimated and that paper-based MEWS systems are unreliable for triggering timely medical reviews. The reason for the incorrect scoring may simply be due to insufficient training, as suggested by Lawson et al. [64], and various authors [16, 55, 97] have noted the importance of effective training prior to introducing a track and trigger system. Alternatively, Subbe and Gao [105] hint at the possibility that the complexity of the MEWS scoring chart itself causes errors, showing that inter- and intra-rater reliability was lower for multiple parameter scores than for single parameter warning scores.

Overall, the weight of evidence in the literature suggests that Track-and-Trigger systems that use MEWS criteria are a useful tool when implemented correctly. However, the vastly varying results reported in different studies suggest either that Track-and-Trigger systems have varying effectiveness depending on the type of patients being monitored, or that the systems are being applied inconsistently. This second reason is considered to be very likely, and in many of the studies where no improvement was found, the researchers hinted at possible problems with the practical implementation of Track-and-Trigger, rather than problems with the system *per se*.

1.4. Continuous Monitoring

Patients in the Majors or Resus sections of the ED commonly have their vital signs monitored continuously through the use of integrated bedside monitors. We now discuss

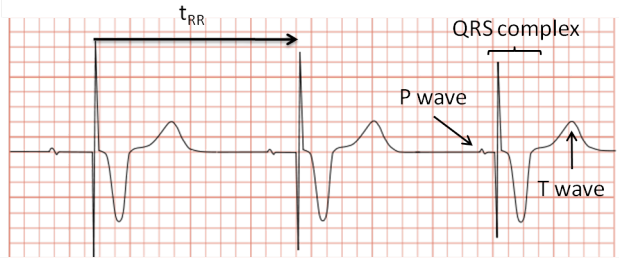

Figure 1.4.1.: An ideal ECG waveform. The diagram indicates the period between consecutive R-peaks in the signal, and the HR can be derived from the average of t_{RR} over several heart beat cycles.

how each of the vital sign parameters can be monitored electronically, and then how the information from each of the vital sign channels can be used within intelligent monitoring systems.

Heart Rate and Respiratory Rate

Heart rate is typically measured by using three or more Electrocardiogram (ECG) chest electrodes to record the ECG waveform from one or two leads. The heart rate can then be derived from the waveform by calculating the time between consecutive R-peaks (see Figure 1.4.1) [114]. The Respiratory rate is derived using impedance pneumography [82], which measures the electrical impedance between two of the ECG electrodes at a frequency between 10kHz and 100kHz. The impedance increases as the patient inhales due mainly to the increased resistivity of the air-filled lungs but also because of changes in the volume of the chest cavity.

Blood Pressure

Blood pressure varies with the contractions of the heart muscles. Blood pressure is recorded using two parameters, the Systolic (SBP) and Diastolic (DBP) blood pressure. The systolic blood pressure is the highest value, corresponding to the time at which the heart's ventricles contract, whereas the diastolic blood pressure is the lowest value, recorded during the lull between contractions. Manual observations of the blood pressure are made by inflating a blood pressure cuff around the upper arm in order to occlude blood flow. The cuff is gradually deflated, and the SBP and DBP can be derived by listening for the presence of so-called Korotkoff sounds [85].

Automatic measurement of the blood pressure is typically measured using oscillometry [85]. The blood pressure cuff is automatically inflated to occlude blood flow in the upper arm, and slowly deflated. This time, however, oscillations in the cuff pressure caused by oscillations in the blood flow are used to determine the Mean Arterial Pressure (MAP). The blood pressure is equal to the MAP when the cuff pressure oscillations are at a maximum. The SBP and DBP values are then derived heuristically from the MAP when the amplitudes of the oscillations reach a set fraction of the oscillation amplitude of the MAP.

Oxygen Saturation

Oxygen saturation is a measure of the oxyhaemoglobin (that is, oxygenated haemoglobin) in the bloodstream, and is reported as a percentage of the total haemoglobin. Typically, a range of 95% to 100% oxygen saturation is considered to be normal . Peripheral arterial oxygen saturation is measured using a pulse oximeter attached to the patient's finger. The oximeter contains two LEDs, usually at wavelengths of 660nm (red light), and 910nm (infrared light), and a photo-diode. The absorption of light at these wavelengths differs for oxygenated and deoxygenated blood due to the different absorption coefficients for oxyhaemoglobin and reduced haemoglobin, hence the oxygen saturation can be determined [60].

Temperature

In clinical practice, core body temperature is measured using a tympanic, rectal, or oral thermometer. It is not possible to obtain continuous measurements of core body temperature with these techniques. Instead, the studies described in this thesis initially used thermistors applied to the skin as a proxy for core temperature. In order to minimise the effect of the ambient conditions, the thermistors are secured with an adhesive under the blood pressure cuff. This precaution both shields the sensor and helps to maintain firm contact with the skin.

1.4.1. Continuous Monitoring Systems

Bedside monitors are designed to generate an audible alert when a patient is deemed to suffer from physiological deterioration. In the simplest case, single-channel alerts can be set for each vital sign, so that the bedside monitor will alert when any one of the vital signs is outside of a set range. These single-channel systems are particularly prone to producing false alerts, and Tsien and Fackler [115] showed that approximately 86% of alerts from a bedside monitor in an ICU setting were false alerts.

Intelligent monitoring can be achieved by combining information from multiple vital sign channels - this process is known as data fusion. The output from a data fusion algorithm can then be used to generate alerts that may provide a truer indication of the overall patient's condition, and may also be less prone to false alerts. Most of the data fusion techniques described in the literature have been designed for the ICU where continuous monitoring is standard, but the techniques may also be applied to acute wards outside the ICU.

Oberli et al. propose an expert systems approach to the data fusion problem [78]. An expert, or knowledge-based system is one that uses a direct encoding of human knowledge to help solve complex problems. In Oberli's system, the vital signs are first converted into a set of quantitative classes which describe a physiological condition, such as "bradycardia" or "normal heart rate", based on training information given by a set of clinicians. The classes overlap and are described using fuzzy logic, so it is possible to be "somewhat bradycardic". After this, a patient diagnosis can be arrived at by following a set of logical

rules derived from expert knowledge. For instance, if asystole AND extreme hypotension were present AND no pulse detected, the patient would be classified as having a cardiac arrest. In this way, the system provides both an alert for patient deterioration and a preliminary diagnosis. For 70.5 hours of continuous vital sign data, alerts triggered by the system were classified as true-positive, false-positive or false-negative by two observers, from which they calculated a false alert rate of less than 1%, and a sensitivity of 92%.

Schoenberg et al. use expert knowledge differently [96]. In their scheme, a customisable "logic engine" is designed, which is able to interpret information from multiple single-channel vital sign monitors. The purpose of the system is to filter out clinically insignificant results. The system works by analysing a set of user-defined features that can be extracted from the raw vital signs. For instance, this may include the change in average heart rate between the current minute and the three previous minutes. Thresholds are set for each feature, based on expert advice, and a feature is assigned a score if it exceeds the threshold value. The sum of the scores is then compared to a critical total, which triggers an alert if exceeded. The aggregate scoring system has many similarities to MEWS systems, and this technique can be considered as an automated generalised scoring system. During tests on 120 hours of ICU data, the logic engine had a positive predictive value of 32% compared to 3% for standard monitors.

Data fusion techniques can also be used to analyse trends in continuous data. Charbonnier and Gentil have attempted to incorporate historical data into an alert system by making use of trend analysis [20]. In their system, each parameter is converted into a semi-quantitative temporal feature. Typically, the features are {Increasing, Decreasing, Steady}, and the quantitative data are the start and end time of the event, as well as the start and end value of the vital sign parameter. In order to make best use of the data, the trend features are aggregated to form the longest possible episode. A set of rules is used to accomplish this task. The extended features can then be used in a rule-based system that triggers alerts when the trend is persistent and severe. The system can also be trained to recognise artefactual events, and tests using this scheme in an ICU resulted in a 33% reduction in false alerts, without missing any clinically relevant events.

Temporal information can also be used to detect artefactual readings and reduce the

number of false alerts. Hoare and Beatty analyse the time series for a set of physiological features, and attempt to predict the next values [47]. A new data point can then be classed as artefactual if its value is outside a predetermined range. Williams et al. extend this work, tackling the problem of infant monitoring immediately after birth by using a Factorial Switching Kalman Filter (FSKF) to model vital sign data in neonatal intensive care [118] . A Kalman filter is a recursive filter that calculates the optimum estimates of process variables in the presence of noise. The FSKF extends this by allowing the filter to use different linear dynamic models that are selected by a switch variable, allowing "normal" and "artefactual" conditions to be modelled.

The switch variable itself changes depending on a number of individual factors such as the presence of bradycardia, or recognition of a probe disconnection. Given a set of observations, the FSKF is then used to calculate the most likely switching state. By establishing which factors activate the switch, the specific cause of the artefact and an estimate of the true value of the observed data can be computed. In testing on eight 28-week old infants, a total of nine parameters were monitored. Results were presented showing examples in which the start and end times of specific physiological conditions were accurately detected.

Few attempts have been made to create personalised models that cater for differences in physiology caused by factors such as age, lifestyle or diet. Zhang proposed a personalised model that attempts to increase the alert specificity by automatically tuning alert thresholds on a per-patient basis [121]. In his study, both neural networks and classification trees were tested and used to generate personalised alert thresholds. From these, neural networks performed consistently better. Typically, the system required eight hours of data, and the performance improved as the length of training data increased. Overall, the algorithm had a sensitivity of 0.96 and a specificity of 0.99. Although the performance of the model appears to be excellent, we note that the metrics are particularly sensitive to the way in which positive and negative events were defined.

1.5. Thesis Overview

In response to the NICE guidance [31], the Oxford Radcliffe Hospitals Trust carried out a review of the available scoring systems and after development work formally adopted in 2008 a hospital-wide Track-and-Trigger (T&T) system. The system makes use of a MEWS score that will be referred to as a T&T score. This was introduced into the John Radcliffe (JR) Hospital ED for all adult patients with the aim of providing a standardised system to detect patient deterioration and to provide continuity from the ED through to the wards.

The effectiveness of the T&T system at the John Radcliffe Hospital ED has not previously been tested. More generally, very little information has been gathered about the effectiveness of T&T in the ED context. In response, a single centre observational cohort study was conducted in the emergency department. The study was conducted with the approval of the UK National Research Ethics Service, reference number 08/H1307/56. During the study, the vital signs and T&T scores were collected, as recorded on nurses' observation charts. In addition to this, continuous vital sign data were acquired from bedside monitors. In order to achieve clinically relevant conclusions, demographic data were recorded for each patient, including their age, gender, diagnosis, whether they were admitted to the hospital. Any important clinical events during their stay on the ward were also noted.

In the following chapters, we describe in greater detail how the vital sign data were collected in the ED, and then how the T&T chart was analysed in such a way as to minimise the number of transcription errors. We then investigate the effectiveness of the T&T system within an ED setting, by examining whether the practical problems in recording observations that result in incorrect T&T scores are exacerbated in the fast-paced ED environment in comparison to other hospital wards. After establishing the limitations of the current system, a further analysis will be performed to quantify the extent of errors made when calculating a T&T score.

To relate the T&T errors to clinical outcomes, we will use information about the important clinical events that occurred during each patient's stay. This will allow us to quantify the effect of incorrect score calculations on patient outcome. In conjunction with this, we

will retrospectively calculate a computer-generated T&T score directly from the patient observations, a process which will eliminate some of the T&T errors. Using a novel analysis framework described in Chapter 2.5, it will then be possible to work out the sensitivity and specificity of both the manual and computer-aided systems. By selecting the more accurate system, we will then consider how improvements in vital sign monitoring can be made in the ED.

In addition to improving the way T&T is used on the ward, a major aspect of this thesis is an investigation of how *continuous* data may be used to detect patient deterioration in the Majors and Resus areas of the ED. By using continuous vital sign data collected during the same study, it will also be shown how the T&T score criteria could be applied continuously in order to detect patient deterioration in real-time, between nurse observations. The advantages of the continuous system will then be quantified by comparing it to the manually-scored system.

In order to create more robust measures of patient deterioration, we investigate alternative methods to the existing T&T systems using continuous vital sign data. An existing baseline data fusion model [111] will be introduced and applied to the ED study data set. A number of alternative methods will also be investigated, and the effectiveness of each system will be compared to the continuous and manual T&T systems. The system with the best performance will then be optimised for use within the ED.

The final part of this thesis is concerned with further improving data fusion models by taking into account time dependencies within the data. A time series analysis method known as Gaussian processes will be used to infer missing data and to predict a patient's physiological status in the short-term, with the aim of allowing earlier detection of deterioration. Finally, there will be a discussion on future directions that this research may take. This will include an overview of how an electronic T&T system may be implemented alongside an intelligent continuous monitoring system within the ED.

2. Vital Sign Observations in the Emergency Department

In Chapter 1, we highlighted how a typical Emergency Department (ED) was organised. In particular, we described how management of patients in the more acute areas of the ED included intermittent observations of the vital signs, which were then assessed with the benefit of a Modified Early Warning Score within a so-called "Track and Trigger" (T&T) system. Observations are "Tracked", and an appropriate intervention can be immediately "Triggered" when certain criteria are met. In addition to intermittent observations, patients in the Majors, Resus and CDU areas also had their vital signs continuously monitored using a bedside monitor. Typically, an audible alert is generated when any of the vital signs crosses a pre-set threshold.

This chapter describes a single-centre prospective study in which both intermittent observations and continuously monitored vital sign data were acquired from an ED at a local hospital. Initially, we describe how the data were collected, and then proceed to explain how the data from three disparate sources were processed to provide an error-checked database containing continuous vital signs, nurse observations, and patient demographics for each of the patients.

The analysis that follows focuses on the manual observations and T&T scores; analysis of the continuous vital signs will be addressed in future chapters. In this chapter, we firstly examine how well the observations and scores are recorded on the observation chart. We expect these manually recorded observations to be imperfect as they are recorded by nursing staff in a high pressure environment, and our analysis will quantify the number of errors and attempt to determine the cause of the errors.

To count the number of errors, some measure of "ground truth" is required. "Ground

truth" was generated by calculating a retrospective error-free version of the T&T score by directly computing the score using the ED T&T scoring chart previously shown in Table 1.2. By using this chart, T&T scores were also generated for sets of observations that had not been assigned a T&T score while the patient was in the ED. In this case, if one of the observation parameters was missing from a set of observations, no T&T score was assigned for that parameter, but a total score was still calculated.

In the second half of the analysis, which is presented in Section 2.6, we investigate whether T&T scores can identify or predict deterioration within the ED. As we have already discussed, deterioration of a patient is difficult to assess, and the most obvious metrics, such as mortality at 30 days, are inadequate as the patient only stays in the ED, on average, for four hours. Instead, we chose escalation of care as our outcome marker, which we define as any documented event requiring intervention from clinical staff and include all the actions that may be triggered when the T&T alert threshold is exceeded. Each patient can therefore have multiple escalations.

The escalations of care are classified as one of six types: A1, A2, B1, B2, C1, and C2. Type 1 escalations are those that are caused by events that occur at presentation to, or prior to arrival in, the ED, whereas Type 2 escalations are those due to patient deterioration during their time in the ED. Type A escalations are due to abnormalities in the vital signs; type B relate to neurological dysfunction as determined by the Glasgow Coma Scale (GCS) score and by other factors recorded in the clinical record (e.g., epilepsy), and escalations that are neither type A or B are classified as type C. For example, a type C escalation is deemed to have occurred if a patient complains to staff of chest pain, a possible indication of myocardial infarction, which warrants some type of intervention despite the lack of any physiological signs at the time. The T&T criteria include the physiological vital signs, and a measure of neurological function, and should therefore be able to identify type A and type B escalation events.

The effectiveness of the retrospective T&T scores at identifying escalation events will also be assessed, and the results will be compared to the manual T&T scores.

2.1. The ED Study

The study had no direct effect on patient care, and participants in the study were cared for using standard hospital procedures. The study was approved by the Central Office of Research Ethics Committees (COREC), and funded by the National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford.

All adults entering one of three areas, the Resuscitation Room (Resus), Majors, or the Clinical Decision Unit, of the ED were eligible for inclusion in the study. Participants were selected for inclusion in the study if the last digit of their randomly allocated seven-digit hospital number was either '0', '5' or '7'. This was performed to limit the number of patients being monitored concurrently. Participants were excluded from the study if they were under 18 years old, unable to tolerate vital-sign monitoring for any reason, unable to understand English, did not consent to participate, or if they had fewer than three nurse observations of their vital signs while they were in the ED. A minimum of three recorded sets of vital-sign observations was deemed necessary *a priori* to reduce the influence of spurious observations, and to allow trends in physiology to be examined.

Each patient eligible for the study was provided with an information leaflet and verbal information about the study, and was then required to sign a release form before their data could be stored and used. For patients lacking in capacity to consent, the next of kin were asked to act on behalf of the patient, in line with the provisions of the Mental Capacity Act (2005). In the event of a full patient recovery, further attempts were made by the nursing staff to gain written consent.

The nurse vital sign observation data were collected from each participant during their stay in the ED, with the GCS score being used as the measure of consciousness. In addition to these, the pupil dilation and urine output were also recorded when nursing staff considered it to be appropriate. The GCS is a scoring system that measures the level of consciousness of a patient according to their response to visual, verbal, and motor stimuli. These are assessed according to the criteria given in Table 2.1. The subscores are summed so that the GCS takes a value between 3 and 15, with a fully comatose patient scoring 3 and a fully alert patient scoring 15.

T&T scores were calculated by nursing staff during their periodic observations, in keep-

	Eyes	Verbal	Motor
1	No eye opening	No verbal response	Makes no movements
2	Eye opening in response to painful stimulus	Incomprehensible sounds	Extension to painful stimulus
3	Opens eyes in response to verbal command	Utters inappropriate words	Abnormal flexion to painful stimulus
4	Opens eyes spontaneously	Confused, disoriented	Flexion/ Withdrawal to painful stimulus
5		Oriented, converses normally	Localizes painful stimulus
6			Obeys commands

Table 2.1.: Glasgow Coma Score criteria for the Eyes, Verbal, and Motor subscores (Teasdale and Murray [113])

ing with standard procedures using the scoring system of Table 1.2. Both the T&T scores and the manually-recorded vital-sign data were written down on an observation chart, which was collected after the patient was discharged. The clinical notes for each patient were also collected at the end of their stay in the ED, so that the observed vital signs could be linked retrospectively to a patient's clinical context.

Continuous measurements of RR, HR, SpO_2 and temperature were recorded at a sampling rate of 30 seconds. Intermittent measurements of BP were recorded at intervals related to the condition of the patient. The most acute patients had BP recordings taken every 5 minutes, whereas recordings were as infrequent as once per hour for patients with less serious conditions. The data were collected using a Phillips Intellivue® bedside monitor and then saved to a hospital data server.

A more detailed description of how the vital signs are recorded by the bedside monitors can be found in Chapter 1. The modern Phillips monitors are also designed to consider probe connectivity, and have specific alerts that indicate when probes are disconnected or when the signal is erratic. In these instances, the corrupted vital sign data are not output by the monitor, to the study server, until the probes are reattached. The time after probe disconnection at which probe-off alerts are generated is not documented in the user manual, but these alerts were found generally to occur within 30 seconds for continuously monitored variables except for blood pressure.

We note that the measure of temperature initially used, skin temperature, is not typically recorded in standard care but was used in preference to core body temperature because it is easier to monitor continuously. The skin temperature has an offset with respect to core body temperature, depending on peripheral perfusion.

The continuous vital sign measurements were displayed on the bedside monitors in real time because they were part of standard care for the most acute patients and nursing staff were already familiar with the monitors prior to the start of the study. The data were stored on a central Philips server, transferred to a local study server using a Health Level 7 (HL7) standard interface and then stored in an SQL database.

The study data were collected between January 2009 and January 2010. All patients who attended the ED on more than one occasion during the study period were included as separate episodes for the purpose of analysis. Over the study period, a small proportion (13/476) of patients attended the ED on more than one occasion, accounting for 33 separate attendances.

In the first part of the study, which included the period from 15$^{\text{th}}$ January 2009 to 5$^{\text{th}}$ May 2009, participants were enrolled into the study at any time of the day. However, during the second part of the study, which included all dates between 1$^{\text{st}}$ September 2009 to 15$^{\text{th}}$ January 2010, only patients that attended the ED between 9am and 6pm were enrolled due to changes in the working practices of the research nurses. No patient data were recorded between May 10$^{\text{th}}$ and September 1$^{\text{st}}$ 2009 because of unavoidable administrative and technical problems.

Data were collected with the aid of a team of three research nurses. The research nurses were directly in charge of consenting patients to the study and entering the patient's hospital number into a bedside monitor. Additionally, a senior clinical team consisting of an ED nurse consultant, an ED consultant and an ED specialist registrar also assisted in the study. The senior clinical team was responsible for analysing the hospital records for each patient to determine whether an escalation event had occurred, and then assessing the type of escalation. Both the research nurses and the senior clinical team were responsible for creating electronic backups from the manual observation charts data and then transcribing the data into one of two electronic databases.

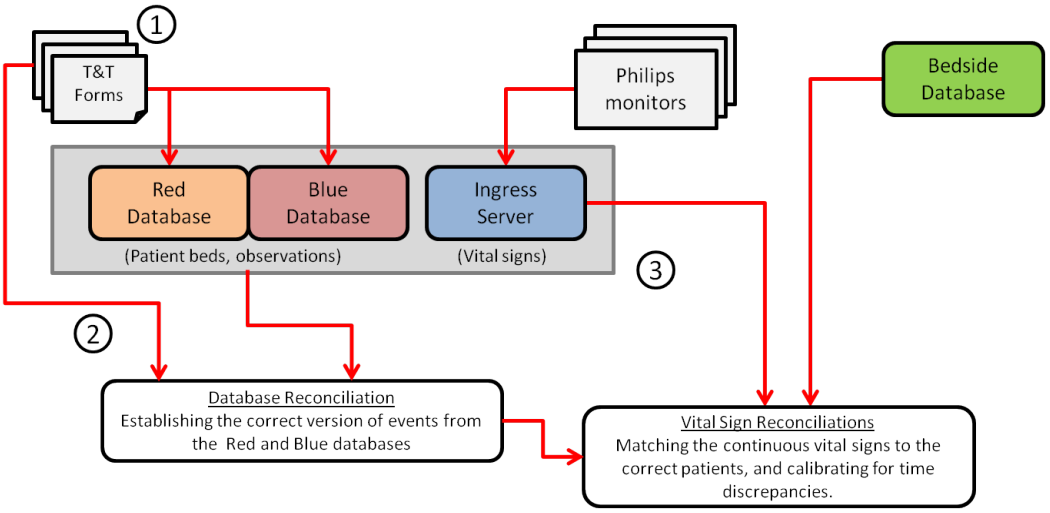

Figure 2.2.1.: Flowchart showing the reconciliation of manual observations and continuous vital sign data into a single error-corrected database

As well as the two hospital-based teams, a university research team consisting of two doctoral students and a post-doctoral researcher created the database infrastructure for the study. Once database entry had been completed, the university research team also reconciled the two databases to provide a single error-corrected database.

2.2. Data Reconciliation

The data collection process provided three sources of data: the observation charts, the patient notes, and the continuous vital sign data from the Philips monitors. To cross-correlate and analyse the disparate sources, it was first necessary to create an electronic record of all the data, and to assign the continuous data to the study patients. The process is shown in Figure 2.2.1 and described below.

Firstly, each individual observation from the paper observation charts was transcribed into an electronic database. To minimise the number of transcription errors, every observation was transcribed into "red" and "blue" databases by two teams who worked independently of each other, and each team was given the task of entering observations into a single database only. The two teams were generated by randomly selecting a total of three members from the research nurse team and the senior clinical team. The paper observation charts were also scanned and saved for reference. In addition to the observation charts, selected anonymised clinical data (presenting complaint, triage category,

location in ED, discharge destination) and demographics (age, gender, date and time of ED attendance, and time of ED discharge) were also entered into both of the databases by both the "red" and "blue" teams.

During the transcription process, the data for each study patient were anonymised to conform to the ethics requirements. This was achieved by replacing the patient's name and hospital ID with a 5-digit study ID that had a prefix beginning "ED". The study ID was generated automatically and sequentially, so that the first patient consented to the study had the ID "ED00001", the second patient, "ED00002", and so on. The study ID was also written on the paper observation charts by the research nurses so that patient notes could be recovered easily if required. The mapping between hospital ID and study ID was kept in a separate file that was not accessible to anyone apart from the research nurses and the senior clinical team.

Secondly, the university research team checked red and blue databases for transcription errors to provide a final reconciled database of the observation chart information. Any vital sign values that differed between the two databases by an insignificant amount, as determined by the senior clinical team, were attributed to difficulties in interpreting the observation charts.

HR and BP were recorded graphically on the observation chart, on which each division on the chart represents 10 beats/min or 10 mmHg. The divisions on the observation chart are closely spaced physically, so that it is difficult to determine the value of the observation if it is between divisions. Therefore, an error was considered to be insignificant if it was less than to be less than 10 beats/min (bpm) and 10 mmHg for HR and BP respectively. RR, temperature, and SpO_2 are typically written numerically on the observation chart. Consequently, the constraints were much more stringent on these observations: 1 respiration/min (rpm) for RR, 1^oC for temperature and 1% for SpO_2. If the vital sign measurements were within these boundaries, the value recorded in the reconciled database was computed as the mean of the values in the red and blue databases. This accounted for 96.2% of the vital sign observations.

The remaining 3.8% of observations had discrepancies that were greater than the acceptable errors. These were independently checked by a member of the university research

team, who manually examined the relevant scanned observation charts. In most cases, this allowed the university researcher to select the correct value when the value in either the red or blue database had been entered incorrectly. In the remaining cases, differences in the red and blue databases occurred when the observations were illegible or ambiguous. If a decision could not be reached satisfactorily after studying the charts, the vital sign observation was discarded. The output of this process was an error-corrected set of observations for each patient.

The third stage of the reconciliation process was to assign continuous vital sign data from the Ingress server to the correct study patient. Each vital sign was stored with a date stamp and the bed number from which the data were recorded. However, this stage of the reconciliation process was non-trivial as specific patient identifiers were not recorded for the data set.

The solution of this problem required the patient admission and exit time information from the reconciled database. In addition to this, another independent source of admission and discharge time was acquired using an additional bedside database, in which times were recorded from the computer's internal clock when the attending nurse pressed a button to indicate the arrival or discharge of a patient. Using these admission and discharge times, it was possible to estimate the start and end time of each vital sign record for each study patient, as the admission and discharge times should approximately match with periods of vital sign activity. In addition, this process allows us to confirm which beds the patient occupied. The relevant continuous vital sign data could then be extracted from the Ingress server and displayed alongside the manual observations. The continuous data were assigned to the patient if:

- Continuous data recording started and ended within 30 minutes of the admission and discharge times
- There was a high degree of similarity (see below) between the continuous vital sign values and the manual observations during the same period.

The flexible start and end times were required because the times recorded in the reconciled database were often estimates that had been written down retrospectively, soon after the patient had left the ED. The degree of similarity between manual and continuous

observations was determined by expert review from a member of the university research team. The reviewers were specifically told to look for instances when the blood pressure values were the same in both data sets. This was a strong indication of a match, as the manual observations of the systolic and diastolic BP values that were entered on the paper T&T chart were usually the same as the values displayed on the Philips monitor, which corresponded to the last inflation of the blood pressure cuff prior to the nurse observations.

This process required extra care to ensure that data were correctly matched during April and October, when daylight saving time changes affected the times recorded on the T&T charts with respect to the Philips monitors and Ingress server, whose internal clocks were set to go forward one hour at 1a.m. on 29th March 2009 and back one hour at 2a.m. on 25th October 2009.

The process was further complicated by the fact that the Philips bedside monitors were regularly serviced and then re-installed in a different ED cubicle so that the bed from which the continuous data had been collected was no longer clearly identifiable. To solve this problem and to maximise the amount of matched continuous data, we adopted a more thorough search strategy which involved reviewing the continuous vital sign data in all the beds when it was suspected that monitor locations had been switched.

In addition to the continuous vital sign data and the observation data, the final data set also contained the escalations of care for each patient and the error-free retrospective T&T scores, which were both defined at the start of the chapter. These were identified using the clinical notes, according to the criteria which were also provided at the start of this chapter. For each patient, the time and the cause of any escalations were assessed independently by the ED consultant and ED nurse consultant from the senior clinical team. Where disagreement arose, the third team member made an independent decision. The error-free retrospective T&T scores were generated using the reconciled vital sign observations, based on the JR T&T scoring criteria of Table 1.2

The completed data set contained vital sign observations, the corresponding T&T scores, error-free retrospective T&T scores, continuous vital sign data, copies of the original observation charts, and a list of the times and causes of escalations of care for each study patient.

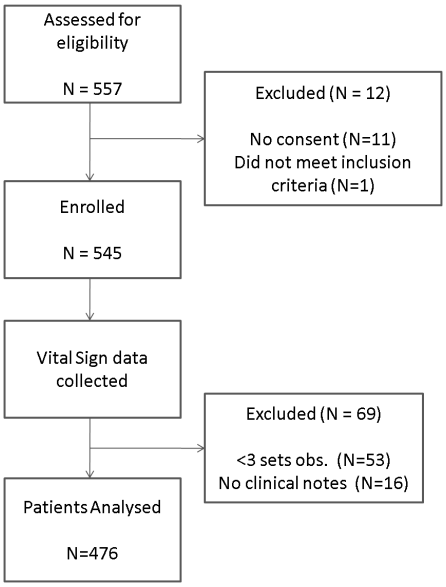

Figure 2.3.1.: A consort diagram showing the patient inclusion for the ED study

2.3. Data Overview and Completion Rates

557 patients were recruited to the study, and 476 patients fulfilled the study inclusion criteria. The breakdown of patients is shown in the consort diagram (Figure 2.3.1). The mean age of the patients was 61 years (range 18-108, IQR 43-79), and the patient demographics shown in Table 2.2 show that a valid representative sample of patients was selected. The 53 patients excluded from the analysis for having fewer than three sets of vital-sign observations were similar to the study group, but were less likely to be admitted to hospital (46.7% admitted).

3025 sets of vital sign observations were recorded from the 476 study patients. Over 99% of patients had at least one full set of observations (HR,RR, BP and SpO_2). GCS and temperature were not considered as part of the "full set" as they are typically not indicated to be recorded as frequently as the other observations if the parameters are within normal limits. Despite this, 89% of patients also had their temperature and GCS recorded at least once whilst in the ED. In the overwhelming majority of cases, the urine output and pupil dilation were not recorded because they were deemed clinically irrelevant. These measurements were therefore excluded from the analysis.

The histogram for the time between consecutive observations in the data set is shown in Figure 2.3.2. The mean time between observations was 65 minutes. The longest time between any two observations was 9.8 hours, for an elderly patient who had fallen and

Gender	255 male (52%) 236 female (48%)	
Age Range	18-99	
Age mean (s.d.)	61 (21.8)	
Initial Location of patients in ED	Resus Majors CDU Other (unspecified bay in CDU or Majors, Minors)	122 276 6 87
Admitted in hospital	290 (59%)	
Discharged from hospital	201 (41%)	
Of those discharged:	GP follow up Outpatient Department No follow up Left before/ refused treatment Died in department Unknown	102 22 56 2 2 17

Table 2.2.: Patient Demographics for the ED study

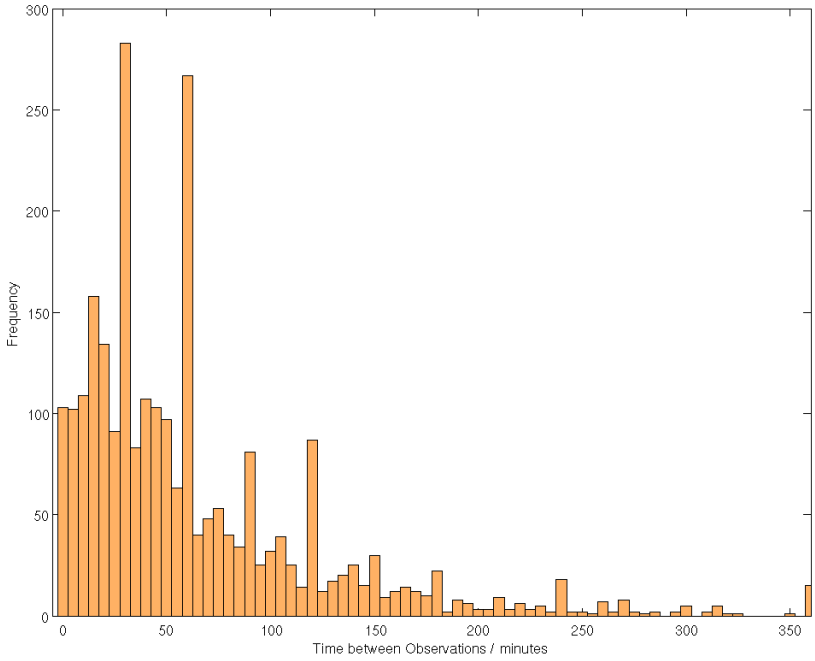

Figure 2.3.2.: Histogram showing the time between consecutive nurse observations (including CDU patients)

had minor bruising. He was kept in the CDU overnight, but regular observations during the night were not deemed necessary. If patients in CDU are excluded, the maximum time between observations decreased to 5.4 hours.

The spikes in the histogram correspond to 15, 30, 60, 90 and 120 minutes between observations. This matches the expected observation frequency in the ED, where patients are observed every hour in the Majors area if vital signs are normal, and at 30 and 15 minute intervals when the T&T score is greater or equal to 2.

Of the 3025 sets of vital signs, only 1037 (34.3%) had corresponding T&T scores, and T&T scores were recorded at least once for just 60.3% of the study patients during their stay at the ED. Of the 1037 T&T scores, 44.4% were scored as a zero, and 20.1% exceeded the T&T threshold. In comparison, by using the arithmetic-corrected retrospective T&T scores we calculated that 43.7% of the 3025 sets of observations should have been scored as a zero, and the T&T thresholds should have been exceeded in 26.0% of cases.

The percentage T&T completion by month is depicted in Figure 2.3.3. The T&T completion rate significantly improved during the second part of the study, with a 51.8% completion rate in comparison to 20.1% for the first part. The improvement in completion rate was expected by the clinical staff during the study as a result of changes in the way that nurses were trained, but it may also have been partly due to the fact that patients were only admitted during the day for the study in the later months. As an aside, we also note the relatively high completion rate in January 2009, and suggest that this may be caused in part by the Hawthorne effect (see Section 1.3.4 for definition). The completion rate between June to August was zero due to the data collection problems mentioned in Section 2.1.

The percentage T&T completion per hour of the day is shown in Figure 2.3.4. To remove the bias caused by the fact that only day patients were accepted during the second half of the study, the Figure only shows the T&T completion for the patients attending the department before 5th May 2009. Overall, T&T completion is slightly better during the day between 9a.m. and 5p.m. (23.6%), compared to at night between 11p.m. and 7a.m. (17.0%). In addition to this, there are significant increases in T&T completion rates between the hours of 8-9a.m., and 7-8pm, and 9-10pm, which coincide with changes in

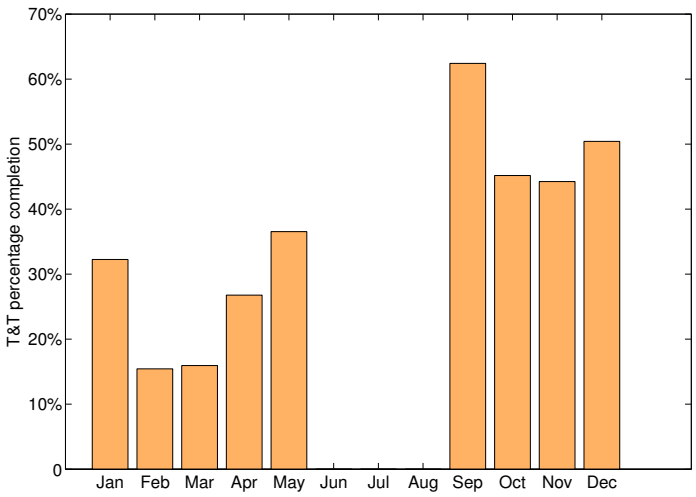

Figure 2.3.3.: Percentage of T&T scores calculated for each month of the ED study.

nursing shifts. There is also a significant spike between the hours of 12a.m. and 1a.m., which can be attributed to the fact that observations that contained a date without a specific time were automatically assigned to midnight due to a limitation of the method (Matlab's *datestr* function) used.

2.4. Incorrect T&T Calculation

In addition to a low T&T completion rate, the T&T score was also calculated incorrectly in 211 out of 1037 cases when compared to the retrospective T&T scores. Of those that were incorrect, 34 were calculated incorrectly in situations when the retrospective T&T score exceeded the triggering threshold, and the nurse-recorded T&T score was below the threshold. These indicate instances where errors in T&T calculation could potentially have led to deterioration being missed. Figure 2.4.1 further shows that most erroneous T&T scores differed from the retrospective score by one or two points (16.2%), but that a sizeable proportion, 4.1% of observations, had scores that differed by three or more points. The skew on the graph indicates that nursing staff tended to under-score their patients, reducing the overall number of T&T alerts, but making it more likely that patient deterioration could be missed.

Using the reconciled database and scanned observations charts, four potential reasons for errors in the T&T totals were identified, and are described below with some examples

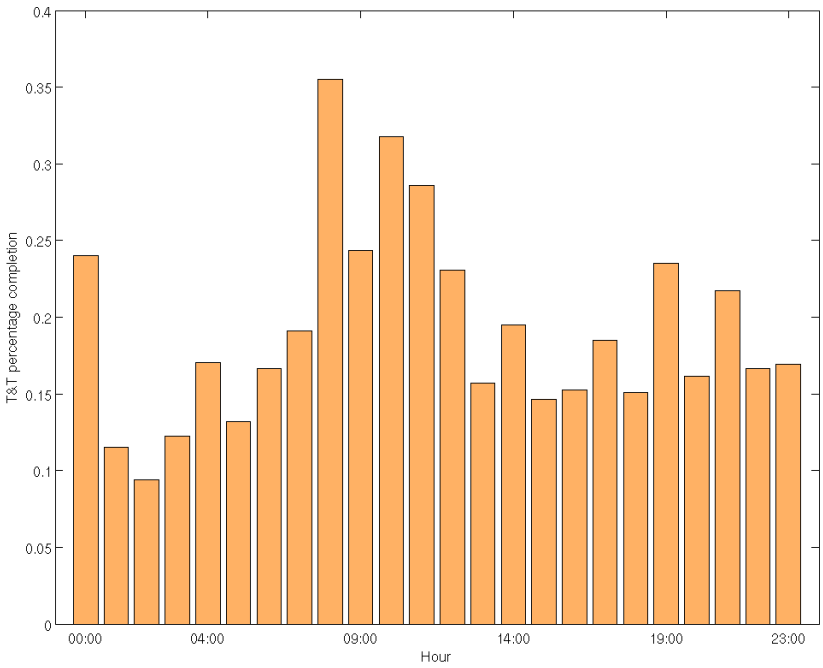

Figure 2.3.4.: Percentage completion of T&T scores per hour in the ED during the study.

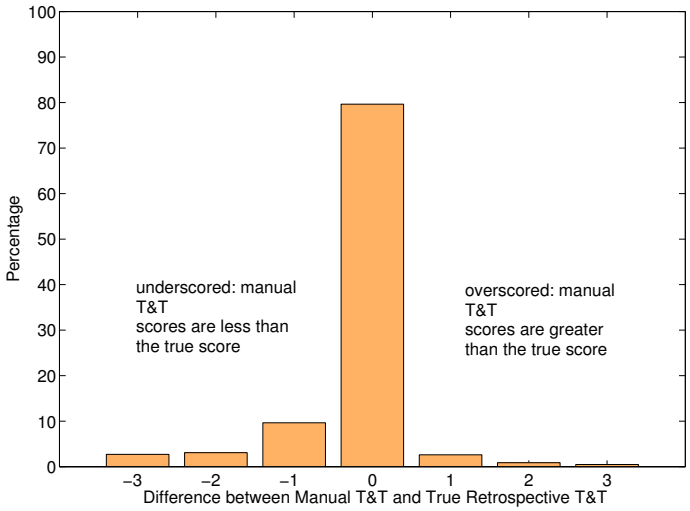

Figure 2.4.1.: Bar chart showing the difference between the score as calculated by the nursing staff and the retrospective error-free T&T score.

taken from the study.

Incorrect T&T Parameter Score Allocation

We describe an incorrect parameter score allocation as a situation where the recorded vital sign parameter and the corresponding T&T parameter score are inconsistent, based on the T&T criteria of Table 1.2. For instance, patient ED00223 was given a score of 1 for high respiratory rate (Figure 2.4.2) whereas the actual recorded respiratory rate of 33 respirations/minute should have led to a score of 3, and senior staff should have been notified. In this case, no action was taken, and the patient's respiratory rate remained high at the next observation, indicating that it was not a spurious measurement. Two members of the clinical team independently reviewed this case and noted that a trigger should have been raised and may have led to a change in patient management, although the clinical significance of any potential intervention is unclear from the notes, as the patient was later discharged from CDU.

Further examination of the patient records shows that incorrect parameter score allocation often occurred when patients were known to be hypertensive or to have Chronic Obstructive Pulmonary Disease (COPD), and were therefore expected to have high systolic blood pressure, or low SpO_2 respectively. In such cases, it is possible that the attending nurse will note the abnormal vital sign, which should have scored 3 and triggered an alert, but then correctly deemed the result to be clinically irrelevant, and scored it as a zero.

This could explain why incorrect T&T scores appear to be underscored, as shown in Figure observation in Figure 2.4.1, and may also explain why the discrepancies between the manually-calculated and retrospective T&T scores were greater than 1 in many cases. Incorrect parameter score allocations were found for 202 of the 1037 T&T scores, making it the most common source of errors.

Incorrect T&T Score Addition

T&T scores were deemed to be incorrectly added if the total T&T score did not equal the sum of the individual parameter scores, regardless of whether or not the parameter scores

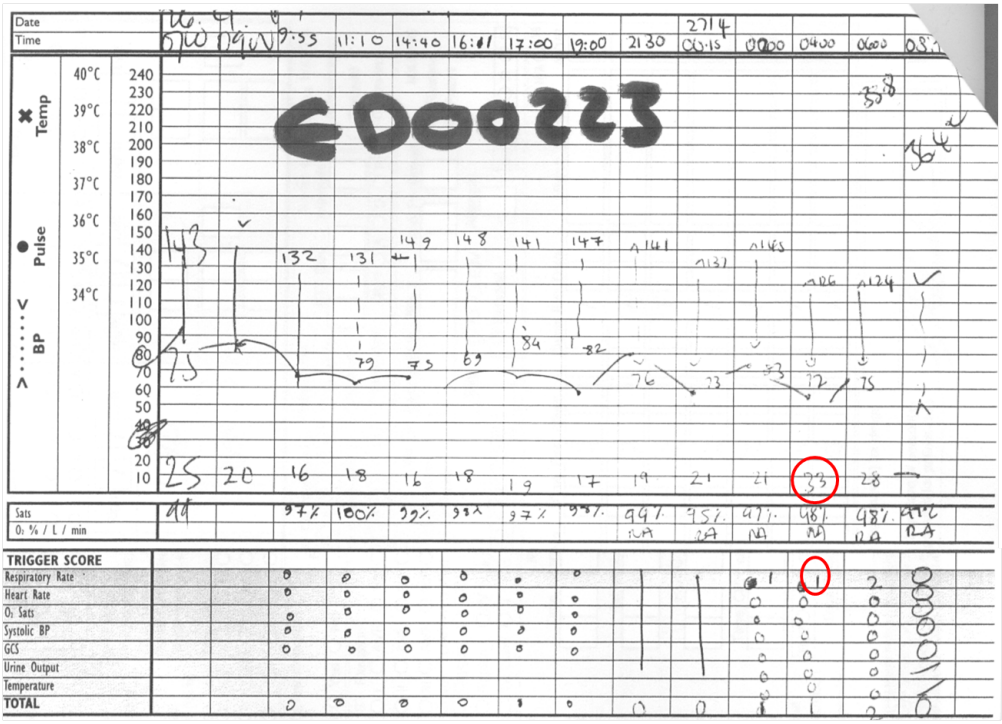

Figure 2.4.2.: Part of the observations chart for patient ED00226, showing an incorrect T&T total caused by incorrect allocation of the respiratory rate sub-score.

were correct to begin with. This type of error accounted for 14 observations in the study. An example of this type of error is shown in Figure 2.4.3 for patient ED00360, a 79-year old man who entered the ED with a consistently high heart rate because of paroxysmal atrial fibrillation.

At 15:45, the patient was administered with Amiodarone to treat the symptom, and at around the same time, the patient's systolic blood pressure dropped below the normal range. This, in combination with a high heart rate and slightly elevated respiratory rate, should have led to a total T&T score of 6, but the observations were incorrectly scored as a 4. The T&T total score of 4 met the T&T criteria for triggering further action, and so this mistake was unlikely to have effected change in patient care. However, the correct score may have alerted staff to possible further deterioration in the patient's status, as the correct T&T score of 6 is more severe than the patient's previous scores of 4 and 5 at 15:10 and 15:15 respectively.

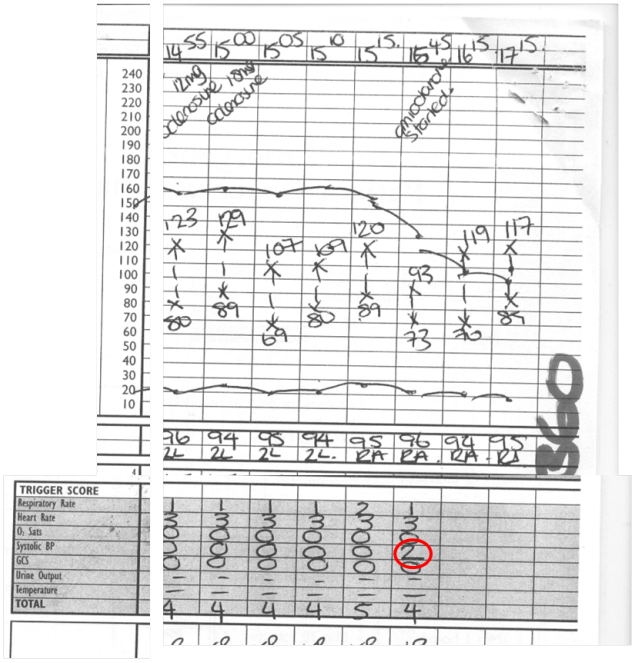

Figure 2.4.3.: Example of incorrect T&T addition for patient ED00360. The score for the observations at 15:45 should have been a 6, but was incorrectly calculated as a 4.

Incorrect GCS Recording

The GCS is comprised of three subscores, and is thus also prone to incorrect addition. For the 2686 GCS observations, 13 had been added up incorrectly. Eight of the thirteen mis-scored observations were due to "follow-through" error, where a new observer simply copied the previous score without doing a new addition. An example of both of these errors is shown for Patient ED00196 (Figure 2.4.4), where the attending staff incorrectly calculated the GCS total as 12 at 18:35, and made the same mistake 35 minutes later during the next observation. The medical impact of this error was limited by the fact that the patient had previously been transferred to the highest acuity area of the ED for neurological reasons, so the condition of the patient was being heavily monitored at the time.

The recording of GCS is also prone to error during its measurement. To measure the GCS, an observer must be present to judge a patient's response to Pain, Voice, and their Eye-Movement. This often involves speaking to, and physically moving, the patient. Because the score depends on interaction with the observer, it may be argued that it is difficult to arrive at an objective GCS score. The difficulties with recording GCS are well known, and Rowley and Fielding [?] have reported that while inter- and intra-rater

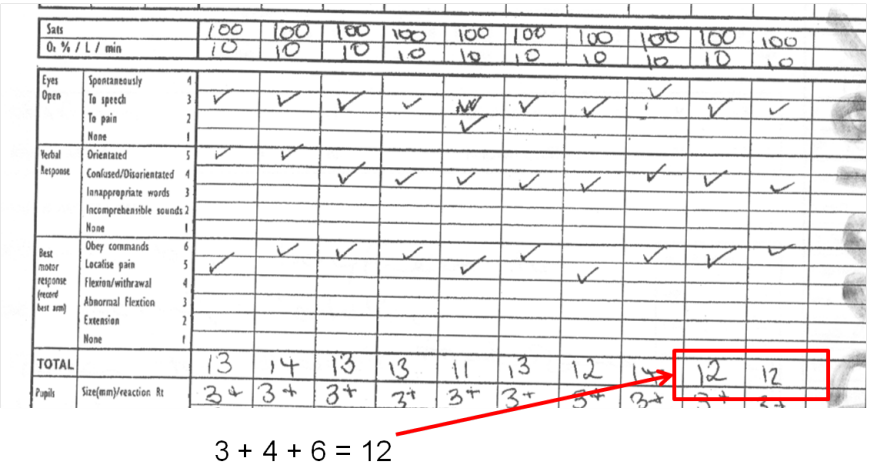

Figure 2.4.4.: Partial view of observation chart for Patient ED00196 demonstrating addition error in GCS total and follow-through of error

GCS estimates were consistent, inexperienced nurses often underestimated the score when compared to a trained expert. The systematic errors in GCS recording are made before the observation is entered on the observation chart, and so our analysis will be unable to detect how prevalent these problems are.

Incorrect Temperature Recording

During the study, manual observations of temperature were made using tympanic thermometers to measure core temperature. Continuous monitoring of temperature was also attempted by using thermistors, and the Phillips bedside monitor displayed the skin temperature values in real-time. This is different to standard practice, for which no temperature measurements are displayed on the bedside monitor. The distributions of the two sets of temperature measurements are shown in Figure 2.4.5. The figure indicates that there are significant differences between the two types of temperature measurements. The skin temperature has a wider interquartile range than the core temperature (1.7^oC in comparison to 1.0^oC), and a median of 34.8^oC, compared to 36.1^oC for core temperature.

The two methods of measuring temperature gives different values, but core temperature is used in the T&T score charts. However, we surmise that the display of skin temperature on the bedside monitors may have led some nurses to record skin temperature on the T&T chart instead of using a tympanic thermometer to measure temperature.

To test this hypothesis, we calculated δ, the difference between the intermittent core

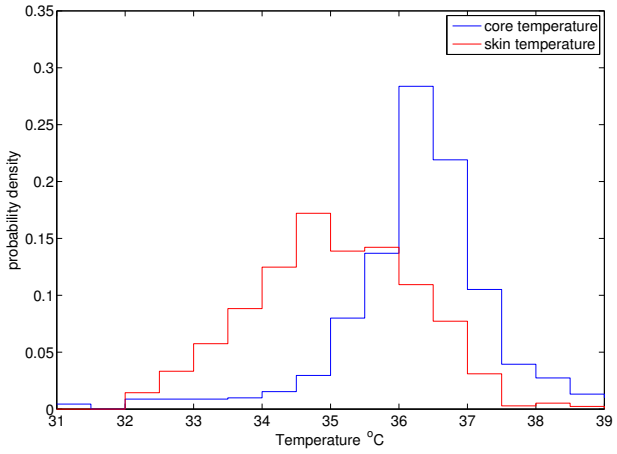

Figure 2.4.5.: Discrete probability density distributions of intermittent core temperature measurements as recorded by nurses (blue), and continuous skin temperature recordings (red). We hypothesise that some of the core temperature measurements, particularly those with unusually low temperature, and where hypothermia was not mentioned in the doctors notes, were not measured using a tympanic thermometer and were instead taken from the skin temperature reading.

temperature measurements and the corresponding continuous skin temperature observation. Due to minor differences between the time recorded on the bedside monitors and the manually recorded time, we assumed that the manually recorded time was correct, and that the time-stamp from the continuous monitor would be accurate to within ± 5 minutes. Therefore, the skin temperature value used to calculate δ was taken to be the mean temperature value within a 10-minute window centred at the time of the manual observation.

The distribution, and the cumulative distribution of δ are shown in Figure 2.4.6. In total, there were 912 temperature observations during the course of the study. However, only 102 observations were used in this analysis, as all other temperature observations did not occur at times when the patient's skin temperature was also being monitored.

If the difference between the core and skin temperature measurements was less than a critical threshold, then we considered it likely that the manual observation had been copied from the bedside monitor. The critical threshold was established by assuming that the core temperature was written down from the bedside monitor if it matched the continuously recorded skin temperature at any point within the ten-minute window; therefore, the core temperature is deemed to be a copy of the bedside monitor value if:

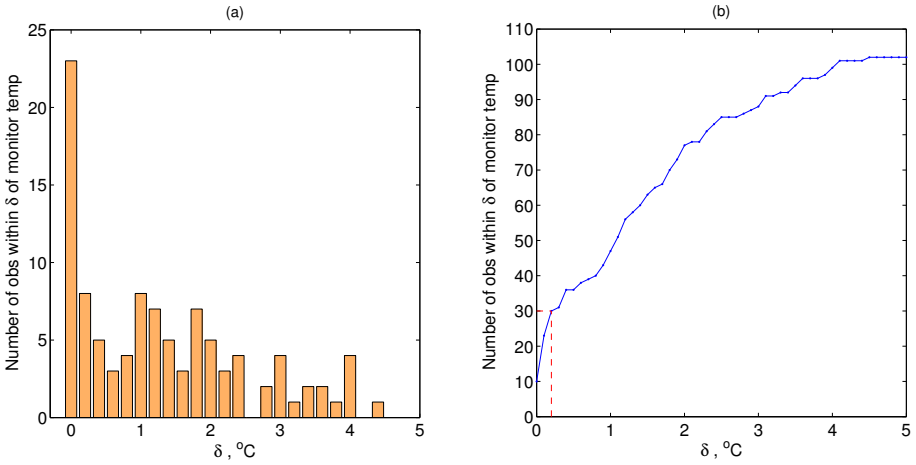

Figure 2.4.6.: (a) Histogram of the temperature observations within δ of the monitor temperature (b) The cumulative distribution showing the number of temperature observations within δ of the monitor temperature, for increasing δ.

$$temp_{skin} - \frac{1}{2}range(temp_{skin}) \leq temp_{core} \leq temp_{skin} + \frac{1}{2}range(temp_{skin}) \qquad (2.4.1)$$

The mean range of the skin temperature over all of the 10-minute windows was calculated to be 0.4^oC, so the critical threshold was set at $temp_{skin} \pm 0.2$. The cumulative distribution of δ is shown in Figure 2.4.6(b). The dotted line highlights that 30 out of 102 temperature observations were within 0.2^oC of the skin temperature as displayed on the bedside monitor, and thus likely to have been erroneously copied from the bedside monitors.

The histogram in Figure 2.4.6(a) also confirms that a significant proportion of temperature observations appear to have been copied from the bedside monitor, showing that 23 manual observations were within 0.1^oC of the continuous skin temperature reading. In addition, the distribution appears to be bimodal, with one major peak occurring when $\delta = 0$, and another peak occurring when $\delta \approx 1.0$, which is approximately the same as the mean difference between the core and skin temperatures.

Illegible Observation Sheets

There were numerous instances where observations could not be read properly from the observation chart. In some cases, the time stamps were illegible, while in other instances,

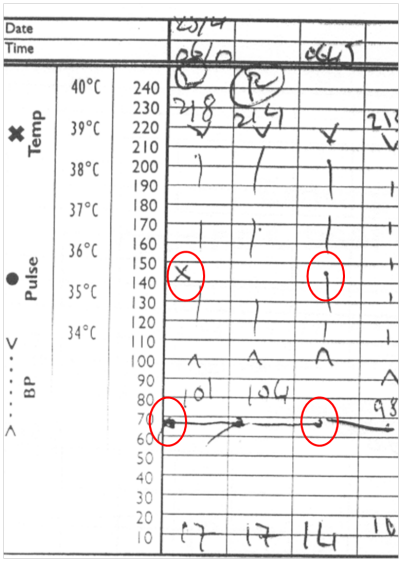

Figure 2.4.7.: Partial observations chart for Patient ED00560, showing an instance where there are spurious crosses and dots (the upper red circles), which may be confused with heart rate measurements (lower circles)

such as that shown in Figure 2.4.7, it was unclear which vital signs were being recorded on the chart. This may lead to confusion when trying to determine long-term trends in the data.

We also observed that the nursing staff used a variety of styles to fill in the observation chart. In some cases, the style changed within the same observation chart, as shown in Figure 2.4.8. In this example, the blood pressure was initially recorded with inverted arrows. Later on, at 1320, the systolic blood pressure may have been written down numerically (although the numbers may also refer to the heart rate), and later on, by 1340, the blood pressure was recorded using normal arrows.

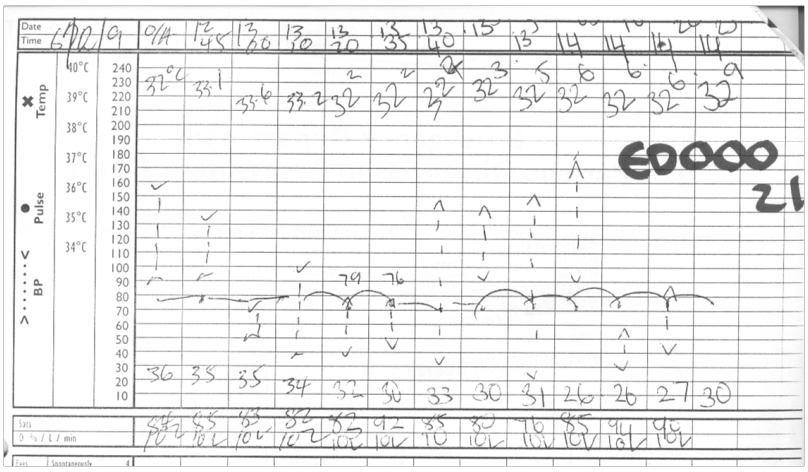

Figure 2.4.8.: Observations chart for patient ED00021, demonstrating multiple styles. Of particular note are the three different styles for recording blood pressure.

2.5. Sensitivity and Specificity Analysis for Multiple Observations

Up until this point, we have considered the types of errors made while recording T&T charts, and how often these errors are made. However, it is unclear what impact the errors have on the standard of patient care. In order to evaluate this, we assessed the effectiveness of the T&T scores at predicting escalation events (as defined in Section 2.2) using the analysis framework described below.

In Chapter 1, we briefly outlined how Early Warning Score (EWS) systems, or indeed any other classification system, could be evaluated using the concepts of sensitivity and specificity. In order to compute these metrics, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) must first be calculated.

In this case, the test is the T&T system, which outputs either a positive (T&T score exceeds one or more alerting criteria) or a negative (T&T score does not exceed one or more alerting criteria) result. To assess whether the test result is true or false, we require an outcome marker that represents ground truth. For instance, in Smith et al. [100], a positive event was recorded if the initial EWS, or Track and Trigger (T&T), score exceeded the alerting threshold, and was considered to be a TP if the patient died before hospital discharge, and a FP otherwise. In this case, the outcome marker was in-hospital mortality.

Unlike Smith et al., our study considers the *multiple* T&T scores recorded during the course of a patient's stay in the ED. Furthermore, the outcome markers were chosen to be the escalations of care defined at the start of Chapter 2, which had been documented according to the procedure described in Section 2.2. Because a patient may deteriorate several times during their stay, *multiple* escalations may be recorded for each patient. These multiple escalation events for each patient are unlikely to be independent. This makes sense clinically, as the initial escalation may trigger further secondary escalations. Similarly, prompt intervention after an initial escalation may prevent further escalations. Therefore, in order to provide the most clinically relevant result, and to avoid the problem of dependent events or outcome markers, the main aim of a classification system such as

T&T should be to correctly identify the first instance of deterioration. This allows the sensitivity and specificity analysis to focus on a single outcome per patient.

Using the T&T score and escalation of care as the test and outcome marker respectively, we would like to build a confusion matrix from which we may calculate sensitivity and specificity. We now have multiple tests over time for each patient, however, whereas sensitivity and specificity are usually assessed for a single (diagnostic) test and a single outcome measure. We now go on to describe how TPs, FPs, TNs and FNs can be defined within our analysis framework.

2.5.1. True Positives and False Negatives

TPs or FNs are only evaluated on the group of patients with one or more escalations, "escalation patients", using the first escalation in time as the event or outcome marker. We assign a TP classification if the first escalation is correctly identified from the T&T score; conversely, we define a FN if the first escalation is not successfully detected. An escalation can be considered to have been detected if the T&T score generates an alert within a time, t, ahead of that escalation, or a time, τ, after the escalation. The time τ is included because, in practice, the timestamps for escalations are inaccurate.

The time, t, is included to reflect our prior belief that an optimised patient monitoring system may be able to provide early warning of the escalation. t cannot be extended indefinitely, because the association between the test and the escalation decreases as t increases. In our initial analysis, we will assume a conservative value of $t = 10$ minutes. Longer window lengths may be more appropriate and are considered later. A pictorial representation of this process is shown in Figure 2.5.1 and explains how this framework deals with an "escalation patient".

2.5.2. False Positives and True Negatives

In this section, we present an equivalent per-patient framework for FPs and TNs. We can achieve this by first defining "normal" patients as the group that comprises all patients who had no escalations. An FP classification is then defined to occur when the data from a "normal patient" generates one or more alerts as a result of the T&T alerting thresholds

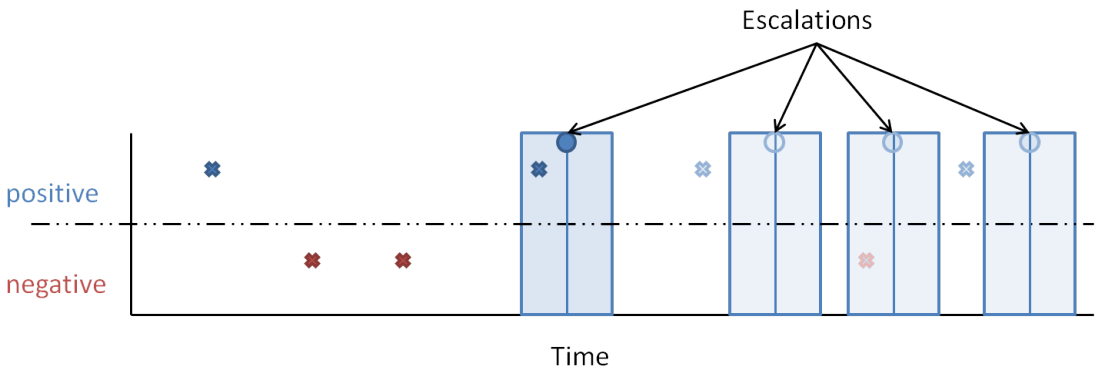

Figure 2.5.1.: Detection of a True Positive for an "escalation patient" within the analysis framework. Test results are shown as crosses, while the escalation events and their associated windows are labelled. The first escalation is classified as a TP as the positive result occurs within the time window associated with the escalation event. The remaining escalations are greyed out to denote that only the initial escalation is considered, and that the framework requires that the "escalation patient" is classified as a True Positive or a False Negative.

being exceeded; conversely, we define a TN classification to occur when the data from a "normal" patient generate no alerts. A pictorial representation of FP and TN patients (with no escalations) is shown in Figure 2.5.1.

2.5.3. Shortcomings of the Framework

The approach we have taken for describing TPs, TNs, FPs and FNs is able to deal with multiple tests and a single outcome marker. This per-patient framework makes clinical sense, as the aim is to generate alerts, as early as possible, for all the patients whose care is eventually escalated. Furthermore, there should be as few false alerts as possible for the "normal patients" group.

While the framework is consistent, there remain some minor drawbacks. Firstly, its accuracy depends on the availability of the information for the outcome markers (escalations). In practice, escalations that relate to transient abnormality are most likely to be missed. Therefore, it is possible that a small percentage of the "normal patients" should be in the "escalation patients" group. The problem of incomplete recording of events in the clinical environment is largely beyond our control, but we note that the effect should be the same for all systems that are tested, so comparisons within the framework are valid.

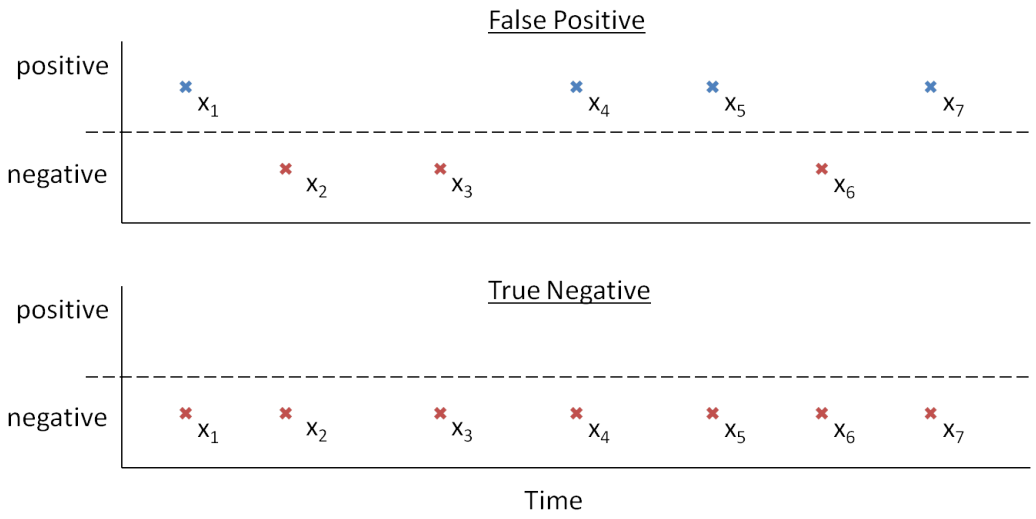

Figure 2.5.2.: Both of these examples show records from "normal patients" with no escalations. In the top example, x_1,x_4,x_5 and x_7 show positive results, and thus the record is classified as a FP. In the bottom example, all of the test results are negative, and correctly reflect that there are no escalations. Therefore the record is classified as a TN.

Secondly, the framework risks oversimplifying the treatment of FPs. In our framework, both a "normal patient" that has multiple false alerts, and a "normal patient" with one false alert will be counted as a single FP. This does not reflect the realities of clinical practice, where each false alert may lead to separate interventions, so the number of false alerts for each patient is important. Similarly, a transient false alert is likely to have a smaller effect than a long-term false alert.

The impact of transient alerts does not need to be considered for T&T systems because each T&T alert does not persist for any length of time. However, the length of alerts is an important issue when we will be considering alerting systems based on continuous vital sign data. In order to achieve a more thorough analysis of these systems, we will report distributions of the duration and number of alerts per patient in addition to the sensitivity and specificity. We will also report a false alert rate, which is calculated as the total number of alerts for the "normal patients", divided by the total time of vital signs recorded for the "normal patients", and has units of alerts/hour per bed. The framework that has been presented here will form the basis of the analysis in the following chapters for analysing the effectiveness of both T&T systems and continuous monitoring systems for detecting patient deterioration.

Escalation Type	A1	B1	C1	A2	B2	C2
No. Escalations	75	22	96	51	13	27

Table 2.3.: Break down of the escalation events according to whether they occurred pre- or post-arrival in the ED, and according to the cause of the escalations. Type-1 escalations were those that were caused by events that occurred at presentation to, or prior to arrival in, the ED, whereas Type-2 escalations were those due to patient deterioration during their time in the ED. Type A escalations were due to abnormalities in the physiological vital signs, type B related to neurological dysfunction as determined by the GCS score and by other factors recorded in the clinical record (e.g. epilepsy), and escalations that were neither type A or B were classified as type C.

2.6. ED Study Results

From the 476 patients in the ED study, there were 284 escalation events in total as shown in Table 2.3. Of these, 193 were Type-1 events that occurred upon arrival, and the remaining 91 represented events that occurred due to patient deterioration after arrival. For the sake of this analysis, only the first escalation for each patient was considered for the reasons given in Section 2.5. The first escalations included all of the Type-1 events, and an additional 14 Type-2 events, where the patient deteriorated after arrival and had no previous Type-1 escalations. In total, 207 first escalation events were analysed.

The first escalations were also classified according to whether they were caused by physiological or neurological issues (type A and B escalations), or non-physiological causes (type C). While the analysis shows results for both of these groups, the escalations were partitioned in this way as the T&T criteria are designed to detect type A and B events, which we call "physiological escalations", only.

2.6.1. Analysis of Initial Escalations

Table 2.4 shows how well the T&T system, as used in the ED, detected initial escalations. Within this analysis framework, an initial escalation was only considered to have been detected (i.e. a true positive) if the T&T score exceeded the triggering criteria for an observation within a ± 10 minute window centred at the time of the escalation. Where there was any ambiguity, the patient notes were used to help determine whether the T&T score was directly related to the escalation event. If the T&T score was below the threshold, it was classified as a FN. If the T&T score was not worked out by the nurse,

	A+B escalations	C escalations	No escalations	Total
T&T was completed and met triggering criteria (at the time of escalation when one occurred)	28	1	22	51
T&T was completed at some point, but did not meet triggering criteria (at the time of the escalation when one occurred)	32	63	141	236
T&T never completed	44	39	106	189
Total	104	103	269	476

Table 2.4.: Overview of the manually-observed T&T system's ability to detect first-time escalation events.

the resulting classification was not included in the sensitivity and specificity calculations.

Table 2.4 shows that 28 out of 104 physiological escalations were detected by the T&T system. In 32 cases of physiological escalation, the T&T scores did not meet the triggering threshold or else the T&T score was not worked out at the time of the escalation. In contrast, only one non-physiological escalation was detected using the T&T system, as one may expect from using a system based upon physiological parameters. In this case, the patient had been moved to Resus because of chest pain, but happened to have an unrelated high systolic blood pressure.

Twenty-two patients exceeded T&T thresholds but were never escalated during their stay (false positives). One of these, Patient ED00277, exceeded the T&T threshold because the T&T parameter scores had been added incorrectly. Twelve of the patients triggered due to hypertension. These were reviewed by the senior clinical team mentioned in Section 2.1. In each case, all three team-members agreed that the hypertensions were not of clinical significance and would not have changed clinical management at the time. Typically, hypertension is ignored in the ED as it is often not relevant to the presenting complaint. More often than not, the hypertension will be a chronic condition that has no associated signs or symptoms, but may increase a patient's risk to cardiovascular disease. The remaining nine patients triggered due to hypotension, tachycardia, bradypnoea, or

	A+B escs	C escs	no escs	total
T&T met triggering criteria (at the time of escalation when one occurred)	98	12	80	190
T&T did not meet triggering criteria (at time of escalation when one occurred)	6	91	189	286
Total	104	103	269	476

Table 2.5.: Overview of retrospective T&T's ability to detect of first-time escalation events.

combinations of vital sign abnormalities. In each case, a review from one member of the clinical team confirmed that no change in treatment would have been advised in each case. In two cases, a doctor was present with the patient at the time of the observation. In one other case, an unusually low RR was attributed to drugs given for pain relief. Three cases of hypotension were borderline, and not considered clinically important. The remaining three cases included vital signs that may have warranted a doctor's review, but the patient was not felt to have been at risk according to retrospective analysis by the senior clinical team.

As we mentioned in Section 2.6.1, the T&T system is only designed to detect physiological abnormality. Therefore, the sensitivity and specificity analysis should only be calculated for the physiological escalations (types A1, A2, B1 and B2). The sensitivity and specificity of T&T scores for detecting physiological escalations for all patients who had the T&T process completed during their stay is:

$$Sensitivity = \frac{28}{28+32} = 0.47, \qquad Specificity = \frac{141}{141+22} = 0.87 \tag{2.6.1}$$

In comparison, the effectiveness of an error-free retrospective T&T (as defined at the start of this chapter) is shown in Table 2.5. In this case, only six patients had physiological events without a corresponding T&T Trigger (false negatives) in comparison to 32 for the nurses' T&T. Of these, four did exceed the T&T thresholds, but the observations had been written down in the patient notes rather than on the observation chart. The fact that patient abnormality was recorded in the notes indicates that these patients were managed

appropriately. One other patient may also have had observations that met the T&T triggering criteria. However, the observation chart was difficult to interpret, and it was unclear whether the time of the relevant observation had been changed retrospectively to match the escalation. The remaining patient was not observed at the time of their escalation, and instead nursing staff were called back to the patient by their relative due to a decrease in the patient's consciousness. The notes state the appropriate clinical action was taken thereafter.

80 patients had a triggering score, but did not escalate at any time during their stay (false positives). 38 of the 80 Triggers were due to cases of isolated systolic hypertension, which were again deemed to be clinically irrelevant. Of the other potentially "missed" alerts, 11 were for hypoxia, 9 for tachycardia, 5 for tachypnoea, 2 for bradypnoea, 4 for hypotension, 1 for pyrexia, and 10 were triggered by a combination of two or more vital sign abnormalities. The clinical notes for the 42 patients without hypertension were reviewed by the clinical team. For 38 of the 42, it would have been appropriate to document that clinicians had been informed of the abnormal vital signs, but it is unlikely that any change in clinical management would have occurred, Of the remaining four, one had bradypnoea that settled by the time of subsequent observations; one had hypoxia, possibly due to undocumented COPD and was later discharged home; one had pyrexia that could have been assessed more promptly than was documented, and another had hypoxia from sepsis and it is unclear when treatment was commenced.

We further note that abnormally low SpO_2 is often caused by the common chronic condition of COPD, and cannot always be treated with oxygen therapy. In nine of the cases here, each alert was deemed to have been appropriately ignored due to pre-existing COPD.

The sensitivity and specificity of error-free T&T, using physiological escalations as the outcome marker are:

$$Sensitivity = \frac{98}{98+6} = 0.94, \qquad Specificity = \frac{189}{189+80} = 0.70 \tag{2.6.2}$$

2.7. Discussion

2.7.1. Observation and T&T Completeness

T&T effectiveness will be firstly limited by the frequency of observation and secondly by how often the T&T total is calculated for each set of observations. In our study, observations were taken frequently and regularly, with a mean time of one hour between observations. Furthermore, 99% of patients had at least one full set of observations (HR, RR, BP, and SpO2 measurements). Section 2.3 clarifies why temperature and GCS were not included in the full set.

In contrast, T&T totals were calculated for only 34.3% of the observations. The poor T&T completion rate can be attributed to two causes. Firstly, the effect of clinical pressures may result in T&T scores not being calculated for the most ill patients because immediate treatment is required, and because deterioration has already been identified without use of the score. Additional results from this study provide evidence for this hypothesis. By calculating the T&T completion rate in each of the ED areas, we observed that T&T observations were completed in 33.1% of cases in the most acute area, Resus, in comparison to 38.5% and 37.0% completion rates for Majors and CDU respectively.

Secondly, poor T&T completion may also be due to staff tiredness. Figure 2.3.4 showed that T&T completion rate increased during day time hours, and also during changes of shift. This can be addressed by regular training, and results from the study show a gradual increase in T&T completeness over the course of the study, suggesting that this issue was adequately addressed (see Figure 2.3.3).

2.7.2. T&T Score Errors

In addition to the poor completion rate, 20.3% of the observations had incorrect T&T overall scores. While the error rate may seem high, it compares favourably with results reported by Prytherch et al. in a controlled setting [89].

The vast majority of errors were due to incorrect assignment of individual vital sign parameter scores. In a few cases, the incorrect assignment appeared to have been caused by simple human error. For instance, one patient was scored 0 for a RR of 19 respi-

rations/min, when it should have scored 1. Mistakes such as these may be rectified by staff training, or by improving the layout of the T&T charts. The likelihood of making incorrect assignment errors is also increased by the fact that T&T criteria vary between hospitals, and even from ward to ward. Subsequent to the completion of this study, the T&T charts have been redesigned to improve the ease of assigning scores. The new chart is shown in Appendix A.

In many other cases, vital sign parameters appeared to be deliberately mis-scored as nursing staff used their clinical judgement to over-rule the T&T criteria. In the majority of these cases, the T&T parameter scores for SpO_2 or for high SBP were scored as a zero to prevent an unnecessary call-out. As mentioned previously, these are usually indications of chronic conditions that have no relevance to the presenting complaint. Adjusting SBP and SpO_2 parameter scores from three to zero explains why there were sometimes large discrepancies between the recorded and error-free T&T scores (see Figure 2.4.1).

Incorrect T&T totals were also caused by errors in arithmetic, though these were uncommon, accounting for only 1% of observations. Arithmetic mistakes tended to be compounded by additional errors which occurred when previous results were simply copied without repeating the addition. Arithmetic errors seem to occur more frequently during night shifts, with six out of eight of the initially incorrect GCS totals (that is, ignoring any copying errors) occurring during the night. However, the small sample means that this result is not statistically significant.

Our continuous measurement of skin temperature was another potential source of error unique to this particular study. By analysing the manually recorded core temperature, and the continuously monitored skin temperature, it was shown that at least 29% of temperature data were likely to have been recorded from the monitor, despite the fact that a cursory look at the skin temperature data would show it to be well outside the normal range for core body temperature. In addition to quantifying the effect of the temperature errors, this result also raises wider issues that go beyond the scope of this thesis. This includes the question of whether nurses place too much confidence in the reliability of bedside monitors at the expense of making physical contact with the patient.

2.7.3. Effectiveness of Track and Trigger

While a 20% error rate for T&T scores may seem high, the result should not be interpreted outside its clinical context. Whether the mistakes in T&T had an effect on patient outcome is the more relevant question.

Section 2.6.1 showed the retrospective, error-free, T&T scores had a much higher sensitivity than the nurse-recorded T&T, and thus detected a greater proportion of physiological events. The retrospective T&T scores detected all but six physiological escalations, of which only one would have been missed if observations had been documented accurately on the observation charts. In contrast, manual T&T calculation resulted in 32 missed physiological escalations.

However, the error-free T&T score was also far less specific than nurse-recorded T&T, which means that it provided a greater number of false alerts. False alerts are particularly troublesome in clinical settings, as studies using audible single-channel alerts have shown that false alerts lead to clinicians learning to ignore the alerts [115]. In order to be truly viable in the ED, the specificity of the retrospective T&T should be at least comparable to the manually recorded T&T.

2.8. Conclusion

In Chapter 1, we described how vital sign measurements were important for determining the condition of a patient, and also described how Early Warning Scores could be used to analyse the data. The ways in which continuous monitoring could be used to enhance patient care were also reviewed. The introduction of both continuous monitoring and T&T systems into the ED are steps towards attempting to identify unwell or deteriorating patients early.

The analysis in this chapter allows us to assess how well the T&T system has been adopted. We have shown that while vital sign observations were taken regularly by nursing staff, T&T scores were calculated infrequently. By using an error-free retrospective T&T score, we were able to highlight that T&T scores were incorrect 20% of the time. We also identified a number of reasons why T&T scores were often erroneous, showing that

the most common problem was an inability to convert vital signs accurately into their corresponding T&T parameter scores. This matches the results reported by Edwards et al. [29], who conducted a study over four medical and surgical wards, and reported that 69.7% of MEWS score errors were due to incorrect score assignment.

The clinical significance of the 20% error rate was evaluated by comparing how well T&T would detect clinically-validated escalation events in comparison to a computer-generated, retrospective T&T score. It was shown that the error-free score minimised the chances of missing patient deterioration at the cost of providing many more false alerts. For manually-scored T&T, there were 22 instances for which the T&T threshold was exceeded, but with no recorded escalation events, whereas there were 80 such instances using the retrospective T&T score.

In conclusion, this chapter has shown the current limitations of manually-scored T&T, and demonstrated that a retrospective computer-generated score increases the effectiveness of T&T at detecting physiological escalations. This increased sensitivity is at the cost of reduced specificity, leading to more false alerts. It should be possible to create an automatic scoring system that emulates our retrospective T&T score. In fact, one such product exists: the VitalPAC^{TM} system, for which nurses enter the values of vital sign observations into a handheld computer and the T&T score is computed immediately. The VitalPAC^{TM} system has so far only been studied in standard ward settings [**?**].

Until this point, the continuous vital sign data that were collected during the study has not been analysed. In the following chapters, we will investigate whether it is possible, through the use of the continuous data, to increase sensitivity to escalation events while maintaining a high specificity. We will begin by investigating how well the T&T criteria perform when applied directly to the continuous data. Following this, we will investigate other methods of combining the continuous vital sign data so that they can correctly identify physiological escalation events. Ultimately, we aim to show that models based on continuous vital sign data provide substantial benefits compared to only using a T&T system based on intermittent observations.

3. Continuous Monitoring with Track and Trigger Criteria

In Chapter 2, we demonstrated that Track and Trigger (T&T) was useful for identifying escalation events and consequently patient deterioration. We also showed that there were serious practical issues in implementing T&T scoring systems within the Emergency Department (ED), including low T&T completion rates, and a significant error rate which limited the system's sensitivity. We concluded that a computer-assisted T&T system may help to improve the standard of care within the ED.

Even if T&T were to be perfectly implemented, however, the effectiveness of the system would be limited by the low frequency of patient observation. Consider the example in Figure 3.0.1, which shows vital sign data collected continuously from a bedside monitor for a patient in a Step-Down Unit at the University of Pittsburgh Medical Centre (UPMC) [49]. Let us assume a perfect T&T system, in which patients are observed at intervals of 60 minutes, and that one such observation was taken at 15:30. The continuous vital sign data shows that the patient appeared to be stable at this time. All of the vital signs were within normal limits apart from Systolic BP, which had been elevated throughout the patient's stay.

A second observation would then be taken at 16:30. At this instant, the vital signs appear to be normal. There is a slight decrease in oxygen saturation compared to the previous observation, but no noticeable difference in the other vital signs. The figure clearly shows that the patient suffers a sudden deterioration between the two hypothetical observations, at 16:10, at which point both HR and oxygen saturation become abnormal, increasing to 95 beats/min and decreasing to 84%, respectively. Similar events occur at 14:20 and 18:00. We can be confident that these sudden events are not artefactual because

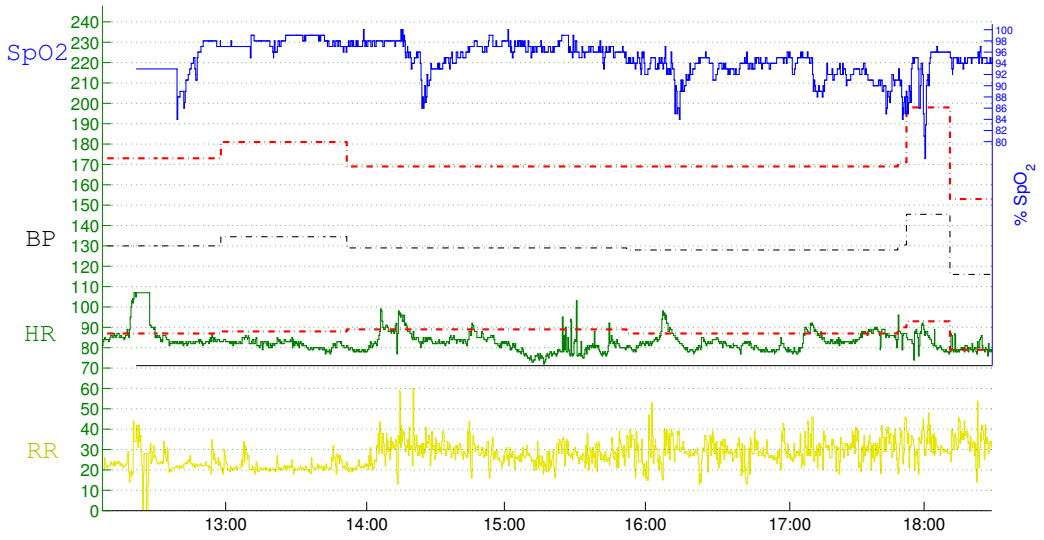

Figure 3.0.1.: An example vital sign record from a study conducted at a UPMC Step-Down Unit. The RR, HR, and SpO_2 are recorded in yellow, green and blue respectively. The Systolic BP and Diastolic BP are indicated by the upper and lower red lines, while their arithmetic mean is shown as a dashed black line. The record indicates that the patient slowly deteriorates between 12:00 and 18:00, but suffers significant short-term deterioration at 14:20, 16:10, and 18:00 (oxygen desaturation and increased heart rate in each case). Between these events, there is partial recovery due to homeostasis.

abnormalities occur simultaneously for two independently collected vital signs. Furthermore, the short-term events are consistent with the long-term gradual deterioration that can be seen during the six hours of monitoring. For example, we observe an increase in respiratory rate and respiratory rate variability over the course of the recording. Continuous monitoring is able to identify short-term deteriorations, which may be missed if the patient is intermittently observed.

Even if short-term patient deterioration is not missed by intermittent observations, continuous monitoring allows deteriorations to be detected in real-time, thus providing early warning with respect to T&T. Let us consider Figure 3.0.1 once more, and this time assume that observations were instead taken at 17:00 and 18:00. In this case, a T&T system would detect the deterioration at 18:00 associated with a low oxygen saturation of 85%. An effective continuous monitoring system would have detected the decrease to 90% of oxygen saturation at 17:15, potentially providing 45 minutes of early warning.

A method that uses the continuous data recorded by the bedside monitors may enable real-time detection of short-term deterioration and facilitate prompt interventions. In

this chapter, we conduct an initial investigation, based on the assumption that the T&T criteria, that were first described in Chapter 1, can be applied to continuous data. We derive a continuous T&T model and then further T&T models which emulate manual observations taken at 15, 30 and 60 minute intervals. We evaluate their performance using the known escalation events described previously as outcome markers, within the sensitivity and specificity framework described in Chapter 2.5.

3.1. Method

3.1.1. Continuous Track and Trigger

The T&T criteria of Table 1.2 defined in Chapter 1 were used to develop a continuous T&T system. The system was designed to emulate nursing practice as closely as possible, which necessitated the modifications described below.

During the study described in Chapter 2, we observed that nurses use the values displayed by the bedside monitors when these are available. In order to ensure a meaningful reading, an experienced nurse will wait to see if there are any momentary fluctuations before recording the "most appropriate" value, thus reducing the influence of any artefacts. To simulate this effect in the continuous model, the median of a one-minute window of each vital sign parameter was used when calculating the T&T score, as shown in Figure 3.1.1. With non-overlapping windows, scores are generated every minute.

The one-minute T&T scores were also sampled at 15, 30 and 60 minute intervals, with a zero time-offset from the start of the record. These sampled scores simulate the effect of recording intermittent observations at regular intervals. Practically, this is the same as assuming that an initial observation would be made as soon as the patient is assigned to a bed, which conforms with our knowledge of standard ED practice. The four T&T models allow us to evaluate the effect of changing the frequency of T&T observations.

The continuous T&T system was tested on the continuous data set recorded from the 476 ED study patients. Only the variables recorded by the bedside monitor can be used in the continuous scoring system. Hence, measurements of Glasgow Coma Score (GCS), pupil dilation, or urine output are not taken into account, and so their associated T&T

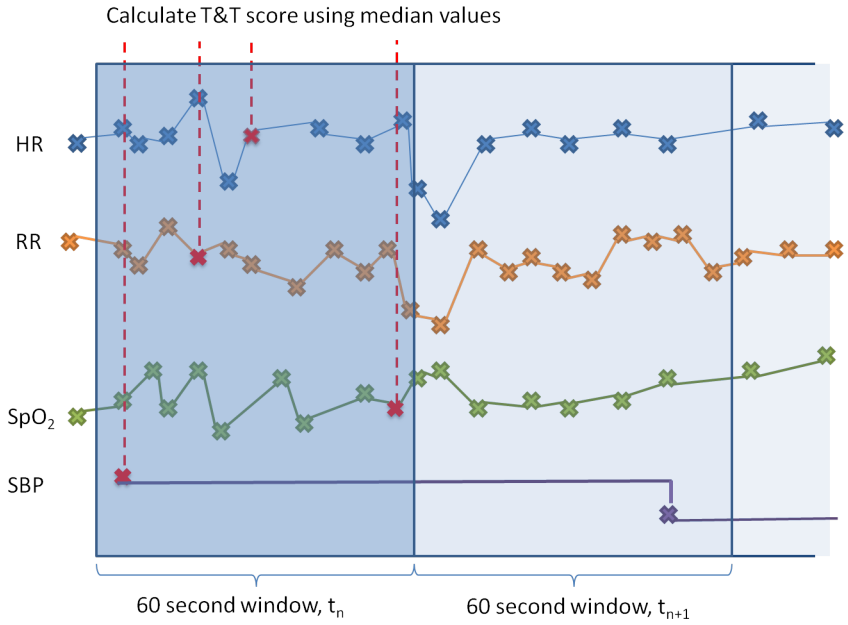

Figure 3.1.1.: Computation of Continuous T&T. The median value is calculated for each of the vital sign channels (red crosses), and then used to compute the continuous T&T score at time t_n using the criteria listed in Table 1.2. The next Track and Trigger score is calculated at time $t_{n+1} = t_n + 60\,\text{sec}$.

scores are always set to zero. Of the five continuous variables collected during the study, only four, HR, RR, SpO_2, and SysBP were used. Temperature was not included, as the bedside monitor recorded skin temperature values, whilst the T&T criteria require the use of core temperature (see Section 2.1).

3.1.2. Analysis Plan for Continuous T&T System

The performance of the continuous T&T System was assessed using the escalation events described in Chapter 2. Using the analysis framework developed in Section 2.5, we again consider only the first escalation, which is classified as either "non-physiological" or "physiological". In doing this, we allow for a direct comparison with the retrospective and nurse-recorded T&T results reported in the previous chapter. The escalation was considered to have been detected, and therefore a True Positive, if the continuous T&T score met the alerting criteria (T&T scores of 3 for a single vital sign, or an overall score of 4 or more for a combination of vital signs) within a window of $t = 10$ minutes before an escalation event and $\tau = 10$ minutes after an escalation event, and a False Negative otherwise. Using the same framework, we generate False Positives (and True Negatives)

by considering only the patients for whom there were no escalations. The resulting sensitivity and specificity are calculated using all first escalations, in order to make a direct comparison with the previous results.

In Chapter 2, we reported that a large proportion of escalations, 198/289 occurred on arrival to the ED and were documented during triage rather than during the patients' stay in one of the ED areas. At this stage, the patient is not connected to a bedside monitor, these escalations are not detected using T&T. Hence, in the second half of the analysis we will apply our framework to the first physiological escalations that occurred after arrival (type 2 escalations). Furthermore, we recognize that the length of the window was set arbitrarily at $t = 10$, $\tau = 10$ minutes, however, other window lengths may be more appropriate.

Rather than trying to optimise the length of the window, in the second half of the analysis we also report the number of True Positives for a range of window lengths. The window length is adjusted in such a way that the window always ends $\tau = 10$ minutes after the escalation to account for the differences in time-keeping methods, but the time before the escalation event is varied between $t = 1$ to $t = 60$ minutes to allow for the possibility of early detection of deterioration.

In summary, we use the following outcome measures:

- False Positives (on a per patient basis for all first escalations initially, then on a per patient basis for the first A2 or B2 escalations)
- True Positives (on a per patient basis for all first escalations initially, then on a per patient basis for the first A2 or B2 escalations)
- Alerts per patient (for those with A2 or B2 escalations and those with no escalations)
- Alert durations (for those with A2 or B2 escalations and those with no escalations)

In addition to these metrics, we will also report the distributions of the number of alerts, and the duration of alerts for patients with physiological escalations and for patients with no escalations.

3.2. Results

3.2.1. Continuous Data Loss

476 ED study patients were eligible for data collection, but continuous vital sign data were only available for 402 of them. The length of stay for each patient was calculated from their admission and discharge times, as recorded in the patient notes. Patients stayed in the ED for 2,170 hours in total, with a mean length of stay of 5.39 hours. It is probable that the length of stay is an overestimate, as patients' documented times of departure were likely to have been filled in retrospectively, after the patients had left the department.

In total, 1708.4 hours of continuous vital sign data were collected. The patients with continuous data often had gaps during which no vital signs were recorded. There are a number of possible causes for this, but the most likely is that data could not be transmitted during certain medical interventions, such as the occasions on which patients were moved from one bed to another or taken for scans.

We can quantify the average data loss per patient as:

$$data\,loss = \frac{length\,of\,stay - length\,of\,record}{length\,of\,stay} \tag{3.2.1}$$

To compute this, the length of each vital sign record was estimated by calculating the period for which at least one vital sign was being monitored. For all of the vital sign channels, apart from blood pressure, measurements were made every 30 seconds (Blood pressure is only measured intermittently, with several minutes between each measurement.)

For the purpose of estimating the length of the vital sign record, each vital sign recording was subjected to a zero-order hold of length 30 seconds. The length of the total vital sign record is then simply calculated from the union of the individual vital sign records.

The ratio of length of record to length of stay for the 402 study patients with continuous data is shown in Figure 3.2.1. The mean value of this ratio (expressed as a percentage) is 79%, indicating that the mean data loss was 21%. This data loss estimate is likely to be an overestimate, as the length of stay is over-estimated.

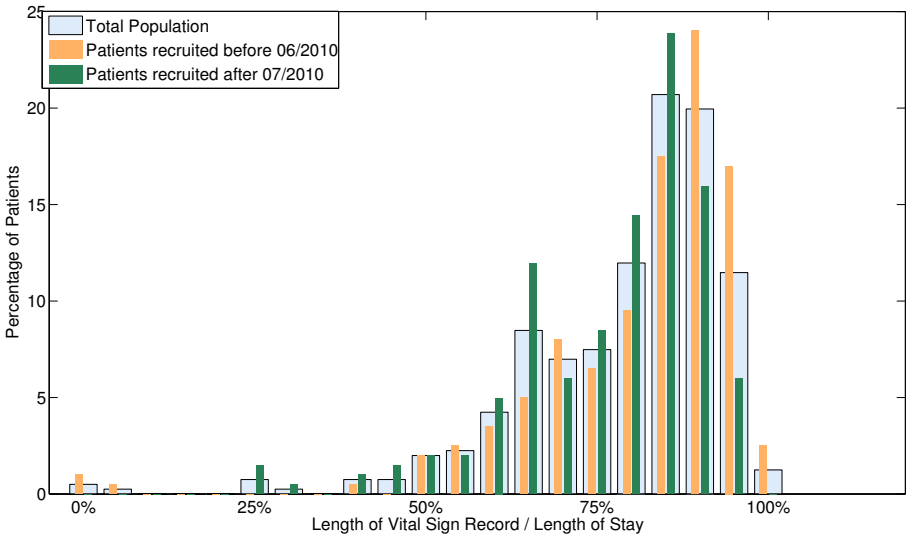

Figure 3.2.1.: The *length of vital sign record / length of stay* ratio for the 402 study patients that had continuously recorded data is shown in blue. The ratios for the first and second half of the study are shown in orange and green respectively, and are broadly similar in shape.

In Chapter 2, it was shown that the completeness of the nurses' observations improved markedly during the study. Unlike manual observations, continuous recording requires very little human intervention and we would therefore expect no differences between the two halves of the study. The distribution of the *length of record/length of stay* ratios in Figure 3.2.1 is relatively flat as expected, with a non-significant decrease in the ratio (from 81% to 77%) between the first and second halves of the study.

In addition to data drop-out on all channels, individual vital signs were absent for periods of time due to probe disconnection. The reason for this type of data loss may be attributed to practical difficulties such as chest electrodes losing their adhesion, or SpO_2 probes being removed by the patient due to discomfort. The extent of this problem can be quantified by calculating the overall recording time for each vital sign, and the percentage data loss as shown in Table 3.1. As an aside, we note that the data loss for temperature, which was not considered for continuous T&T, is 74.0%, far exceeding the data loss for the other channels.

The similarity in percentage data loss values for each vital sign suggests that there is no obvious problem with any one vital sign probe. Given that at least one vital sign is present 78% of the time, and that each vital sign has completion rates of approximately

Vital Sign	HR	RR	SpO_2	SBP
total time in hours	1645	1629	1664	1776
% data loss	24.2	24.9	23.3	18.2

Table 3.1.: Total time of vital signs recorded and percentage data loss for each channel of vital sign data

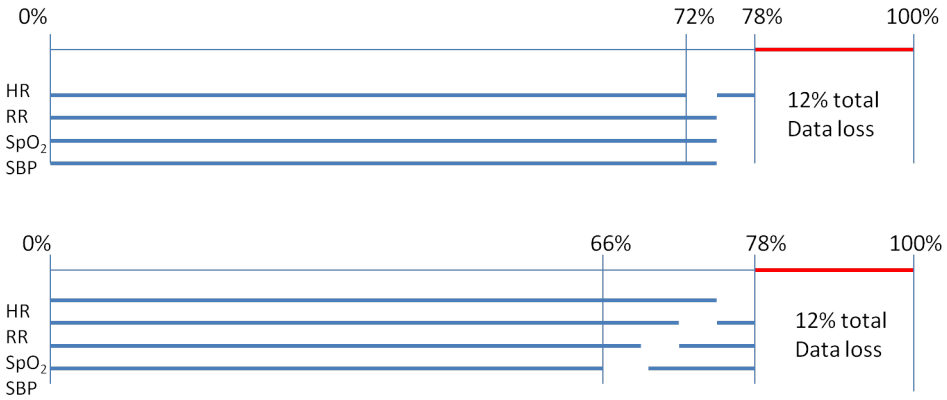

Figure 3.2.2.: Best and worst case scenarios for single channel data loss. In each case, total loss of data occurs 12% of the time, and each vital sign channel contains data for approximately 75% of the time. The upper figure shows the best case, in which 6% of data is affected, and the lower figure shows the worst case, when 12% of data is affected

75% (Table 3.1), we can also conclude that single-channel drop-out is a minor issue. At best, only 6% of the data were affected, and at worst, the single channel drop-out may have had an effect on 12% of the data if data loss on each channel occurred out at different times. In comparison, data loss over all channels occurred 22% of the time.

The distributions of values for each of the four vital signs are shown in Figure 3.2.3, where the red lines indicate the single-channel T&T alerting criteria for each vital sign. The distributions for the whole study population, for the patients with physiological escalations (i.e. A or B escalations), and for those with no escalations have been plotted in blue, grey and red respectively. The distributions for the A/B-type escalations group differ significantly from the distributions corresponding to the other two groups. In particular, the median HR is higher and the median SBP is lower in the A/B escalation group. In addition to this, all of the vital signs distributions from the A/B-type escalation group have longer tails than the other groups. For instance, consider the SpO_2 distribution for the A/B escalation group, which has a flatter distribution with long tails. In comparison to the distributions from the other groups, this distribution has a lower modal value, and

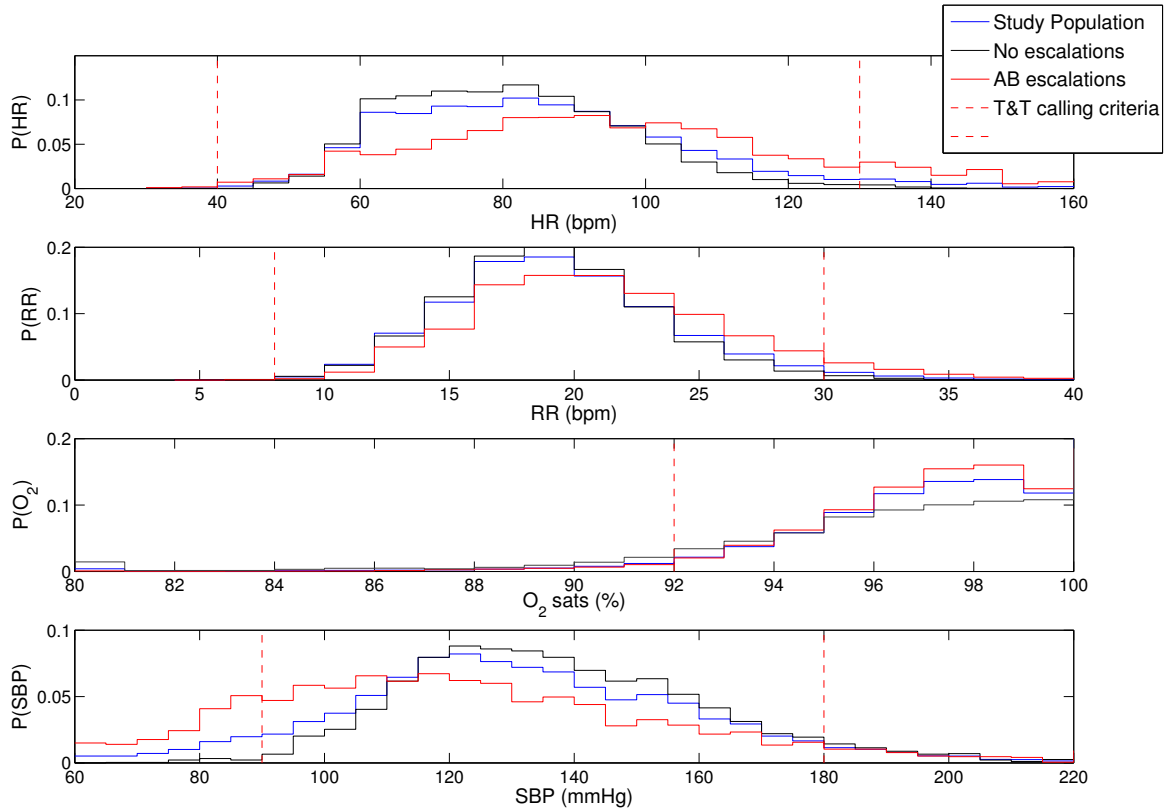

Figure 3.2.3.: Estimates of the discrete pdfs for each of the vital signs calculated by normalising the vital sign histograms for the ED study patients. Estimated pdfs for the sub-populations with *No Escalations*, and *Physiological (A or B) Escalations* are in red and grey respectively.

higher probabilities for all SpO_2 values lower than 94%. The presence of long tails means that there is a relatively high probability that the vital sign values will cause the T&T thresholds to be exceeded. Overall, the vital sign distributions confirm that patients with A/B escalations are physiologically abnormal, particularly with respect to hypotension (low blood pressure) and tachycardia (high heart rate).

3.2.2. Detection of Escalations

The effectiveness of continuous T&T at detecting physiological and non-physiological escalations within 10 minutes of the escalation is shown in Table 3.2, using the same format as in the analysis of Chapter 2 (Table 2.5). All patients without continuous data were assigned to the second row in Table 3.2 to allow comparison to previous results. The accuracy of the result is dependent on the percentage data loss during the ±10-minute window, which is shown in Figure 3.2.4. The graph indicates that a large proportion of

	A/B escs	C escs	no escs	Total
continuous T&T score above alerting criteria*	61	20	168	249
continuous T&T score below alerting criteria*	43	83	101	227
Total	104	103	269	476

*at the time of the escalation if one occurred, and within the test window

Table 3.2.: Initial escalations detected using continuous T&T within a window with $t = 10$ before the escalation and τ=10 minutes after the escalation

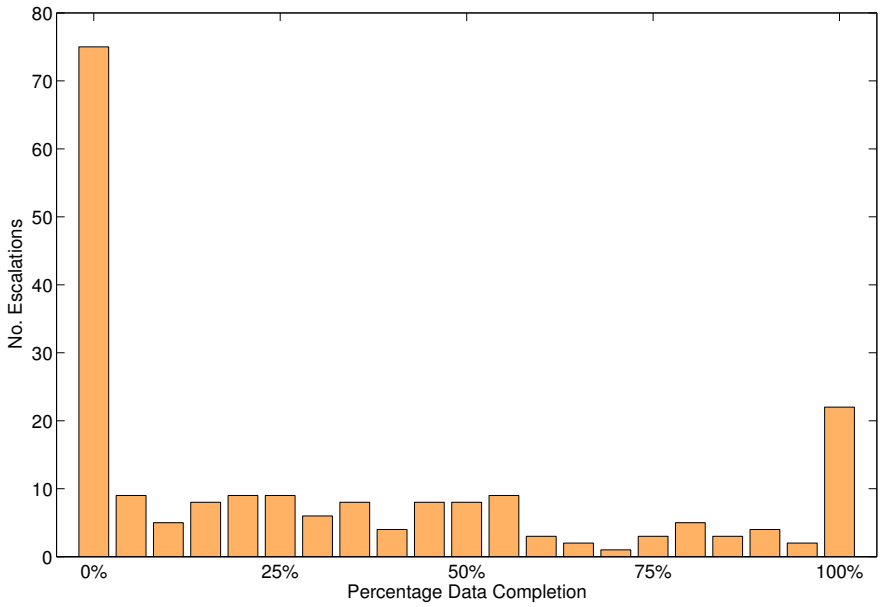

Figure 3.2.4.: Percentage of continuous data available during the ±10 minute window for each of the 207 first escalations

the patients who had escalations had high data loss at the time of their escalation. 61 patients had an initial escalation that would have met the alerting criteria for continuous T&T if the T&T score had been computed continuously. 168 patients had continuous T&T scores that met the alerting criteria, but did not have an escalation during their stay. The resulting sensitivity and specificity are 59% and 38% respectively, which does not compare favourably with retrospective T&T.

As indicated in Section 3.1.2, only the physiological (A2 or B2) escalation events that occurred after arrival to the ED should be considered. We previously reported in Section 2.6 (Table 2.3) that there were 64 such events. Of these, one did not have a time recorded, 4 occurred for patients that had no continuous vital sign data available, 13 were due to

low GCS, and one other escalation was caused by an abnormal temperature. In addition, there were 3 additional escalations (pertaining to patients ED00502, ED00354) for which no vital sign data were recorded during the 60 minutes or 10 minutes after the escalation event time, although vital sign data were available at other times during the patients' stays. This reduced the maximum number of events that could be detected by continuous T&T to 43 events. These escalation events occurred in 29 patients, as some of the patients had more than one physiological escalation event.

The number of escalation events detected by continuous T&T (i.e. the true positive rate), and the sampled T&T scores, $T\&T_{15,30,60}$, are shown in Figure 3.2.5 for window lengths varying between $t = 1$ and $t = 60$ minutes. An escalation event was deemed to have been detected when the initial physiological escalation was detected within the test window. The figure also shows the number of escalation events which would have been detected, that is, the number of True Positives, on a per-patient basis.

The majority of escalations were detected within a 10-minute window length, regardless of the frequency of T&T observations. With a $\pm$10-minute window, the continuous T&T system detects 36 out of 43 escalation events (see Figure 3.2.5). In comparison, the sampled T&T scoring systems perform worse, with $T\&T_{15}$ (i.e. intermittent observations every 15 minutes) detecting 25 out of 43 escalation events and $T\&T_{30}$ and $T\&T_{60}$ detecting 16 and 10 escalation events respectively.

By considering the same data on a per patient basis, Figure 3.2.5 indicates that 24 out of 29 patients would have had their escalations of care identified by continuous T&T in comparison to 14, 8, and 5 patients for the $T\&T_{15}$, $T\&T_{30}$ and $T\&T_{60}$ systems respectively.

At the maximum window length considered, 60 minutes, continuous T&T would have detected 40 escalation events and 26 out of 29 patients. In comparison, the next best system, $T\&T_{15}$, would have detected only 32 escalation events. Although the system's detection performance improves marginally for at the longer window lengths, it is debatable whether or not the vital sign and the event are linked for window lengths of 60 minutes.

Table 3.2 showed that there were 269 patients who did not have any escalation events. These will be True Negatives if the continuous T&T score did not meet the T&T calling criteria at any time during the patient's stay, and a False Positive otherwise. The same

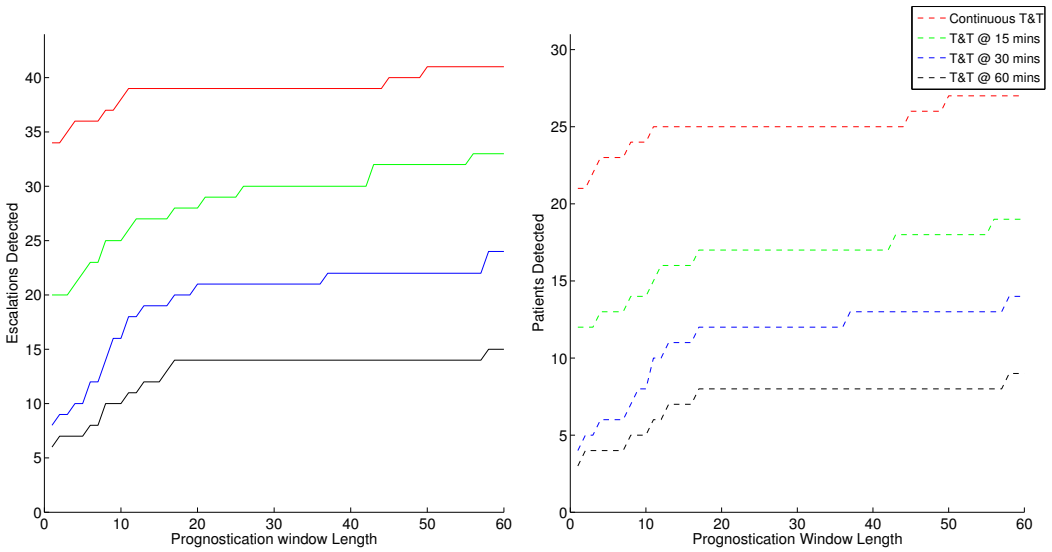

Figure 3.2.5.: *Left*: Number of correctly identified A2 and B2 escalation events (true positives) over a range of window lengths, *Right:* Number of correctly identified first A2 and B2 escalations (True Positives on a per patient basis)

	Continuous T&T	$T\&T_{15}$	$T\&T_{30}$	$T\&T_{60}$
True Negative (Zero Alerts)	101 (49)	168 (116)	183 (131)	200 (148)
False Positive ($\geq$ 1 Alerts)	168	101	86	69

Table 3.3.: Summary of the true negative/false positive rate for patients with no escalation events

analysis was undertaken using the sampled T&T scoring systems ($T\&T_{15,30,60}$) and the results are listed in Table 3.3. We note that the 269 patients include those who did not have any continuous data. The corresponding figures for those with continuous data are shown in brackets.

The per-patient False Positive shown in by Table 3.2 is a slight oversimplification. In a clinical setting, the importance of a False Positive depends both on how often and for what duration the T&T score meets the calling criteria. For instance, long-term false alerts that require intervention may add significantly to the nursing workload, whereas transient false alerts are likely to have resolved themselves before staff can respond. We can assess this effect by first calculating the number of occasions on which the T&T alert criteria were met per patient, when using the continuous T&T score, as shown in Figure 3.2.6. Similar distributions are also plotted for the patients who had physiological escalations, and the patients who had no escalations.

The two sub-populations have markedly different distributions. The no-escalation group

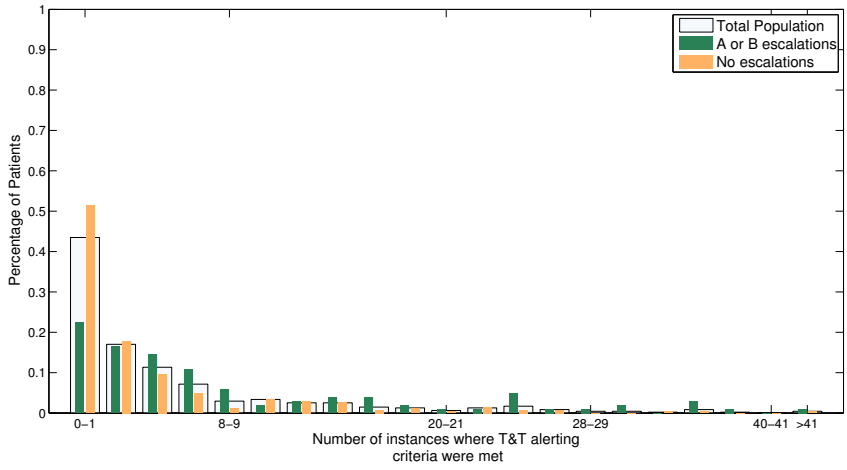

Figure 3.2.6.: Number of alerts per patient when using the continuous T&T system. The plot in white shows the distribution for the entire population, while the plots in orange and green show the alert distribution for patients with physiological (A or B) escalations and no escalations respectively.

contains 78.8% of its data mass between 0 and 5 alerts and has a median of one alert. In comparison, the group of patients with physiological escalations contains 53.4% of its data mass between 0 and 5 alerts, and has a median of five alerts. We can compare the two distributions using the Mann-Whitney rank-sum test, which tests whether their medians differ by a statistically significant amount. The advantage of this test over other tests such as the Student T-test, is that it does not assume any knowledge about the shape of the distribution. The two sets of data are ranked by value. The test statistic, U, is then given by:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{3.2.2}$$

where R_1 is the sum of the ranks for the first set of data, which contains $n_1$elements. U_1thus represents the difference between the actual rank-sum and the expected rank-sum. To gauge whether U is statistically significant, we assume that U is normally distributed for large samples, so that the normal approximation for the Mann-Whitney test can be used. The standardised test statistic is calculated as:

$$z = \frac{U - m_u}{\sigma_u} \tag{3.2.3}$$

where m_u and σ_u, the mean and standard deviation of U, are:

$$m_u = \frac{n_1 n_2}{2} \tag{3.2.4}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{3.2.5}$$

Applying this to our data set indicates that the medians of the two groups are significantly different ($p = 1.4 \times 10^{-9}$). Although the no-escalation group meets the T&T alerting criteria far less frequently on average, there is nevertheless a wide range, and 37 (13.8%) of these 269 patients exceeded the alerting criteria more than 10 times during their stay in the ED.

Using the information required to generate Figure 3.2.6, we can also calculate an "alert rate" by dividing the total bed-time by the number of alerts (i.e. the number of occasions on which the continuous T&T score is greater than the alert criteria for a particular patient group). For the entire study population, 1708.4 hours of data were recorded, and the continuous T&T system would have generated 2503 alerts. This gives an estimate of 1.47 alerts/hour per bed. By considering only the patients with no alerts, and assuming that all alerts for these patients are incorrect, we can estimate an underlying *false* alert rate on the same basis. 1120 false alerts during 1156.7 hours of data gives an estimate of 0.97 false alerts/hour per bed. In comparison, the alert rate for patients that had physiological escalations is 1.77 alerts/hour per bed.

The impact of false alerts can be assessed by assuming that false alerts that persist for a short time are likely to be less problematic than those that sustain for a long period. We therefore calculated, for each patient, the time that the T&T alerting criteria was met as a percentage of the total vital sign recording time (Figure 3.2.7). As expected, patients with physiological escalations had longer alerts on average than patients with no escalations, and their median alert lengths were 36.9% and 5.6% of the vital sign record length, respectively.

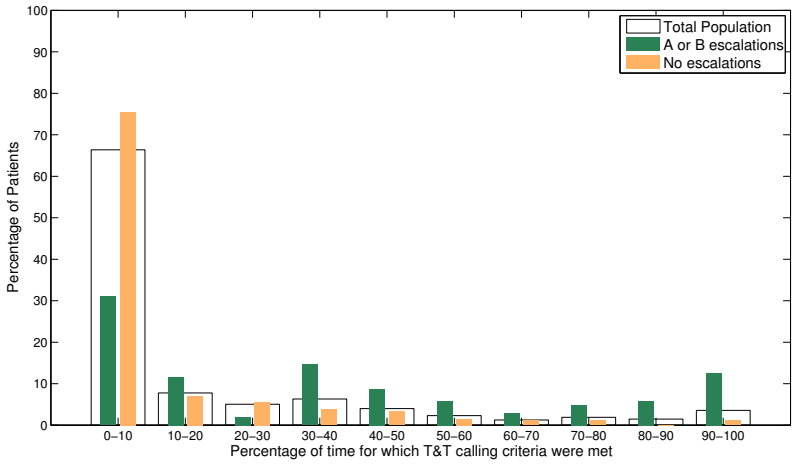

Figure 3.2.7.: Time spent in the alert state as a percentage of the total time on the ward, for the whole population, and the sub-populations with *No Escalations*, and *Physiological (A or B) escalations*.

3.3. Discussion

Our results show that the overall level of data collection was acceptable, with a 21% upper-bound estimate on percentage data loss per patient. Technical problems accounted for the vast majority of the 74 patients with no recorded data, though in a small number of cases, continuous monitoring was not considered clinically appropriate; for instance, if a patient was unable to tolerate monitoring. The exact number of these patients could not be obtained from the ED data set due to the brevity of the written medical notes.

The technical issues included power outages that caused the data collection server to shut down (despite using an uninterrupted power supply unit), and periods during which the hospital's data server was switched off due to overheating. These issues were largely beyond our control, and did not affect whether or not the continuous vital sign data were displayed on the bedside monitors.

The data loss for each vital sign channel was summarised in Table 3.1. During the study period, we received anecdotal evidence that some patients would remove the SpO_2 finger probes due to discomfort; thus, some improvement in data collection rates may be achieved by using more ergonomic pulse oximetry probes. However, the most significant losses of data occur when all vital signs are simultaneously lost, most likely as a result of patients being disconnected from bedside monitors when moving between locations in the department.

3.3.1. True Positives

Table 3.2 showed the number of physiological and non-physiological escalation events that the continuous T&T system would have been able to detect. In comparison to the 90 physiological escalation events detected by retrospective T&T computed for nurse observations, continuous T&T performs relatively poorly, detecting only 61 escalation events. The worse performance can be attributed to a number of factors.

For instance, the continuous T&T system cannot be expected to detect any of the 15 neurological escalation events (B1 and B2 events), as the measure of neurological function, GCS, requires a human observer. Relevant data may also be unavailable for any of the type-1 escalations event (A1,B1,C1), that is, escalation events that occurred at the time of, or prior to, the patient's arrival in the ED. In these instances, the patients will have been manually observed, assessed and escalated before being assigned to a bed with continuous monitoring equipment. In addition to this, vital sign data at the time of the escalation event may be unavailable in some cases due to patients being disconnected from the monitors. The extent of data loss during escalation events can be assessed using Figure 3.2.4, where we see that there was high data loss within the escalation windows.

3.3.2. False Positives

Table 3.2 also showed that 168 out of 269 patients would have generated False Positives with continuous T&T. These patients had no escalations of any type, but their vital signs still exceeded the T&T thresholds on at least one occasion. This again does not compare favourably with retrospective T&T, which generated only 80 false positives.

In our analysis, we counted a False Positive every time a patient for whom one of the T&T thresholds was exceeded despite the patient having no escalation event during their stay in the ED. The validity of using this method is heavily dependent on the notion that escalations are directly correlated with physiological abnormality, and that the opposite is also true. The data distributions in Figure 3.2.3 confirm that this is the case, showing that the vital sign data for the patient group with no escalation events is largely within the T&T thresholds.

In Section 3.2.2 we estimated the overall false alert rate for patients with no escalations

to be 0.97 alerts per hour, with alerts typically enduring for 5% of the total stay. While this figure may at first sight appear reasonable, this changes when we consider the false alert rate in the context of a typical ED such as the 20-bed unit at the John Radcliffe hospital. We would then expect 20 alerts per hour, or one every 5 minutes if alerts were spread uniformly. Thus, if continuous T&T were to be included in an audible alarm system, a constant and unacceptable level of background noise would be generated.

The high sensitivity of the system may be attributed to the observation frequency, because a high frequency allows T&T to detect short-term transient events which would be missed at lower frequencies of observation. Table 3.3 clearly demonstrates that the number of False Positives increases as the observation frequency is varied from 60 to 1 minute. Similarly, Figure 3.2.5 shows that the number of detected escalations increases with the observation frequency.

High numbers of alerts can occur when a patient is borderline abnormal, and many transient events occur. For instance, consider Figure 3.3.1, which shows the continuous T&T data alongside the vital sign data for study patient ED00571, an 89 year-old female patient who had presented to the ED after falling and dislocating an arm. The SpO_2 record is fairly stable, and fluctuates around a mean of 94%. This is verified by the manual observations, all of which are 94% or 95% during the patient's stay. However, for short periods of time, between one and two minutes, the SpO_2 intermittently dips below 92%, thus meeting one of the T&T thresholds. During the 90-minute monitoring period, the vital signs exceed the T&T thresholds on six separate occasions. However, there were no documented escalations during this period, which indicates that the six alerts were probably False Positives.

3.3.3. Persistence Criterion

From our previous example, it is clear that one of the reasons for the high number of false alerts is the frequent occurrence of short-term transient changes in the vital signs. These transients are most likely due to external factors such as patient movement. In order to reduce the effect of these transients, we now modify the continuous T&T system by introducing a persistence criterion, and assess its effectiveness using the same metrics.

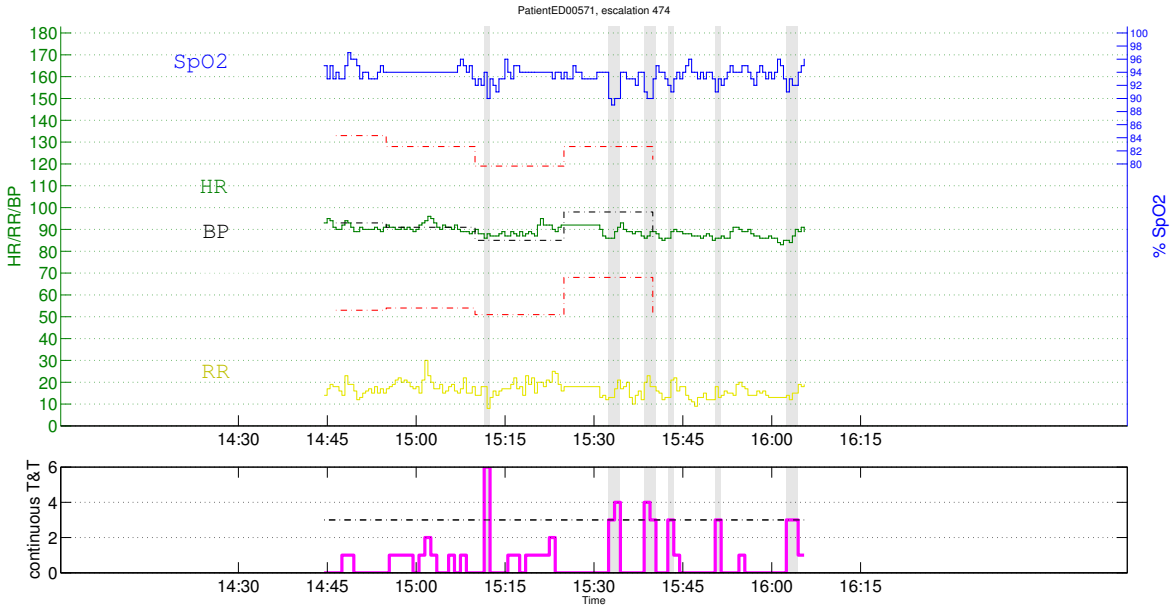

Figure 3.3.1.: Continuous T&T scores for study patient ED00571, showing multiple calls for intervention (highlighted in grey), based on a fluctuating SpO2, near the critical threshold.

In the original continuous T&T system, an alert was generated for any instance that the T&T alerting threshold was exceeded. In the modified model, an alert is only generated if the T&T alerting threshold is exceeded for any 4 minutes out of a sliding 5 minute window. The only exception to this rule occurs for blood pressure. If the T&T threshold is exceeded due to a change in blood pressure, then the alert is generated immediately, as the blood pressure measurements are typically recorded relatively infrequently.

The alert is stopped when the continuous T&T value drops below the alerting threshold for at least 2 minutes out of a 3-minute sliding window. The parameter values used here are arbitrary, and may be considered as merely an initial solution. The sensitivity and specificity of the continuous T&T system including the persistence criterion is:

$$sens_{T\&T} = \frac{24}{24+5} = 82.7\% \quad spec_{T\&T} = \frac{49}{49+168} = 22.6\%$$
$$sens_{T\&T\,persist} = \frac{19}{19+10} = 65.5\% \quad spec_{T\&T\,persist} = \frac{74}{74+143} = 34.1\%$$

We can examine the effect of the persistence criterion more widely by recalculating the time spent in the alert state and the number of alerts per patient, and replicating the graphs of Figures 3.2.6 and 3.2.7 for a modified continuous T&T system that includes a persistence criterion. Figure 3.3.2 can be directly compared to Figure 3.2.6, where it can

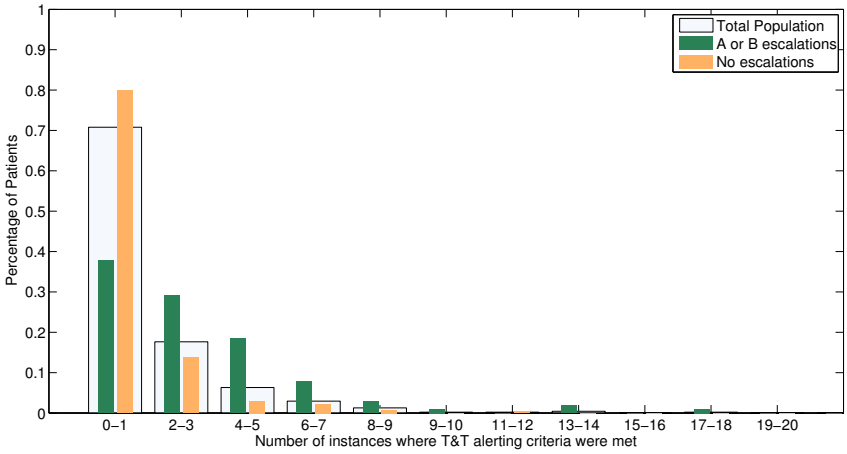

Figure 3.3.2.: Number of alerts per patient, using the continuous T&T system with a persistence criterion.

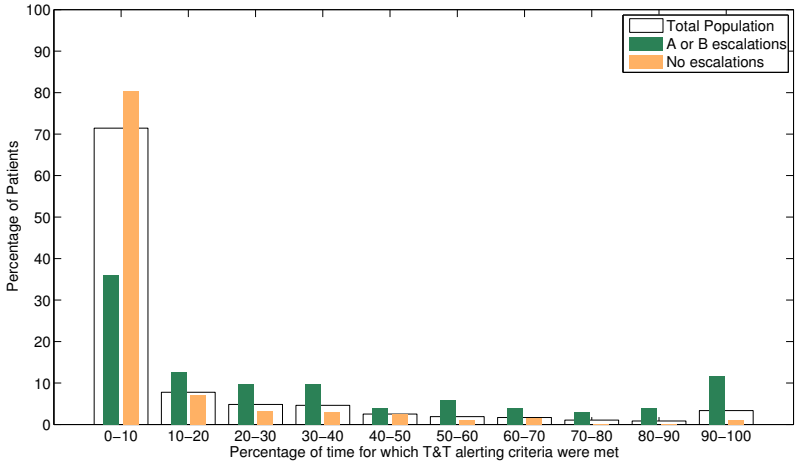

Figure 3.3.3.: Time spent in the alert state as a percentage of the total time in the ED, when using the continuous T&T system with a persistence criterion

be seen that far fewer alerts are generated by the modified system. Whereas previously the false alert rate was estimated at 0.97 false alerts/hour per bed, the effect of the persistence criterion reduces this number to 0.57 false alerts/hour per bed. In comparison to the original continuous T&T system (see Figure 3.2.7), the percentage of time spent in the alert state after the introduction of the persistence criterion, which is depicted in Figure 3.3.3, decreases slightly for each of the groups considered. The small changes in length of alert, and large changes in number of alerts, are entirely expected, as the persistence criterion acts to eliminate the many transient alerts.

The effect of the persistence criterion on an individual case can best be seen by considering once again patient ED00571. In Figure 3.3.4, the continuous T&T score including

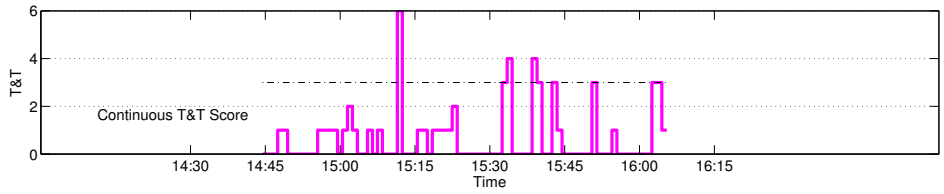

Figure 3.3.4.: Continuous T&T score including the persistence criterion for Patient ED00571. The corresponding vital signs and unmodified continuous T&T score are shown in Figure 3.3.1

the persistence criterion is plotted. In comparison to Figure 3.3.1, the modified system contains no greyed-out areas representing periods of alert for this patient, as each increase in the SpO_2 score lasts for a no more than one minute and are not frequent enough to meet the persistence criterion.

While the persistence criterion solves the problem of transient alerts, we also observe that continuous T&T has the further drawback of producing large changes in score even though there may only be relatively minor changes in the patient's vital signs. For instance, the T&T score for patient ED00571 fluctuates from a completely normal score of zero, to a severely abnormal score of 3. This behaviour highlights the coarseness of the T&T scores, which takes only integer values between 0 and 3. While this simplifies the calculation of the T&T score by the nursing staff, automatically computed continuous T&T scores do not have be subject to these limitations, and a better detection of deterioration may be achieved by a more finely-graded score.

3.4. Conclusion

We have attempted to use the T&T system in a continuous manner to try and identify deterioration events in ED patients. The continuous T&T system developed here did not fully replicate the manual T&T system as used on the ED because temperature and GCS measurements were not available, and so our implementation relied on only four vital signs. A more complete vital sign monitoring system may be able to use continuously recorded skin temperature measurements, and additionally make use of intermittent manual observations such as the GCS score.

Our results show that the four vital-sign continuous T&T system would have been

able to detect the majority of the escalation events, the latter being considered to be an appropriate proxy for patient deterioration. However, the system also misclassified a large proportion of the patients in the "no escalations" group, with many of these patients, 13.8%, having at least ten false alerts. The number of False Positives, 168, was significantly higher than the number of False Positives for the retrospective T&T system used in Chapter 2, where only 80 out of 269 patients were incorrectly classified.

In order to effectively detect patient deterioration in real time, a monitoring system must have high specificity, such that it does not generate large numbers of false alerts, while still being able to correctly identify escalations. We showed that the number of false alerts in the continuous T&T system may be reduced through the introduction of a simple persistence criterion. However, we further note that the current T&T scoring system has a coarse scale. For instance, a drop in SpO_2 from 100% to 93% will have no effect on the overall T&T score for that patient, but may have clinical value in helping to detect future deterioration. A finer-grained scoring system may lead to improved detection of patient deterioration.

In the following chapter, we begin to investigate whether other techniques may be more effective than continuous T&T at detecting escalation events while still maintaining a low False Positive rate. We start by analysing a baseline data fusion technique and then introduce three alternative techniques which address some of the baseline technique's weaknesses.

4. Data Fusion for Patient Vital Sign Monitoring

In Chapter 3, we demonstrated that it was possible to identify deterioration by applying the ED T&T criteria to continuous data. We also observed some drawbacks of this approach. Firstly, the continuous T&T system generated many false alerts. These were attributed to brief changes in the measurements caused by insignificant transient changes in the patient, or else by measurement artefacts caused by, for instance, patient movement. The number of false alerts was greatly reduced by the introduction of a persistence criterion.

Secondly, we also noted that the T&T scoring system is coarse-grained, as the T&T score for each vital sign parameter may only take integer values between 0 and 3. While this enables the total score to be calculated quickly by nursing staff, it may also prevent gradual deterioration from being detected.

In this chapter, we begin to investigate whether alternative methods may be more effective at identifying patient deterioration, while keeping the number of false alerts at a manageable level. Firstly, we will focus on an intelligent continuous vital sign monitoring system that uses a previously developed data fusion model. After describing how the model is derived, we then discuss its advantages and limitations. Following this, we outline two alternative methods for improving on this baseline data fusion model.

4.1. A Baseline Data Fusion Algorithm for Patient Monitoring

We firstly consider the vital sign data fusion system previously developed in our research group to identify patient deterioration in real time. The system, described by Tarassenko et al. [110], uses a model initially trained on pilot study data acquired at the John Radcliffe Hospital, Oxford, which has since been tested using data collected from the Clarian Methodist Hospital in Indianapolis and at the University of Pittsburgh Medical Centre (UPMC) [49, 56].

The central premise of the system is that acute patient conditions are strongly associated with uncommon, or novel, *vectors* of vital signs. A vital sign 'vector' is defined as the set of all vital sign parameters values recorded at one instant in time. The system attempts to distinguish between novel and normal vectors. By considering vectors, rather than each vital sign parameter individually, the method takes into account associations between parameters, as explained below.

There are two main approaches to classification. In "supervised" learning, the training data are labelled, and the classification algorithm attempts to group the training data such that data points with the same label are assigned to the same class. In this application, the simplest labels that we can assign the data are either "patient stable", or "patient unstable", which can then be used to classify the data.

However, in our case, most of the data will come from the "patient stable" group as events, such as escalations of care, are rare, even in acutely-ill patients. In addition, the "patient unstable" group is unlikely to cover all possible unstable conditions. In such a case, supervised learning may lead to incorrect classification in regions with few data points in the training set. The alternative is unsupervised learning, in which a labelled data set is not required, and which instead uses the distribution of features of the data set to learn the boundaries of the one class represented in the data set.

The algorithm described here uses unsupervised learning, or a one-class classification, assuming that the training data comes from patients with normal physiology. Any new data vectors that are sufficiently dissimilar to the training data are then considered to be

novel. To achieve this, it is assumed that any vital sign vector may be modelled as an independent selection from some underlying N-dimensional joint distribution over the N vital sign vectors. When a new vector from a continuously monitored patient is presented to the system, the probability of the vector being selected from the estimated underlying distribution is calculated, and then converted into a Patient Status Index (PSI). The PSI is designed so that novel vital sign vectors are assigned a high score. A threshold on the PSI is then determined so that any vectors above the threshold are assigned to the "patient unstable" group.

This method avoids two of the major drawbacks of the T&T early warning scores. Firstly, in contrast to the subjective T&T scores, PSI scores are calculated using a model derived from training data collected from a large, representative population of acutely-ill patients, thus providing an objective, data-driven score. In addition, the method used to evaluate the underlying model's probability density function (p.d.f.) allows for small changes in vital signs, unlike the T&T scores, which provide coarse estimates of vital sign abnormality as the scores may only take integer values.

The following section describes the data fusion algorithm in greater detail, showing how the model is derived from training data and how new vital sign data are interpreted to provide alerts as a result of patient deterioration. An overview of the training procedure is provided in Figure 4.1.1.

4.1.1. Training Data and Pre-Processing

The original model was trained on 3,500 hours of continuous vital sign data collected from 150 high-risk patients at the John Radcliffe Hospital, Oxford between 2001 and 2003 as part of an observational study [110]. The patient group included those who had severe heart failure, acute respiratory problems (such as acute asthma or pneumonia or pulmonary embolism), trauma, and those who were being continuously monitored following a myocardial infarction. The vital signs measured were HR, RR, temperature, SpO_2, and Systolic and Diastolic blood pressures (SBP and DBP), using the continuous monitoring methods described in Chapter 1. The HR, RR, temperature and SpO_2 values were sampled at a frequency of approximately 1Hz. SBP and DBP were measured at

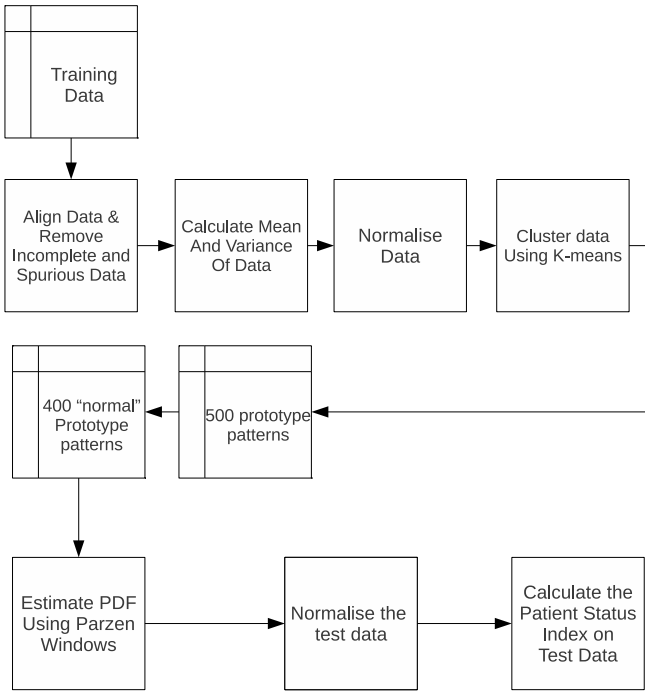

Figure 4.1.1.: Flow diagram showing the steps involved in constructing a model of patient normality and calculating a Patient Status Index for test vectors

	Lower Threshold	Upper Threshold
HR (bpm)	30	300
SDA (mmHg)	20	180
SpO_2 (%)	60	-
Temp. (°C)	32	39
RR (rpm)	3	45

Table 4.1.: Physiological upper and lower bounds for the five vital sign parameters. Systolic and Diastolic blood pressure have been combined into one parameter, Systolic-Diastolic Average (SDA)

30-minute intervals during the day, and at hourly intervals during the night, when the patient was asleep.

Because the channels of data were recorded asynchronously, the vital sign data were first aligned into vectors, and then sampled at 5-second intervals. This produced 2.6×10^5 vectors of vital signs, with each vector having five elements (one per vital sign parameter).

Vectors with elements which had physiologically implausible values were rejected according to the criteria shown in Table 4.1. Any SpO_2 readings below 85% were also discarded, as the pulse oximeter is considered to be inaccurate for SpO_2 measurements below this value. The effect of this was to reduce the number of available training vectors to 2.4×10^5. The distributions of each of the vital signs in the training set are shown in Figure 4.1.2.

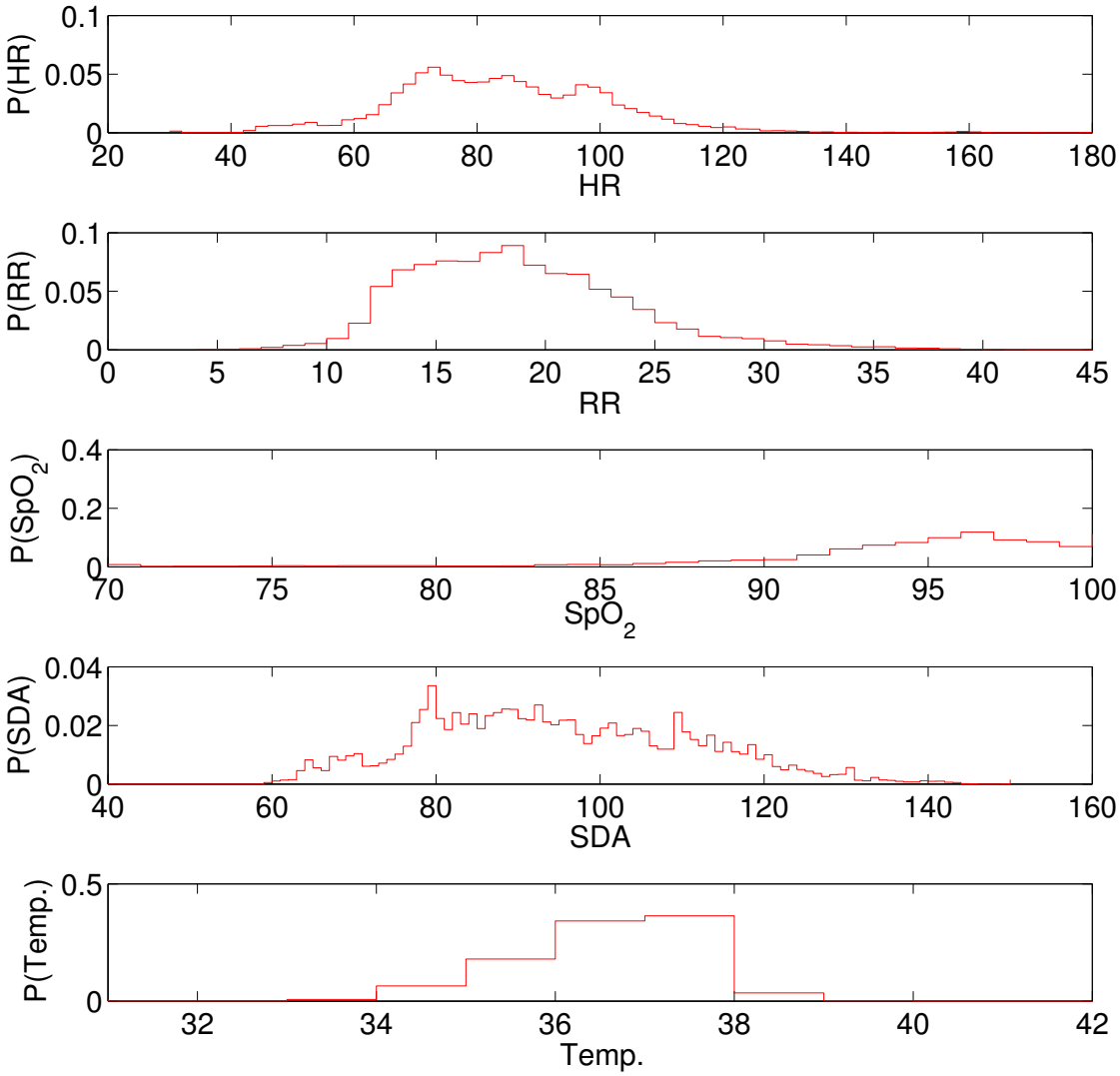

Figure 4.1.2.: Vital sign distributions for the training data set

	μ	σ
HR (bpm)	83.7697	17.4831
SDA (mmHg)	94.6828	16.5471
SpO_2(%)	95.2000	3.4900
Temp. (°C)	36.0459	1.2767
RR (rpm)	18.3043	5.0568

Table 4.2.: Mean (μ) and standard deviation (σ) of each vital sign parameter in the training set

In the data fusion model, it is implicitly assumed that each of the vital signs has equal importance, and should therefore have an equal weighting in the model. The final two steps of data pre-processing were undertaken according to this assumption. Firstly, only one measurement of blood pressure was used in the vital sign vector (now five-dimensional) to avoid placing more importance on blood pressure than on the other vital signs. SBP and DBP, were combined into one parameter, the Systolic-Diastolic Average (SDA), by calculating their arithmetic mean. As a result, SBP and DBP have equal influence on the blood pressure parameter, despite the fact that SDA is not a standard measure in clinical care.

The final stage of pre-processing involved scaling each of the parameters by applying the zero-mean unit-variance transformation on each vital sign measurement v:

$$v_n = \frac{v - \mu}{\sigma} \tag{4.1.1}$$

where v_n is the scaled value, μ is the mean value for that vital sign in the training set, and σ is the training set standard deviation. The mean and variance of each vital sign parameter are shown in Table 4.2.

4.1.2. Parzen Windows

Parzen windows allows the underlying 5D vital sign distribution, or p.d.f., to be estimated from training data points. While other methods, such as Gaussian mixture models, were considered, Parzen windows was chosen as it has the advantage of being a non-parametric technique. This means that no *a priori* assumptions are made about the form of the probability distribution.

In the Parzen windows scheme, we estimate the p.d.f. of a random variable, X, by placing a kernel function on each training data point. The estimated p.d.f. is the linear combination of the kernels, which is then normalised by the number of kernels so that the integral is 1.0. Mathematically, Parzen windows can be described as follows: if $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N \sim f$ is an independent and identically distributed sample of a random variable, then an approximation of the p.d.f. evaluated at a new data point, $\mathbf{x}$, can be written as:

$$\hat{f}_h(\mathbf{x}) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{4.1.2}$$

where K is some kernel function, N is the number of training data points, and h is a smoothing parameter. Analysis has shown that, in the limit, the shape of kernel is not crucial for estimating the p.d.f. in the case of independent and identically distributed random variables [28], but it must be symmetric and integrate to 1.0. For this application, a multivariate Gaussian kernel with dimension, d, and zero mean and unit variance was chosen:

$$K_x = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\mathbf{x}^2}{2}\right\} \tag{4.1.3}$$

so that the Parzen windows estimate for this problem is:

$$\hat{f}_h(x) = \frac{1}{N(2\pi)^{d/2}\sigma^d} \sum_{i=1}^{N} \exp\left\{-\frac{|\mathbf{x} - \mathbf{x}_i|^2}{2h^2}\right\} \tag{4.1.4}$$

The smoothing parameter, h, is also equivalent to the kernel width, and has the effect of controlling the level of detail in the Parzen windows p.d.f. As Figure 4.1.3 demonstrates, a large value of h leads to a distribution that is too general and does not capture the details of the data, while too small a value causes the p.d.f. to be over-fitted to the data. In theory, a risk function metric such as the Mean Integrated Squared Error may be used to derive the optimum value of h. However, in practice, the true state of the underlying p.d.f. is not known, and data-based methods such as cross-validation, or maximum likelihood estimates are often used.

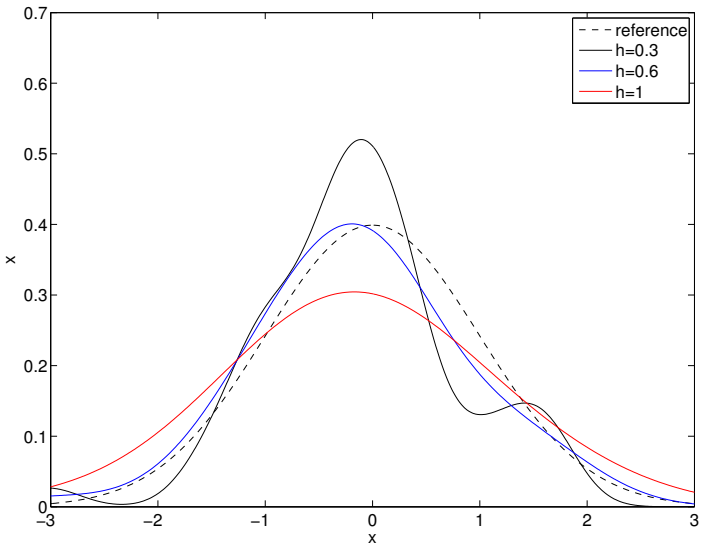

Figure 4.1.3.: Parzen windows model of a zero mean, unit variance Gaussian distribution using 30 kernels and various width parameters. The original distribution is shown in dashed lines. For h=1, the Parzen windows p.d.f. is wider than the underlying distribution, whereas for h=0.3, the thin kernels lead to an over-fitted model.

4.1.3. Application of Parzen Windows

While the Parzen windows technique is simple and generalisable, it is impractical to apply it directly to our training data set due to the large number of training vectors ($> 10^5$). In particular, such a solution would require storage of all the training vectors, $\mathbf{x}_i$, and each estimate $f(\mathbf{x})$, would require a number of additions and computation time proportional to the size of the training set.

To circumvent this problem, the number of training vectors is reduced to 400 "prototype vectors". The reduction in training vectors is implemented in two stages. Initially, the training data vectors were clustered into 500 prototype vectors using the K-means algorithm (see algorithm 4.1); the resulting centroids of each cluster were defined to be

Algorithm 4.1 The K-means algorithm

1. Place K initial points into the space represented by the objects that are being clustered
2. Assign each object to the group that has the closest centroid
3. When all objects have been assigned, recalculate the positions of the K centroids
4. Repeat until convergence

a prototype vector. The number of prototype vectors selected at this stage was chosen empirically, but 500 centres have since been shown to provide a reasonable model [26].

The training data from which the p.d.f. is derived contains vital signs from *all* monitored patients. In the majority of cases, these vital signs correspond to times when the patient was stable. However, in a few instances, vital signs will have been recorded from patients who were unstable, even for a short period of time. These abnormal events in the data set will be captured as clusters during the K-means procedure, and we should therefore remove the outlying cluster centres to ensure that only normal physiology is captured .

The second stage of the data reduction was implemented by discarding the 100 prototype vectors with the greatest Euclidean distance from the origin so that only the 400 most "normal" vectors are used in the subsequent Parzen windows model. An investigation by Hann [41] using patient data from a separate study confirmed that this adaptation was an improvement over the 500-vector model.

The removal of 100 points was based on the empirical observation that roughly 20% of the vital sign values recorded for acutely-ill patients represent abnormal physiology. The removal of 20% of values can be further justified by calculating the percentage of manual observations, for the ED data, that resulted in a single vital sign channel with a T&T score of 3. In total 722/3025 (23.9%) of the ED observations met this criterion.

The effect of this process is visualised in Figure 4.1.4 using a 2D Sparse Approximated Sammon Stress (SASS) map of the cluster centres. The SASS visualisation is based on the Sammon map, which attempts to maintain Euclidean distances between points in the original feature space, and the (typically 2D) visualisation space. Further details on both Sammon maps and SASS visualisation are given in Appendix B.

The baseline model of normality estimates the p.d.f. of the vital sign distribution using the Parzen windows method with 400 prototype vectors. The kernel width parameter was set using a heuristic suggested by Bishop [11], who recommended calculating the mean of the local estimate of the variance at each vector location:

$$h = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m}\sum_{j\in Q_i}\left|\mathbf{x}_i - \mathbf{x}_j\right|^2\right) \tag{4.1.5}$$

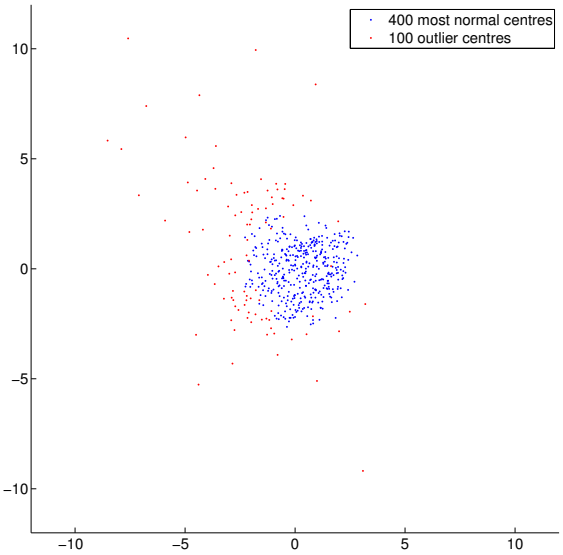

Figure 4.1.4.: SASS map of the 400 prototype centres in blue, and the 100 removed prototype centres in red.

where N is the number of prototype vectors, and Q_i are the m nearest neighbours for each vector. $m = 10$ members was chosen, which produced a value of $h = 1.49$, and so the Parzen windows estimate for the baseline model, $f(\mathbf{x})$, is given by:

$$f(\mathbf{x}) = \frac{1}{400(2\pi)^{\frac{5}{2}}1.49^5}\sum_{i=1}^{400}\exp\left\{-\frac{|\mathbf{x}-\mathbf{x}_i|^2}{2\times 1.49^2}\right\} \tag{4.1.6}$$

4.1.4. Patient Status Index

In order to provide a score corresponding to the novelty of a patient's vital sign vector, the Patient Status Index (PSI), which is also known as the Visensia Status Index and Novelty Index in related literature [50, 49], is calculated from the p.d.f. as follows:

$$PSI = \log\left(\frac{1}{p(\mathbf{x})}\right) - \log\left(\frac{1}{p_{max}(\mathbf{x})}\right) = log\left(\frac{p_{max}(\mathbf{x})}{p(\mathbf{x})}\right) \tag{4.1.7}$$

where $p(\mathbf{x})$ is the p.d.f. evaluated at $\mathbf{x}$, and $p_{max}(\mathbf{x})$ is the maximum possible value of $p(\mathbf{x})$. This point can be determined using gradient descent methods, and is approximately located at the origin and has a value of 6.08 when a 5-dimensional Parzen windows model is generated using the training data. $p_{max}(\mathbf{x})$ is subtracted to adjust the scale so that the PSI is close to zero when all the vital signs are normal. The log transform of the

probability is used so that a low probability corresponds to a high PSI score, so that a high score indicates highly improbable, abnormal physiology.

4.1.5. Frequency of Score Calculation and Missing Data

To use the data fusion model with test data, a vector of data containing all channels of vital sign data is required so that a point on the p.d.f. can be evaluated. However, with most vital sign monitors, each channel of data is treated independently and hence data are received asynchronously. To deal with this, values of all the vital sign parameters are sampled-and-held, and a new PSI is calculated each time new data from a single channel becomes available.

In practice, vital sign data may be unavailable for extended periods of time due to disconnection of the sensors; this often occurs as electrodes become poorly attached over time, or as patients actively remove the pulse oximeter finger probe. To deal with this situation, a simple heuristic is used. If a vital sign parameter value is missing over a one minute period, the median value of that vital sign over the last five minutes is used instead. This heuristic is used for all the vital signs apart from blood pressure, which is sampled far less frequently.

If data are missing for 30 minutes or more, the mean value of the vital signs in the training data set is used instead. This has the effect of reducing the dimensionality of the data fusion model by limiting vital sign vectors to a 4D cross section of the 5D data space. The short term median filter and population mean methods are only deemed valid in the cases where up to two vital sign parameters are missing. In the case of any further data drop-out, no PSI is calculated.

4.1.6. Alert Generation

By setting a suitable threshold on the PSI, alerts can be generated that are associated with vital sign abnormality. The baseline model uses a threshold of PSI = 3.0, and the suitability of this threshold was assessed by considering how the PSI responded to single channel events. This was achieved by varying only one vital sign parameter at a time, between -4 and +4 standard deviations from the mean, while keeping the others

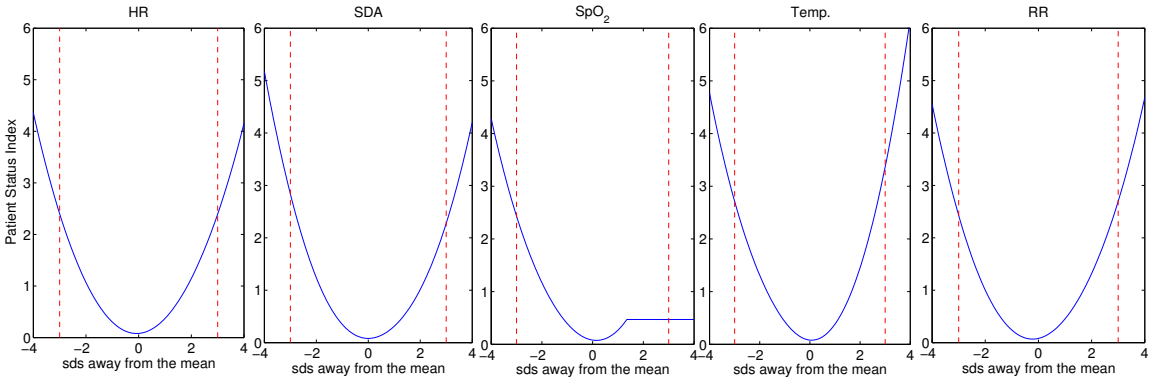

Figure 4.1.5.: Patient Status Index for the case when each vital sign parameter is varied from -4 to +4 standard deviations in turn, while fixing the remaining vital signs at their normalised mean value of zero.

	Lower Threshold		Upper Threshold	
	Vital sign	Deviation from mean (σ)	Vital sign	Deviation from mean (σ)
HR (bpm)	25.29	-3.34	142.90	3.38
SDA (mmHg)	43.72	-3.08	151.29	3.42
SpO_2(%)	83.52	-3.35	N/A	N/A
Temp. (°C)	32.01	-3.16	39.67	2.84
RR (rpm)	1.59	-3.30	34.37	3.18

Table 4.3.: Values of individual parameters that cause the novelty to exceed $PSI = 3.0$, when the other parameters are set to the mean value in the training set.

fixed at their mean value (Figure 4.1.5). SpO_2, was only varied between -4 and +1.37 standard deviations away from its mean value, as 100% oxygen saturation corresponds to $\mu_{SpO_2} + 1.37\sigma_{SpO_2}$.

The value of each vital sign at the candidate threshold is shown in Table 4.3, and demonstrates that a PSI of 3 is reached when any the vital sign parameters are between 2.84 and 3.42 standard deviations from their mean value in the training set.

Tests conducted on a data set collected from a similar patient population to the training data showed that the candidate alerting threshold was highly effective for single-channel alerts and also for alerts that were caused by multiple vital signs [110]. Although a fixed threshold of PSI 3.0 has both theoretical and empirical support, more complex methods of generating alerts may produce better results [21].

In order to reduce the number of false alerts due to transient spikes in the PSI, a heuristic persistence criterion that was first introduced for continuous T&T in Section

3.3.3, is also introduced to the data fusion system here so that alerts are only generated if the PSI exceeds the alert threshold for four minutes within the previous five minutes. The alert will remain active until the PSI drops beneath the threshold for two minutes out of three. The only exception to this is for alerts generated due to blood pressure. Any PSI score that exceeds the alerting threshold due to changes in blood pressure generates an immediate alert.

4.2. Shortcomings of the Data Fusion Algorithm

In the algorithm described in 4.1, we noted that directly applying Parzen windows to the entire training data set was unfeasible due to computational costs. Instead, a sub-set of 400 "prototype" centres was extracted under the assumption that this reduced set would be representative of the full training set. Further analysis, presented here, shows that this assumption is not entirely correct.

In the first instance, it is simple to show that the assumption underlying the use of the clustering step using the K-means algorithm is not true for all cases by using a 1-D example. Consider the example in Figure 4.2.1, which shows a number of training data points. The black line in the upper figure shows the kernel density estimate from Parzen windows, which appears to be an adequate estimate. The lower figure shows the effect of reducing the same data set by using K-means clustering with two centres, following the procedure described in Section 4.2. The output of the K-means algorithm is represented by two red crosses, and the subsequent Parzen windows estimate is shown as a black line. The p.d.f. estimates using the two methods are considerably different.

The difference can be explained by noting that the K-means clustering algorithm can produce clusters of unequal population. When Parzen windows is used subsequently, the least populated cluster will have the same influence as the most populated cluster. This can be seen in Figure 4.2.1, in which the right-hand cluster has support from only four of the twenty four data points, yet contributes to 50% of the probability mass estimate.

This argument holds true as long as there are differences in the cluster populations, and the size of the effect will depend on how different the cluster populations are. It therefore remains for us to show whether differences in cluster population exist when the algorithm

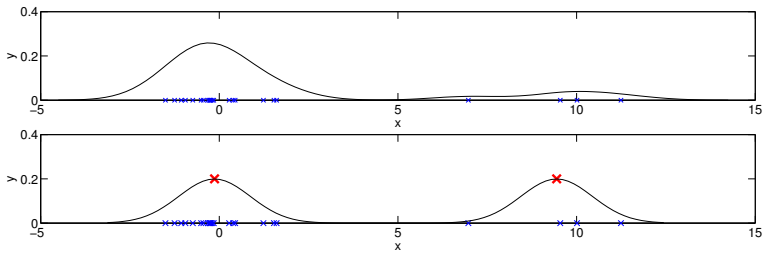

Figure 4.2.1.: The top graph shows a Parzen windows estimate of the underlying p.d.f. based on the training data (in blue). The lower graph shows the effect of applying an intermediate clustering step to provide two "prototypes" (in red). The Parzen windows estimate in this situation is considerably different to that in the top graph, and does not model the data correctly.

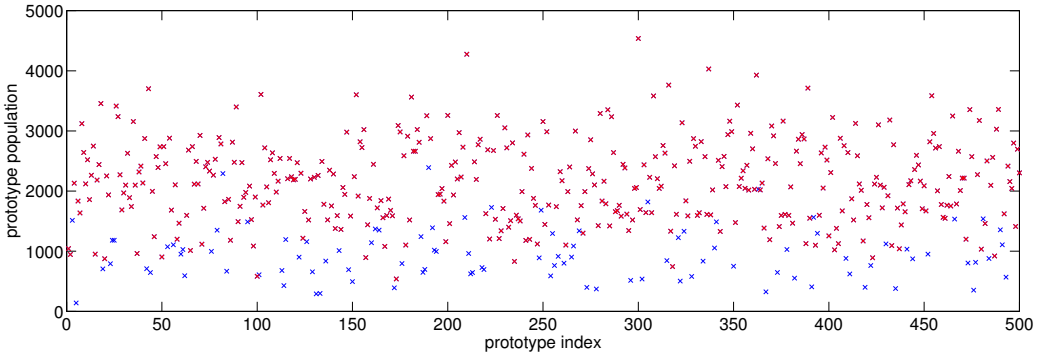

Figure 4.2.2.: K-means cluster populations for the 400 prototype centres and the 100 "outlying" centres, as described in Section 4.1.

is applied to real training data. For the training procedure and the data set described in Section 4.1, the K-means cluster populations of the 400 prototypes are highlighted in red in Figure 4.2.2. In addition, the remaining 100 'outlying' clusters are shown in blue. As expected, the 100 removed 'outlying' clusters have much lower populations, which serves to reduce the error in the Parzen windows estimator.

However, the population of the remaining 400 prototype clusters still varies from 539 to 4538 data points. In the original algorithm, all of these prototypes have equal influence on the model, whereas in a more principled scheme, the prototype with 4538 data points associated with it should have approximately eight times the influence of the prototype with 539 data points.

In summary, the use of the K-means algorithm, in combination with Parzen windows, may cause the baseline model's effectiveness at detecting physiological deterioration to be sub-optimal. Hence, there is a need to improve the baseline model's training procedure, or else alternative machine learning techniques should be investigated. The remainder

of this chapter considers two alternatives to the current training procedure. The first method, weighted Parzen windows, is a natural extension to the original algorithm that addresses the issue highlighted in this section. The second method uses one-class Support Vector Machines (SVM), to construct the vital sign data fusion model of normality.

4.3. Weighted Parzen Windows

In the previous section, we highlighted the fact that the baseline model incorrectly estimates the p.d.f. because each prototype centre is assigned an equal prior, despite the fact that each prototype represents a different number of training data. The most natural solution to this problem is to allow each prototype a prior based on the number of patterns associated with that cluster. This is known as Weighted Parzen Windows (wPw), and was first proposed by Babich and Camps [5] to deal with the problem of the significant processing time and data storage needed to compute a kernel density estimate as a data set becomes large. Mathematically, the wPw approximates Parzen Windows using a set of m prototype patterns, and is described by:

$$p_m(\mathbf{x}) = \sum_{i=1}^{m} \frac{\omega_i}{h} K\left(\frac{|\mathbf{x} - \mathbf{x}_i|}{h}\right) \tag{4.3.1}$$

where ω_i is the i^{th} cluster weighting and is equal to population of the i^{th} cluster, divided by the total size of the training set. All other variables take the same meaning as in equation 4.1.4. The result of this process is that the kernels that lie in low density regions of data space are assigned lower weightings, reducing those kernels' influence on the overall shape of the estimated p.d.f., thus providing a more accurate estimate. Once the wPw model is generated, the PSI score can be calculated using the transform given in Equation 4.1.7.

The accuracy of distributions generated by wPw was assessed on a simple 2D example. 10,000 samples were selected from a 2D Gaussian distribution with unit variance in both directions (Figure 4.3.1(d)). The data were then down-sampled using the K-means algorithm to generate a subset of 50 prototype centres. The populations of each cluster were stored and used to compute a kernel density estimate using wPw as described in

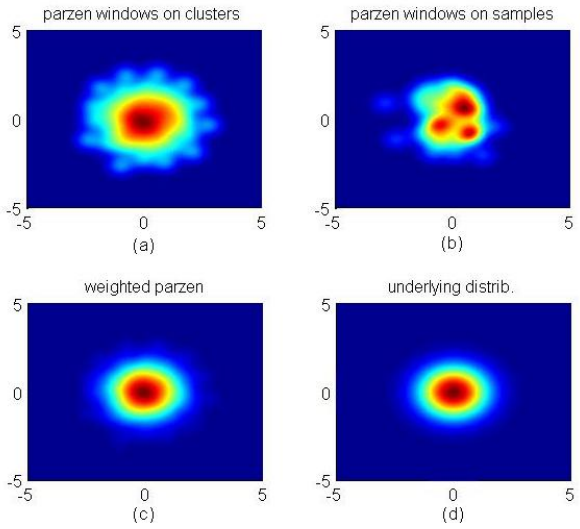

Figure 4.3.1.: Kernel density estimates using 50 kernels derived from a 2D Gaussian distribution. (a) shows the Parzen Windows result using the original training procedure (b) shows the Parzen Windows estimate using 50 kernels selected directly from the underlying distribution, (c) shows the result of wPw, and (d) shows the original distribution.

Equation 4.3.1, and the result is shown in Figure 4.3.1(c). The result from using the original algorithm is shown in Figure 4.3.1(a). For comparison, 4.3.1(b) shows the effect of applying Parzen windows to a randomly selected subset of 50 of the 10,000 samples. In each case, the Parzen width parameter was set empirically at $h = 0.1$.

Through visual inspection, we can see that wPw is a better estimate of the original Gaussian distribution than the limited centres Parzen Windows (b) and the original training method (c). The degree of error for each of the models can be quantified using the Bhattycharyya distance, which measures the similarity of two probability distributions. The Bhattycharyya distance is defined for discrete distributions as:

$$D_b(p,q) = -ln(BC(p,q)) \tag{4.3.2}$$

where:

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{4.3.3}$$

The Bhattycharyya distance between the original distribution and the estimates in (a),(b) and (c) are 0.0470, 0.0662 and 0.0045 respectively, again demonstrating that the

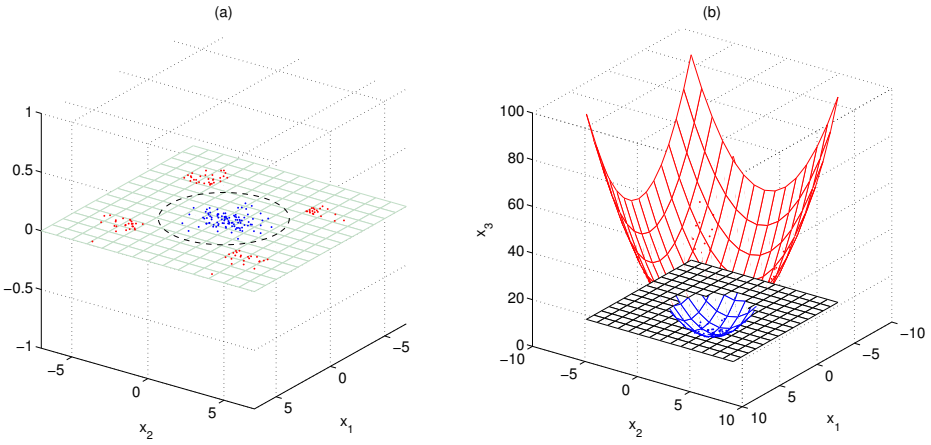

Figure 4.4.1.: (a) an example of a 2-input (x_1, x_2) two-class data set, show by the red and blue dots. The data are not linearly separable, as indicated by the black dashed line (b) shows the same two-class data set, with an additional feature $(x_3 = x_1^2 + x_2^2)$. The inclusion of this extra feature allows the data set to be linearly separated by a plane through $x_3 = 20$, as indicated in black.

wPw estimate is the most accurate. The distances were calculated over the range $-5 \leq x_1 \leq 5$ and $-5 \leq x_2 \leq 5$ in steps of 0.1 in each direction.

4.4. Support Vector Machines

One recent machine learning technique that has proved to be popular for pattern recognition problems is the Support Vector Machine (SVM). Unlike the baseline model, SVMs are unable to produce a probabilistic output, and cannot therefore provide a meaningful PSI score. However, the method has been shown to provide accurate classifications in numerous applications (for example, see [54, 81, 4]) which may be used, in the vital sign monitoring context, to generate patient alerts. Introduced by Vapnik in 1995 [22], SVMs were originally developed for two-class classification. Although the basic technique uses labelled data and is thus a form of supervised learning, SVMs can also be adapted for use in novelty detection applications by using a one-class unsupervised version of the method.

The SVM attempts to linearly separate two classes of data (i.e. creating an optimal hyperplane) in some feature space that may be high or infinite dimensional. Such a separation can be described mathematically by:

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + \mathbf{b} \tag{4.4.1}$$

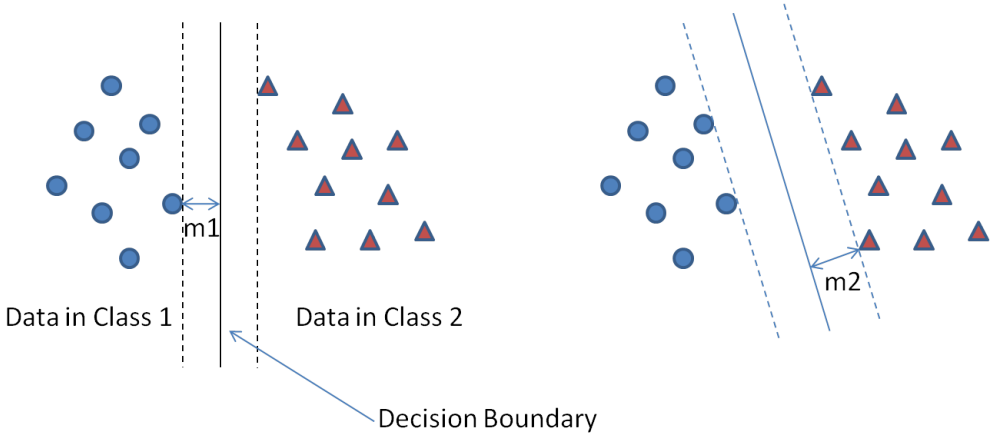

Figure 4.4.2.: One margin, $m1$, is shown for a simple synthetic 2D data set in which the feature space is the same as the input data space. The optimal solution, which maximises the margin, is shown as $m2$

where $\phi(\mathbf{x})$ describes the transformation from the input data space to the feature space, $\mathbf{w}$ is a vector of weights and $\mathbf{b}$ is a bias term. $y = 0$ then describes the decision boundary plane. The transformation of the data into a high-dimensional feature space increases the likelihood that the data become linearly separable. For instance, consider the case shown in Figure 4.4.1(a). In this example, the training data is two dimensional (x_1, x_2) and the two classes of data, show in red and blue, are not linearly separable. By the inclusion of an additional feature, $x_3 = x_1^2 + x_2^2$, the data can be separated by a plane in the 3D feature space (Figure 4.4.1(b)).

The degree of separation between the two classes is defined with reference to the "margin", which is the minimum perpendicular distance between the decision boundary and any of the transformed data points in the training set. Optimal separation occurs when the margin is maximised. For instance, the example in Figure 4.4.2 shows two possible separating planes, but the margin $m2$ is larger and thus corresponds to a better solution.

Figure 4.4.2 also demonstrates that the optimal solution in this case is only affected by the four data points that are closest to the decision boundary. This hints at the fact that general solutions to this type of problem may only depend on a small subset of the data, the so-called "support vectors". We now show how this problem can be formulated mathematically, following the treatment of the problem by Bishop [12].

4.4.1. Primal Formulation

Initially, we consider a two-class problem for which each of the data points is assigned a target, t_n, of -1 or $+1$ according to their class. Furthermore, we assume that the data can be linearly separated in the feature space. The decision boundary is defined by the hyperplane $y(\mathbf{x}) = 0$, and all new data are classified according the sign of $y(\mathbf{x})$, so that any data for which $y(\mathbf{x}_n) < 0$ is assigned $t_n = -1$, otherwise it is assigned to $t_n = +1$. Consequently, the following inequality must hold:

$$t_n y(\mathbf{x}_n) > 0 \tag{4.4.2}$$

The perpendicular distance of a point $\mathbf{x}$, from the separating hyperplane $y(\mathbf{x}) = 0$ can be shown to be $\frac{y(\mathbf{x})}{\|\mathbf{w}\|}$ through geometric arguments [12]. Using this identity with Equation 4.4.2, the scalar distance between some point, $\mathbf{x}_n$, and the plane is given by:

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + \mathbf{b})}{\|\mathbf{w}\|} \tag{4.4.3}$$

where the multiplication of $\frac{y(\mathbf{x}_n)}{\|\mathbf{w}\|}$ by t_n merely ensures that resulting distance is positive. The minimum distance to the hyperplane (i.e. the margin) can be found by minimising this expression over all n data points. The SVM solution seeks to maximise the margin by adjusting the hyperplane through the parameters $\mathbf{w}$ and $\mathbf{b}$. The SVM objective function is thus the maximin:

$$arg\,max_{w,b}\left\{\frac{1}{\|\mathbf{w}\|}\min_n[t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + \mathbf{b})]\right\} \tag{4.4.4}$$

where the term in the curly brackets is the margin. Without loss of generality, $\mathbf{w}$ and $\mathbf{b}$ can be rescaled to derive the canonical representation of the problem. We rescale the problem such that the minimum distance to the hyperplane is fixed as 1, and thus all of the n data points, $\mathbf{x}$, must be subject to the constraint:

$$y(\mathbf{x}_n) = t(\mathbf{w}^T\phi(\mathbf{x}_n) + \mathbf{b}) \geq 1 \qquad n = 1, ..., N \tag{4.4.5}$$

Thus, the maximin optimisation in equation 4.4.4 can be reduced to the maximisation

of $\frac{1}{\|\mathbf{w}\|}$, subject to the *constraint* in equation 4.4.5. Rather than maximising $\frac{1}{\|\mathbf{w}\|}$, we instead choose to solve the equivalent problem, the minimisation of $\frac{1}{2}\|\mathbf{w}\|^2$, where the coefficient of $\frac{1}{2}$ has been introduced to simplify the mathematics. It should be noted that the original objective function called for a maximisation over both $\mathbf{w}$ and $\mathbf{b}$. However, in practice we only need to optimise over $\|\mathbf{w}\|$, as all changes in $\mathbf{w}$ will affect $\mathbf{b}$ via the constraints.

The constrained optimisation problem can be reformulated using Lagrange multipliers. Lagrange multipliers are a mathematical technique which allow constrained optimisations to be rewritten as an unconstrained optimisation in terms of an objective function and a weighted sum of the constraints. The reformulation is known as the Lagrangian [24]. The Lagrangian for this problem is:

$$L(\mathbf{w}, \mathbf{b}, a) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\phi(\mathbf{x}_n) + \mathbf{b}) - 1\} \tag{4.4.6}$$

where a_n is the Lagrange multiplier for the constraint on the data point $\mathbf{x}_n$. There is a minus sign in front of the Lagrange multiplier, as we are minimising with respect to $\mathbf{w}$ and $\mathbf{b}$, and maximising with respect to a. The Lagrangian can then be solved directly using computational optimisation techniques (such as gradient descent) to give values for $\mathbf{w}$ and $\mathbf{b}$. However, we instead choose to reformulate the problem into a form that is easier to solve, and allows us to work more easily in high dimensional feature space.

4.4.2. Dual Formulation

The constrained SVM problem can be converted into an alternative, dual, form by noting that the partial differentials of the Lagrangians must be zero at the solution. The resulting expressions may be substituted back into the Lagrangian, eliminating w, and producing an optimisation in terms of the Lagrange multipliers instead. In this case, the partial derivatives must satisfy:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum a_n t_n \phi(\mathbf{x}_n) = 0 \\ \frac{\partial L}{\partial a} &= \sum a_n t_n = 0 \\ \frac{\partial L}{\partial \mathbf{b}} &= 0 \end{aligned} \tag{4.4.7}$$

By substituting these back into the Lagrangian, the dual representation of the problem can be derived, which expresses the optimisation in terms of the Lagrange multipliers and a kernel function $k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x})^T \phi(\mathbf{x})$:

$$L(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \tag{4.4.8}$$

The Lagrangian is now maximised with respect to a. In order to ensure that there is a solution, the kernel must be positive definite, such that the second term in the equation is always negative. In this case, the solution of the constrained optimisation can be solved using the Karush-Kuhn-Tucker (KKT) conditions, which are a set of general constraints that are used when the problem constraints contain inequalities [24]. The KKT conditions in this case are:

$$\begin{aligned} a_n &\geq 0 \\ t_n y(\mathbf{x}_n) - 1 &\geq 0 \\ a_n \{t_n y(\mathbf{x}_n) - 1\} &= 0 \end{aligned} \tag{4.4.9}$$

and can be solved computationally.

The advantage of the dual representation is that all of the data ($\mathbf{x}$) terms appear within a dot product, or kernel. Rather than having to specifically choose a high dimensional transformation $\phi(\mathbf{x})$, and solve the optimisation in the transformed space, the so-called 'kernel trick' allows us instead to pick a simple function that returns the dot product of some implicit nonlinear transformation. To classify new points using the dual form, we substitute for $\mathbf{w}$ using Equation 4.4.7 to transform the expression in Equation 4.4.1 to one that only contains the kernel k:

$$y(x) = \sum a_n t_n k(\mathbf{x}, \mathbf{x}_n) + \mathbf{b} \tag{4.4.10}$$

we can then calculate $y(\mathbf{x})$ for a new input ,$\mathbf{x}$, and classify the point according to the sign of $y(\mathbf{x})$.

An example SVM output showing the classification between two sets of data in red and blue is shown for the 2D case in Figure 4.4.3. The support vectors are circled, and lie close to the decision boundary in the data space and would be the closest points to

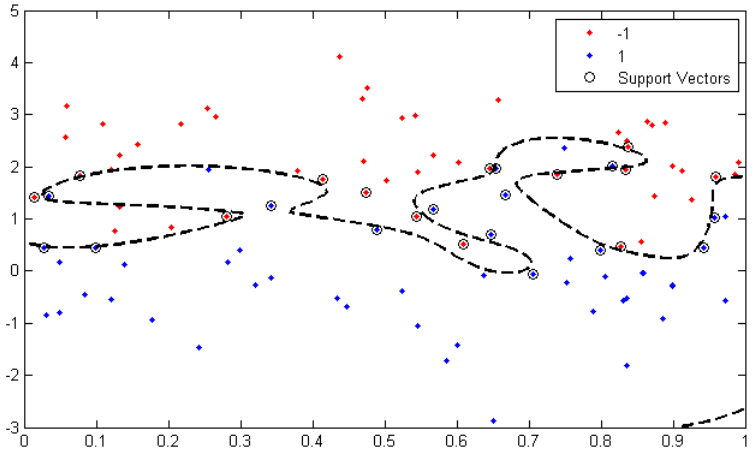

Figure 4.4.3.: A non-linear classification using SVMs with Gaussian kernels on a synthetic 2D data set. The support vectors are the small subset of points that lie closest to the decision boundary in the feature space. In this case, the support vectors also appear close to the decision boundary in the input data space.

the decision boundary in the feature space. Note that the data are linearly separated in the high-dimensional feature space, which in this case corresponds to a highly non-linear separation in the input data space.

4.4.3. Slack Variables

So far, we have assumed that it is appropriate to separate the data in the feature space. Even in the high-dimensional space, this may not be possible, and even in cases where separation is possible, the resulting solution may generalise poorly and so the concept of "slack" variables was introduced by Vapnik (1995). In this scheme, each data point is assigned a slack variable which takes a value of zero for each data point on the correct side of the margin, and $|t_n - y(\mathbf{x}_n)|$ for every point on the wrong side of the margin. This means that any data point on the decision hyperplane has $\xi = 1$, and any misclassified data point has $\xi > 1$ (see Figure 4.4.4)

The mathematical formulation of the SVM optimisation changes only slightly. Rather than minimising $\frac{1}{2}\|\mathbf{w}\|^2$, the objective function is now:

$$C\sum \xi_n + \frac{1}{2}\|\mathbf{w}\|^2 \tag{4.4.11}$$

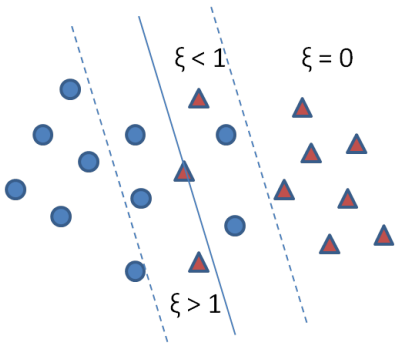

Figure 4.4.4.: Depiction of slack variables for the triangular class. All data points in the correct class and outside the margin are scored zero. Data points in the correct class but within the margin are scored linearly between one and zero. Data points in the incorrect class have a slack variable greater than one.

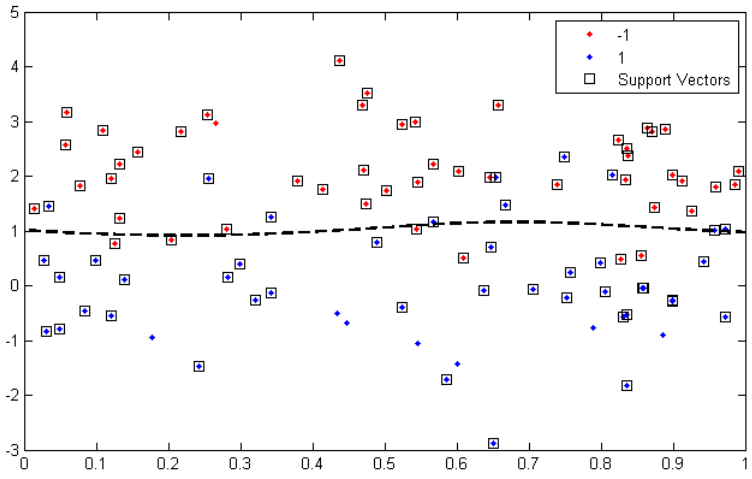

Figure 4.4.5.: A non-linear classification using SVMs with soft margins on the synthetic 2D data set. The slack variable takes the value of $C = 1$.

In this case, the slack variables allow a 'softer' hyperplane boundary such that, the optimal hyperplane is now the one that best classifies the training data, while still maintaining a large margin, or separation, between the classes. The effect of the slack variables is controlled by C; when $C \to 0$, the solution returns to the "hard" margin solution. The effect of using soft margins ($C > 0$) is shown in Figure 4.4.5 for a 2D example with a conservative value of $C = 1$, which provides a much smoother boundary.

4.4.4. One-Class Support Vector Machines

With the vital sign monitoring application considered in this thesis, there are relatively few instances of vital signs from deteriorating patients. Scholkopf [95] extended SVMs for

single-class (i.e. novelty detection) problems, and the technique has been used successfully in applications such as document classification [69], and seizure analysis [36].

The main idea in this scheme is that the vast majority of the data come from a single class, which is to be separated from the origin in the feature space with maximum margin. By using an appropriate kernel function, the origin maps to infinity in all directions in the feature space. One such kernel function is the Gaussian kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) = e^{\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\gamma}} \tag{4.4.12}$$

where γ is a free parameter that controls the size of the kernel. As $\mathbf{x}_i$ and $\mathbf{x}_j$ move further apart in the data space, then the value of the kernel function in feature space becomes closer to the origin. In order to separate the data from zero, we assume that the margin is always measured with respect to the origin, and so the hyperplane takes the form:

$$\begin{aligned} \phi(\mathbf{x}_i).\mathbf{w} + \mathbf{b} &= 1 \\ \therefore \phi(\mathbf{x}_i).\mathbf{w} - \rho &= 0 \end{aligned} \tag{4.4.13}$$

This changes the SVM optimisation problem to:

$$\min \frac{1}{2} \left\|\mathbf{w}\right\|^2 + C \sum_n \xi_n - \rho \tag{4.4.14}$$

subject to the constraints

$$\begin{aligned} (\mathbf{w}.\phi(\mathbf{x}_n)) &\geq \rho - \xi_n \\ \xi_i &\geq 0 \end{aligned} \tag{4.4.15}$$

This formulation can be solved using the same methods as before.

4.5. Conclusion

In this chapter, we have described the baseline data fusion algorithm for vital sign monitoring, showing firstly how a probabilistic model of normality is created using K-means and Parzen windows, and then how the resulting probabilities may be converted into a

Patient Status Index score. We then showed that there are theoretical problems with the algorithm that may result in an inaccurate model.

Two alternative methods have been proposed that address these problems. The first, weighted Parzen windows, is a simple modification of the baseline algorithm that applies priors over the Parzen windows prototype centres, and thus corrects for the outlier bias error that had previously been observed in Section 4.2. The second method is the SVM classifier, which was then extended for novelty detection using Scholkopf's one-class classifier. The next chapter is concerned with assessing how well each of these methods performs when tested on the continuous vital sign data acquired from the ED patients.

5. Application of the Data Fusion Models

5.1. Introduction

In Chapter 4, we introduced the baseline popular data fusion algorithm for calculating the Patient Status Index for vital sign data, which is a score used to facilitate in the assessment of a patient's physiological condition. This chapter is concerned with the application of weighted Parzen windows (wPw) and one-class Support Vector Machines (SVMs) to the same problem.

In the first instance, we will describe the training, validation and test data sets for the data fusion models. By examining the quality of the data, we will then determine whether it is appropriate to include all of the vital sign channels in the models. After this, we describe how each of the data fusion models were trained using the training set data, and then explain how any free parameters within the model were set by using the validation data set. Finally, we use the test data to assess how well each of the methods perform, using the analysis framework described in Chapter 2.5. As part of this analysis, we will ascertain how well the new models detect escalation events in comparison to the baseline model, and to the continuous T&T system developed in Chapter 3.

5.2. Data Sets

Training Data Set

Each of the data fusion models was trained using the vital sign data set that had been used to train the baseline model as described in Section 4.1.1. The distributions of each of the vital signs in the training set are reproduced in Figure 5.2.1 (for later comparison with the distributions of the validation and test data sets).

Validation Data Set

The one-class SVM does not require labelled data to make a classification, as long as there is a prior estimate of the percentage of data points in the abnormal class. However, Hayton et al. [44] note the accuracy of the classifier may be improved with some abnormal data. The values of the parameters were estimated simultaneously by maximising the accuracy, as calculated on a balanced validation subset.

The validation data set was collected from Phase I of a three-phase clinical study at the University of Pittsburg Medical Centre (UPMC) [49]. The data set consists of continuous observations from 333 patient admissions within a 24-bed Step Down Unit, for patients stepping down from the Intensive Care Unit (ICU). HR, RR, SpO_2 and temperature data were recorded using the Hewlett Packard "Viridia 24" bedside monitors, and the data were sampled approximately every 20 seconds. SBP and DBP were recorded every 30 minutes while the patient was awake and once every hour while they were asleep in order to minimise patient discomfort. In total, this provided 28,782 hours of vital sign data. Distributions of each of the vital signs for the study population are also shown in Figure 5.2.1, where the SDA has been computed as the arithmetic mean of the SBP and DBP values, so that the validation data set can be directly compared to the training data set.

To be useful as a validation set for one-class classification (or novelty detection), the data vectors must be labelled as "normal" or "abnormal" (for the few instances of physiological deterioration in the thousands of hours of data). Initially, abnormal data vectors were labelled using computer assistance; all continuous vital sign data that met the UPMC's Medical Emergency Team (MET) calling criteria were labelled as a C event (see Table 5.1

	Lower Limit	Upper Limit
HR (bpm)	40	140
RR (bpm)	8	36
SBP (mmHg)	80	200
DBP (mmHg)	-	110
SpO_2	80	-

Table 5.1.: Single-channel Medical Emergency Team criteria for the UPMC, used to generate C events

for MET criteria). The C events were then checked manually by two independent clinicians and any events that were considered to be non-artefactual and physiologically plausible were relabelled as C' events. The C' events were further evaluated to classify those with "serious, persistent, and generally displaying multiparameter abnormality" into a C" class. The C" events were initially classified independently by two critical care medicine clinicians. The inter-rater variability of this process was 61%, and a second review was required, in which the clinicians met together to reach a consensus classification.

In total 237 C' and 112 C" events were identified, which had a mean length of 24 minutes. We note that the C' and C" labels differ from the "escalation" events that were previously used to label the ED data set, in that the C events in the validation data set were derived from the continuous data, whereas escalation events were documented cases of clinical interventions.

In practice, the labelled data set should contain equal amount of "normal" and "abnormal" data so that the resulting model will not be biased towards having either a high sensitivity or high specificity. To achieve this, a subset of the validation data was created by first selecting all the data during the C' and C" events to be abnormal data vectors. An equal number of data vectors containing only normal physiology was created by randomly selecting vital sign data from any of the patients who did not have any C events. The size of the subset was thus 1.4×10^5 vectors of data, which corresponds to roughly 800 hours of data.

Test Data Set

The data fusion models were tested on the ED data set which was previously used to assess the performance of continuous T&T in Chapter 3. The set contains 1708 hours of data

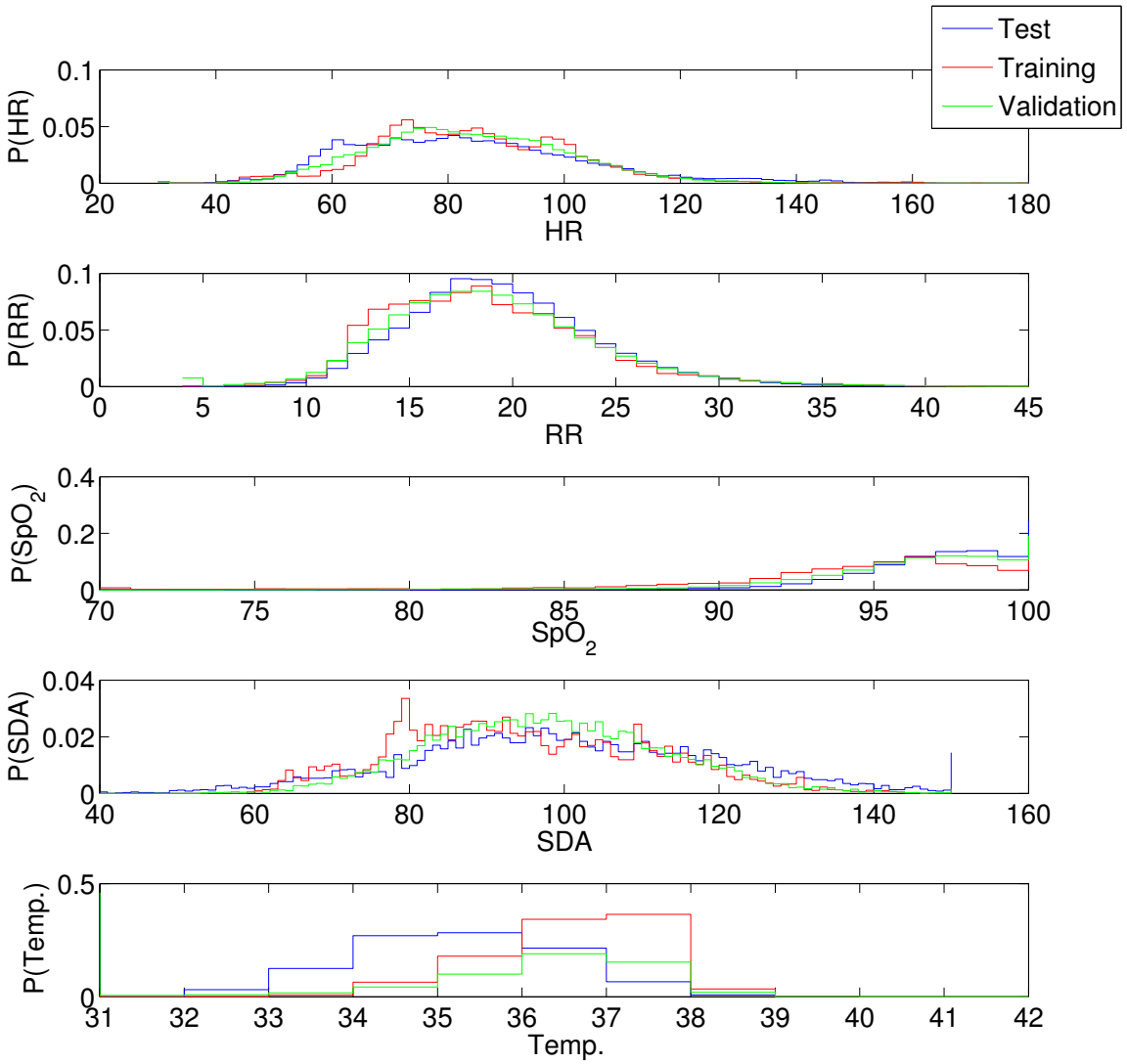

Figure 5.2.1.: Vital sign distributions for the training data set (red), the validation data set (green), and the test data set (blue)

recorded from 402 patients presenting to the John Radcliffe Hospital ED. A more detailed description of how the data set was recorded and labelled has already been provided in Section 2.1. The distributions of each of the vital signs for the test data set population are shown in Figure 5.2.1.

5.2.1. Data Set Summary

Each of the data sets was checked for physiologically implausible measurements. For instance, a quick examination of the training and validation data showed many instances during which the respiratory rate and heart rate were recorded as zero, when the ECG

Vital Sign	Lower Threshold	Upper Threshold
Heart Rate (bpm)	30	300
Resp. Rate (rpm)	3	45
SDA (mmHg)	20	180
SpO_2 (%)	10	101
Temp ($°C$)	32	39

Table 5.2.: Criteria used for removing spurious data in the Training and Validation data sets

	Training	Validation	Test
Location	JR	UPMC Step-down Unit	JR Emergency Dept.
Hours of Data	3500	28,782	1,708
No. Patients	150	333	476
Male/Female	50/50	58/42	52/48

Table 5.3.: Summary of the data sets

electrodes attached to the patient are likely to have become disconnected. The criteria used for removing spurious data are shown in Table 5.2. More modern monitors, such as those used for collecting the test data set, recognise when probes are disconnected, making the removal of spurious data unnecessary.

An overview of the training, validation and test sets, including information on the size of the data sets and the hospital locations, are given in Table 5.3. The histograms for each of the vital signs are shown in Figure 5.2.1 for the three data sets. A visual inspection of these graphs indicates that the shape of the vital sign distributions are similar for each of the data sets, apart from the temperature distribution. The modal temperature for the training set is approximately $37°C$, whereas the test set has a modal temperature of $35°C$. The shapes of the temperature distributions for the training and test sets also differ significantly. The difference in mean value between the training and test set temperatures is $\mu_{train} - \mu_{test} = 1.24$ (see Table 5.5), equivalent to approximately 1 standard deviation, as $\sigma_{train} = 1.26$ and $\sigma_{test} = 1.19$. In comparison, the difference in means for RR is 0.73 rpm, or approximately 0.14 standard deviations.

We can quantify the similarity of the data sets using some distance metric such as the Kullback-Leibler divergence. The K-L divergence can be considered as the relative entropy, that is the level of disorder, of a distribution P with respect to Q.

Formally, the Kullback-Leibler divergence is described as:

	HR	RR	SDA	Temp.	SpO_2
Training - Test	0.1915	0.0724	0.2453	1.5606	0.5240
Training - Validation	0.0809	0.0221	0.1538	0.2767	0.1983
Validation - Test	0.1171	0.0409	0.2419	0.9541	0.0918

Table 5.4.: Pairwise comparisons of the Kullback-Leibler distances for each of the vital signs, using the training, validation and test data sets described previously.

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{5.2.1}$$

However, this is non-symmetric measure such that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and instead we use the symmetrised divergence

$$D_{KL}(P||Q) + D_{kl}(Q||P) \tag{5.2.2}$$

The symmetrised divergence for each vital sign, for each pair of data sets, is shown in Table 5.4, and indicates that the test set is more similar to the validation set than to the training data set. The validation set is equally similar to both the test and training data sets.

	μ_{Train}	μ_{Valid}	μ_{Test}		σ_{Train}	σ_{Valid}	σ_{Test}
HR	83.77	83.22	83.17		17.48	16.60	20.83
RR	18.30	18.61	19.03		5.06	5.12	4.60
SDA	94.68	97.61	102.5		16.54	15.06	19.79
SpO_2	95.20	96.44	96.98		3.49	3.10	3.13
Temp.	36.05	35.87	34.81		1.26	1.24	1.19

Table 5.5.: Means and Standard deviations for the five vital signs recorded in the training (JR) and validation (UPMC) data sets

5.2.2. Removal of Temperature Recordings

In our previous investigation of continuous T&T in Chapter 4, we disregarded the temperature recordings because the T&T temperature criteria were designed to be used with

core temperature measurements rather than the skin temperature that had been recorded by the bedside monitors. In each of the data fusion models presented in Chapter 4, the classification of abnormal measurements is no longer based on a clinical heuristic, but is instead derived directly from the data set. Therefore, we are no longer constrained to using one type of temperature measurement, and it should be possible to incorporate skin temperature recordings.

In the end, we decided not to include temperature recordings in the data fusion models for two reasons. Firstly, as noted in the previous section, the training and test set temperature distributions are very different. Therefore, a model based on the temperature data from the training set is unlikely to be able to discriminate well between normal and abnormal temperatures in the test set.

Secondly, we observed a high level of temperature data loss from the validation and test data sets. Discussions with the clinical teams that collected the validation and test data sets led to the observation that the thermistors tended to detach in the clinical setting, leading to unreliable and missing data. For instance, 48% of the temperatures recorded in the validation data set were outside the expected physiological limits of 32 and $39^{\circ}C$ for skin temperature, as determined by clinical experts during the UPMC study. Similarly, the data loss for skin temperature in the test data set is 74%, in comparison to the data losses for the other variables, which were around 25% (see Table 3.1).

By limiting the data fusion models to the four vital sign parameters (HR, RR, SpO_2 and SDA), we also have the further benefit that we are able to compare the results from the data fusion models to those obtained with continuous T&T.

5.3. Implementation of Data Fusion Models

5.3.1. Baseline Parzen Windows Model

The five vital sign baseline model was trained according to the procedure outlined in Section 4.1. The model can be applied to the appropriate subset of four vital signs by setting the temperature to the mean value of the training set (that is, a value of zero after vital sign normalisation).

The baseline model contains three model parameters: the number of Parzen window centres, the Parzen window kernel width, and the alerting threshold on the resulting probability density function. Of these, none were set using the validation data. The initial number of Parzen window centres was set at 500, to reduce the size of the data set by a factor of approximately 4000. The alerting threshold was set heuristically at $PSI = 3.0$, on the basis that this score corresponds to single-channel values of ± 3 standard deviations away from the mean for any of the vital signs (Section 4.1.6).

The kernel width, which is used to define the size of the kernel in Equation 4.1.3, was set at $h = 1.49$, using the approach recommended by Bishop [11].

5.3.2. Weighted Parzen Windows

The wPw model follows the same training procedure as the baseline model and has the same free parameters. However, unlike the baseline model, temperature data were not included so that a 4-dimensional Parzen windows distribution was created. The number of Parzen windows centres was again set at 400 centres, in keeping with the baseline model. However, we now consider two methods of generating the 400 centres.

In the baseline model, the 400 prototype centres were identified by first applying the K-means algorithm to cluster the training data into 500 prototype centres. Following this, 100 prototype centres, broadly corresponding to clusters of abnormal vital sign data, were removed. The same procedure was applied in the wPw training procedure, again using K-means to generate 500 prototype centres. The relative weights of each centre were computed by recording the membership of each centre. This was achieved by adapting the Netlab [76] implementation of the K-means algorithm (*kmeans.m*), which also extends its ability to deal with very large data sets. However, unlike with the baseline model, we then considered two methods of pruning the 100 most "abnormal" centres to derive the final set of 400 prototype centres.

The first method, which produces a model we define as wPw_{dist}, involves removing 100 prototype centres using the same criterion as for the baseline model: the 100 prototype centres with the greatest Euclidean distance from the population mean were removed. In Section 4.2, we postulated that outlier removal should be based instead on cluster

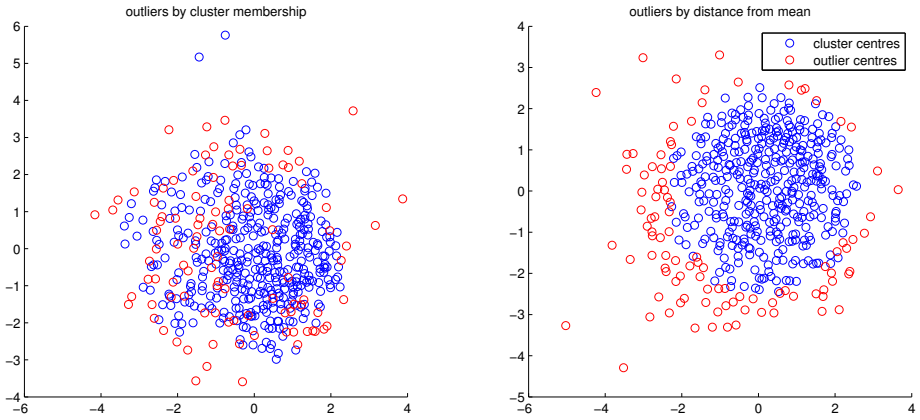

Figure 5.3.1.: Representation of the 400 wPw kernels and 100 outlier kernels using SASS visualisation maps where (a) kernels are selected to be the 400 k-means centres with the largest membership (b) kernels are selected to be the 400 k-means centres with the smallest Euclidean distance to the mean [0 0 0 0].

population. Clusters with few patterns belonging to them denote the areas in data space with lower probabilities, and hence the vital sign data in these sparsely populated regions are likely to be abnormal.

We therefore considered an alternative method of removing 100 prototype centres, by removing the 100 centres with the lowest K-means cluster populations. The model that resulted from this procedure is defined as wPw_{pop}.

The 400 prototype centres and the 100 pruned centres are shown for both methods in Figure 5.3.1, using the SASS visualisation map. Although the maps show the same 500 centres, the visualisations appear slightly different as there are many possible equally-valid visualisations when data dimensionality is reduced, as described by Nabney [76]. In both cases, the 400 centres and their associated cluster memberships were used to generate a probability density function using wPw according to equation 4.3.1.

The kernel width parameter was again set using the method described in Equation 4.1.5, this time giving values of $k = 0.56$ and $k = 0.46$ for wPw_{pop} and wPw_{dist} respectively. The kernel widths are substantially smaller than for the baseline model because the vital sign vectors are of a lower dimensionality. By sampling from the resulting wPw distributions the offset $p_{max}(x)$ that is required to generate a positive Patient Status Index (PSI) in all cases was determined to have a value of 4.30 for both models.

The final free parameter, the alerting threshold, was set using the same method as

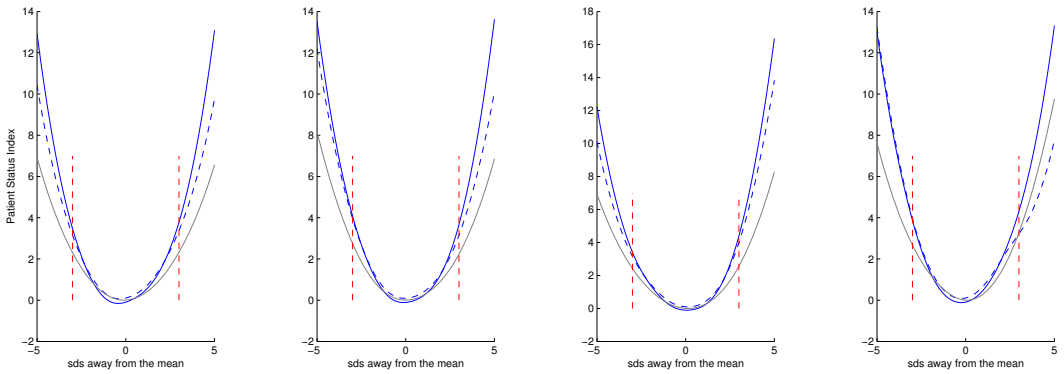

Figure 5.3.2.: Patient Status Indices for the cases where HR, BR, SpO_2 and SDA are varied while remaining variables are held at their mean value. 3sds away from the mean are marked for each parameter.

for the baseline algorithm. 1D slices through the wPw_{pop} and wPw_{dist} models, which are analogous to the slices in Figure 4.1.5, are shown in blue in Figure 5.3.2 for the case where one vital sign variable is varied between ±4s.d.s while the remaining variables are set to their mean value. For comparison, the corresponding slices for the baseline model are also plotted in green. The points at which any of the single variables deviate by ±3s.d.s from the mean is also marked on the diagram in red. An alert which would trigger when any single variable deviated by 3s.d.s from the mean requires a PSI threshold between 3.4 and 4.0 in both the wPw_{pop} and wPw_{dist} models. The alerting threshold was therefore set between these two bounds, at $PSI = 3.7$, for both of the models.

So that the wPw models can be fairly compared to the baseline model, we continued to use the method described in Chapter 4 for dealing with missing data, which involved first setting a missing channel to the local median after 5 minutes, and then to the population mean after 30 minutes of missing data. Furthermore, the persistence criterion for generating PSI alerts, as described in Section 4.1.6 for the baseline model, was also used when generating alerts in the wPw models.

5.3.3. Support Vector Machines

The one-class SVM technique was applied to the JR training data set using the LIBSVM [18] toolbox for Matlab. We recall from Section 4.4.4 that the SVM problem can be defined as an optimisation problem that is dependent on two model parameters, C (Equation 4.4.14), the slack variable weighting, and γ, the kernel size (Equation 4.4.12).

The parameter C is often rewritten in terms of the number of data points, l, so that:

$$C = \frac{1}{l\nu} \tag{5.3.1}$$

An optimal value for the accuracy was ensured by employing a grid-search over the two parameters. The accuracy is defined as:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.3.2}$$

and can be viewed as a summary of the overall classification rate. The True Postives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) were all calculated on a sample-to-sample basis. The true, false, negative and positives were defined as follows:

- True Positive - Any "abnormal" validation data vector that is correctly classified as being outside of the decision boundary by the SVM (i.e. an outlier)
- True Negative - Any "normal" validation data vector that is correctly classified as being inside the decision boundary by the SVM
- False Positive - Any "normal" validation data vector that is incorrectly classified as being outside the decision boundary by the SVM (i.e. an outlier)
- False Negative - Any "abnormal" validation data vector that is incorrectly classified as being inside the decision boundary by the SVM.

An abnormal data vector was defined as any vector of four vital sign measurements ([HR RR SpO_2 SDA]) that was recorded during a C' or C" event. Conversely, a normal data vector was defined as any vector of four vital signs that was recorded from a patient with no C, C' or C" events. The validation subset contained an equal number of abnormal and normal vectors so that there was no bias that could affect the accuracy.

We note that TPs, TNs, FPs and FNs were calculated using a different method than the one outlined in the analysis framework in Chapter 2.5. The reason for this is twofold. Firstly, the outcome marker for the validation set is different from the outcome marker in the test set, and the two cannot be used interchangeably. The C, C' and C" events were

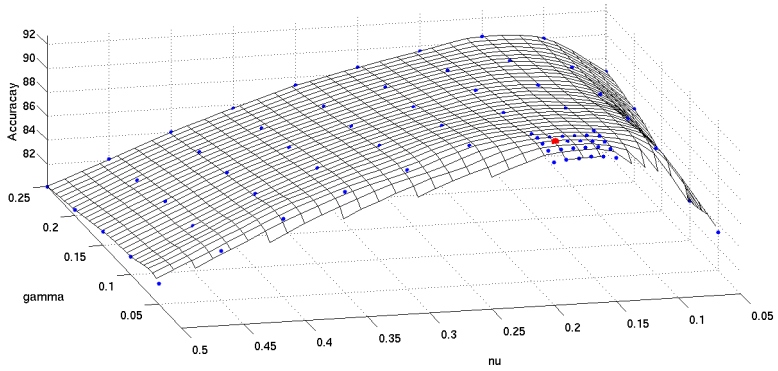

Figure 5.3.3.: A plot of the SVM accuracy, as calculated on the validation data subset. Grid search was performed at the points in blue. The optimum point (in red) was chosen. Interpolation between points shows that optimisation is a smooth function in the limit (of infinite validation data)

based on retrospective analysis of whether or not vital signs met the MET-calling criteria, and are not the same as an "escalation" event, which is a clinical intervention documented at the time of the event and which may have both physiological and non-physiological causes.

Secondly, the analysis framework in Chapter 2.5 was designed to assess the clinical benefit of a system by determining whether *patients* were identified promptly. In this validation step, we do not optimise the system as a whole, but only the model used for classification. A classification is made for each vital sign vector, therefore it is appropriate that the completion of accuracy that is used to optimise the model should be performed on a per-vital sign vector basis.

The accuracy was calculated over the parameter ranges $0.05 \leq \nu \leq 0.50$ and $0.05 \leq \gamma \leq 0.25$. From these results, a more refined search in the range $0.15 \leq \nu \leq 0.20$ and $0.01 \leq \gamma \leq 0.05$ was carried out to improve the estimate of accuracy. The result of the grid search is shown graphically in Figure 5.3.3, where the optimal parameter pair, ($\nu = 0.19$,$\gamma = 0.03$) is highlighted in red.

Alerts were generated using the binary output of the SVM, and the persistence criterion. Alerts were generated if the SVM classified the vital sign data as being abnormal for 4 minutes out of any 5-minute window and alerts were turned off once data had returned to a normal state for 2 minutes out of any 3-minute window. Alerts were also generated immediately whenever a blood pressure reading resulted in the data being classified as

abnormal.

One further heuristic was introduced for the SVM model. Any SpO_2 measurements that were greater than the mean of the training data were replaced by that mean. For instance, an SpO_2 measurement of 99% would be replaced by a value of 95.2%. The heuristic was required because the SVM model correctly determines that SpO_2 values above 100% are extremely unlikely (in fact, impossible). Therefore, the SVM decision boundary is close to 100% saturation, and vital sign vectors that include values of 100% for the SpO_2 measurement could erroneously be classified as abnormal.

5.4. Evaluation of Data Fusion Models

All the data fusion models were applied to the ED test data set. In the section that follows, we firstly demonstrate the output from each of the models on a small subset of patients. Following this, we quantitatively assess the performance of each model, using the analysis framework previously developed for evaluating continuous T&T in Section 3. By using the same metrics, we may compare the results from the data fusion models directly with the corresponding results from the continuous T&T system.

5.4.1. Examples of the Data Fusion Systems

The data fusion systems were first applied to the test data from two example patients. Figure 5.4.1 shows the vital signs and scores from the baseline model, wPw models, and SVM models for a 81-year old female patient who had attended the department with chest pains and shortness of breath, and had been unwell for the previous week. A physiological escalation at 14:05, 20 minutes after arrival, caused by a high heart rate due to atrial fibrillation and a high respiratory rate, is marked on the figure in red. The patient's vital signs are recorded in the upper graph, and the four data fusion models are shown below on separate axes. In this example, where two of the parameters are grossly abnormal, each of the methods correctly assigns an alert (denoted by a grey background) at the time of the escalation. We note that in general, the weighted Parzen windows models have a similar overall behaviour to the baseline model, and a larger dynamic range. The SVM does not provide a score, but instead classifies the vital signs as either normal (SVM=+1),

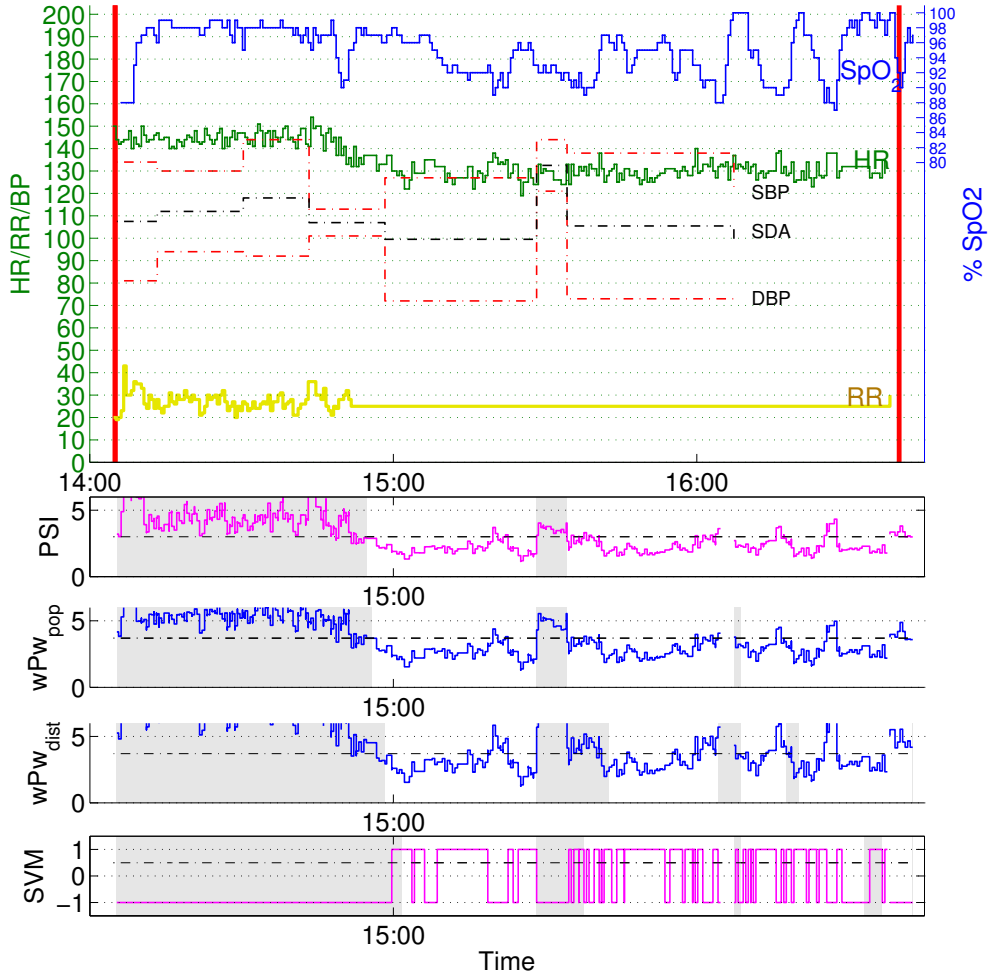

Figure 5.4.1.: PSI scores for patient ED00112 derived using the baseline model, weighted Parzen windows, and support vector machines. The patient deteriorates at the start of the recording, and continues to remain unstable throughout their stay on the bed. The patient was escalated a second time at the end of the record, due to low SpO_2 and continuing atrial fibrillation. The greyed-out areas represents occasions on which each data fusion system generates an alert.

or abnormal (SVM=-1).

Figure 5.4.2 shows the vital signs and data fusion model scores for a 64-year old male patient who was brought to the Emergency Department by ambulance, and had presented with breathing problems, as indicated by initial heart and breathing rates that were extremely high. In addition to this, the patient was hypotensive. At the time of arrival, the patient was extremely unwell, which we deduced from his "red" triage category - the most severe class.

The patient was initially admitted to Resus, after which he was stabilised before being

admitted to a general medical ward for septic shock caused by metastatic rectal cancer. The record shows how the heart rate decreases from 160 bpm to 103 bpm, and how the breathing rate decreases from 34 rpm to a more regular value of 19 rpm. The SpO_2 values also increase throughout the record, which denotes better blood oxygenation. However, the SpO_2 trend should be treated with scepticism, as the low sampling rate may be indicative of intermittent probe disconnection. The blood pressure values also remain abnormally low throughout the patient's stay on Resus.

The improvement in the patient's physiological condition is identified by all four of the data fusion models. At the start of the record, each of the models generates an alert due to the initial abnormal physiology. The PSI and wPw values then decrease until they are below the alerting threshold. Similarly, the SVM switches from an abnormal to a normal state. Over the course of the record, there are brief period of high abnormality, due to momentary measurements of low SpO_2. However, data loss occurs in these cases before the persistence criterion will allow an alert to be generated.

5.4.2. Sensitivity and Specificity of the Data Fusion Models

True Positives

In Chapter 3, an escalation event was identified if the continuous T&T alert was active within some given time window of an escalation event. The method is repeated in this section, again considering only the type-2 escalations caused by BP, RR or HR. The number of escalation events identified by each of the data fusion systems over the range of windows is shown in Figure 5.4.3(a), and the number of patients identified is shown in Figure 5.4.3(b).

As before, the number of True Positives was calculated as the number of patients identified within a window defined by $t = 10$ and $\tau = 10$. The True Positives, along with the False Negatives, are summarised in Table 5.6 for each of the data fusion models. The table indicates that the SVM system identifies the greatest number of True Positives, 17, while both of the weighted Parzen windows schemes perform better than the baseline model, which detects only 12 escalation events. We also note that all of the methods appear to be less sensitive than the continuous T&T system with and without a persistence

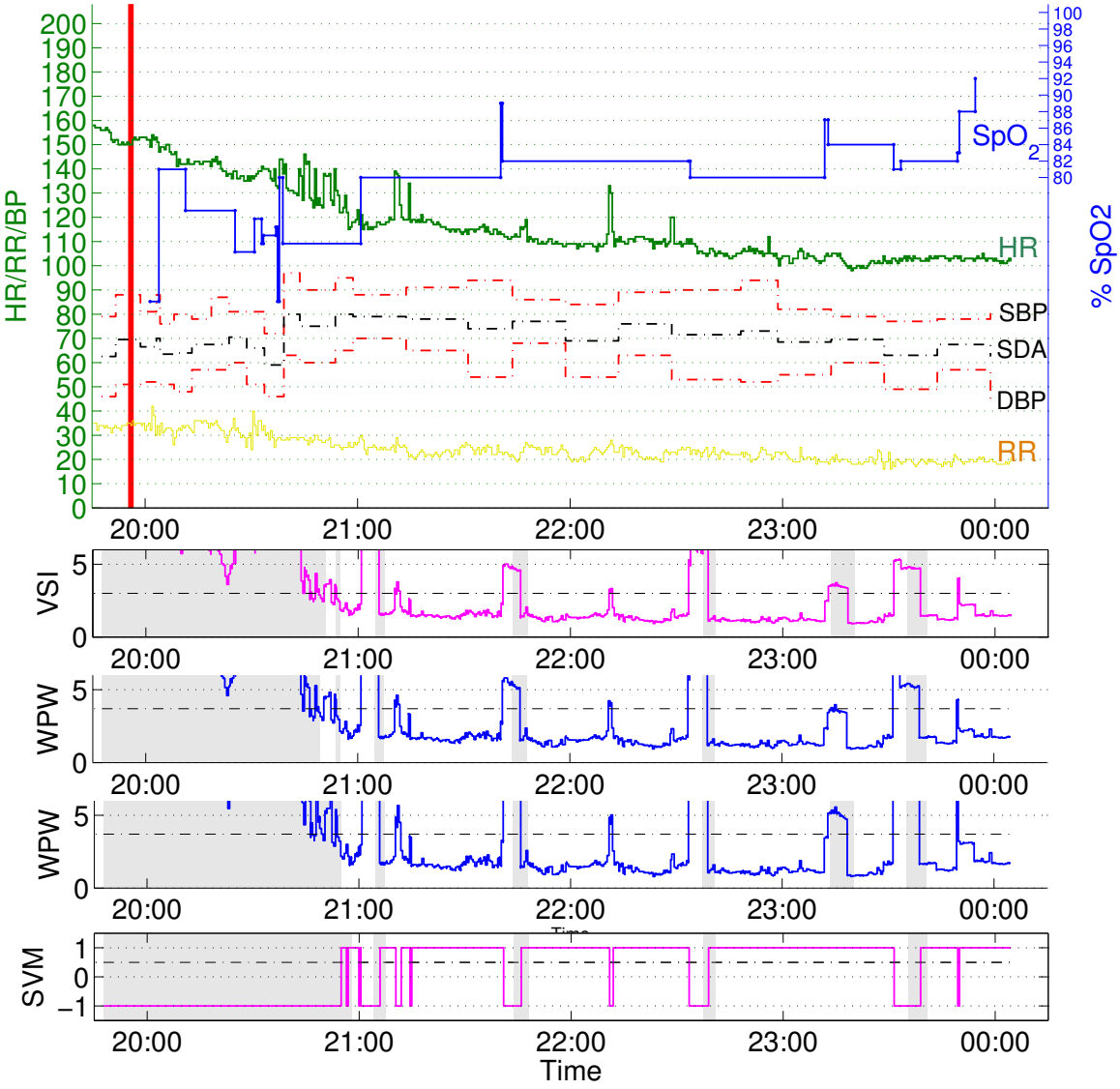

Figure 5.4.2.: PSI scores for patient ED00263 derived using the baseline model, weighted Parzen windows, and support vector machines. The patient stabilises during their stay at the ED, as indicated by the reduction in alerts over the recording period. The greyed-out areas represents occasions on which each data fusion system generates an alert.

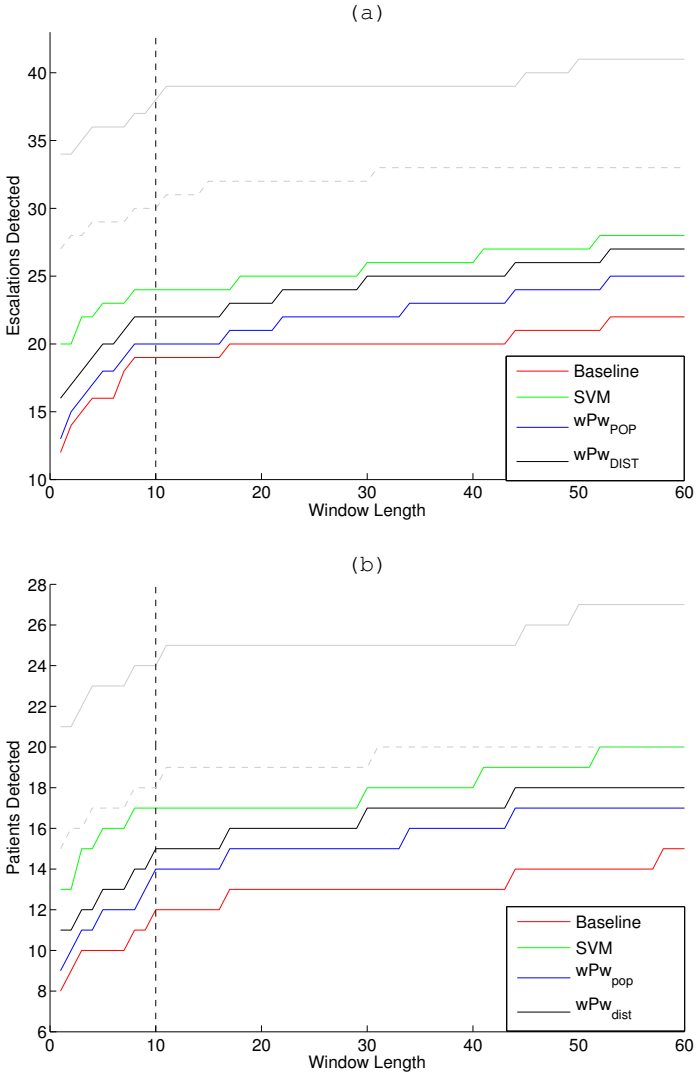

Figure 5.4.3.: The true positive rate for the baseline data fusion model, weighted Parzen windows, and support vector machines. The true positive rate for continuous T&T is also shown in solid grey (without a persistence criterion) and dashed grey (with a persistence criterion).

	Baseline Model	wPw_{pop}	wPw_{dist}	SVM		Continuous T&T
True Positive (first escalation detected)	12	14	15	17		24
False Negative (first escalation not detected)	17	15	14	12		7

Table 5.6.: Summary of the true positives and false negatives for all patients with at least one physiological escalation event that occurred after arrival to the ED, for each of the data fusion methods. The results for continuous T&T have been included for comparison.

criterion, which identified 19 and 24 true positives, respectively.

In total, 9 patients were classified differently to continuous T&T without a persistence criterion. 8 True Positives were detected by continuous T&T, but not the SVM model, and 1 True Positive was detected by the SVM model, but not by continuous T&T. We determined the cause of the differences between the SVM and continuous T&T models by manually reviewing the continuous vital sign data for the patients that were classified differently by the two models.

In 5 cases (Patients ED00031, ED00066, ED00196, ED00320, ED00478), the escalation was due to abnormal systolic blood pressure (SBP). These were not detected in the SVM model (or Parzen windows models) as SBP was not directly included as one of the input parameters and instead, the Systolic-Diastolic Average blood pressure had been used.

The problem with using the SDA can be most clearly seen for the example in Figure 5.4.4. In this example, the SBP fluctuates between 208 mmHg and 148 mmHg. This is statistically significant, denoting a change from the 99.99^{th} centile to the 85.57^{th} centile when compared to the training data. Physiologically, an SBP of 208mmHg indicates extreme hypertension that may require immediate intervention, whereas 148mmHg indicates slightly elevated blood pressure which would cause no clinical concern. Despite the drastic range in the SBP, the SDA, shown in black, only fluctuates between 130 and 120 mmHg, which is equivalent to the 98.10^{th} and 93.66^{th} centiles in the training data, a much smaller change. In this case, the change in SDA is small because the increase in Diastolic Blood Pressure (DBP) occurs at a similar rate to the decrease in SBP.

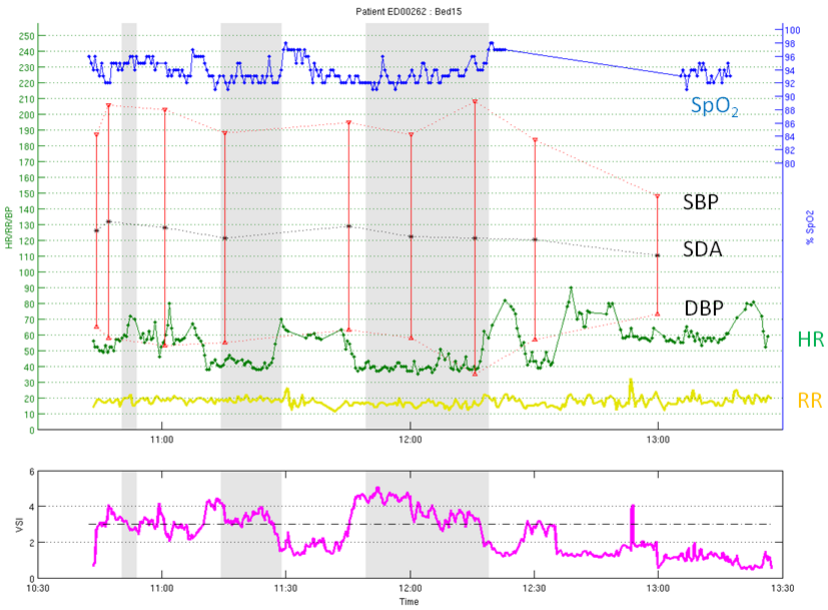

Figure 5.4.4.: Plot of vital signs and PSI score for patient ED000262, a 73-year old male who was admitted to the hospital having being assessed to have symptomatic bradycardia. Again, the greyed-out areas represents occasions on which each data fusion system generates an alert. The patient was initially in majors, and a clinical review of the patient was only prompted (A2 escalation) at 16:00. Only one set of manual observations in majors was taken at 11:00.

The escalation for patient ED00521 was due to the SpO_2 saturation dropping below 90%. None of the data fusion algorithms were sensitive enough to identify these escalations. However, in one of the cases, patient ED00536, an earlier alert would have been generated by each of the data fusion algorithms due to a more prolonged and severe desaturation that occurred 20 minutes before the escalation event. In this instance, a continuous monitoring system may have provided early warning of the escalation event, and it is only due to the limitations of our analysis framework that the patient is not considered to be a true positive.

There were two patients that were identified using continuous T&T, that we would not expect to detect using the data fusion systems. Patient ED00210 experienced momentary vital sign abnormality, but a nurse was on hand to repeat the measurement before the persistence criterion was met, and an alert was generated. Patient ED00077 had an escalation event caused by sustained hypotension which was identified by continuous T&T, but could not be identified by the data fusion systems as more than 2 channels of data

	Baseline Model	wPw_{pop}	wPw_{dist}	SVM		Continuous T&T
True Negative (Zero Alerts)	173	181	171	159		49
False Positive (≥ 1 Alerts)	44	36	46	58		168

Table 5.7.: Summary of the true negative/false positive rate for patients with no escalation events, for each of the data fusion methods. The results for continuous T&T have been included for comparison.

were missing (see Section 4.1.5) at the time of the escalation.

The single case for which the SVM model correctly classified an escalation, but continuous T&T did not, was for Patient ED00139. In this instance, a combination of slightly abnormal blood pressure and heart rate leads to a classification that persists long enough for an alert to be generated. In contrast, the continuous T&T score detects the abnormality, but for a shorter length of time than 4 minutes (out of 5) and so no alert was generated.

True Negatives

As in Section 2.5, a True Negative event was defined over the set of 217 patients who had no escalations during their time in the ED and who had some continuous vital sign data. Any of these patients for which the data fusion system correctly produced no alerts during their stay were classed as True Negatives, otherwise the patient was considered to be a False Positive. The number of true negatives in each of the systems is shown in Table 5.7.

Using these figures, the sensitivity and specificity of each of the models can be calculated using Equations 1.3.1 and 1.3.2, giving the following values:

$$sens_{baseline} = \frac{12}{12+17} = 41.3\% \qquad spec_{baseline} = \frac{173}{173+34} = 79.7\%$$
$$sens_{wpw\_pop} = \frac{14}{14+15} = 48.3\% \qquad spec_{wpw\_pop} = \frac{181}{181+36} = 83.4\%$$
$$sens_{wpw\_dist} = \frac{15}{15+14} = 51.7\% \qquad spec_{wpw\_dist} = \frac{171}{171+46} = 78.8\%$$
$$sens_{SVM} = \frac{17}{17+12} = 58.6\% \qquad spec_{SVM} = \frac{159}{58+159} = 73.3\%$$

in comparison, the sensitivity and specificity for continuous T&T with and without a persistence criterion are:

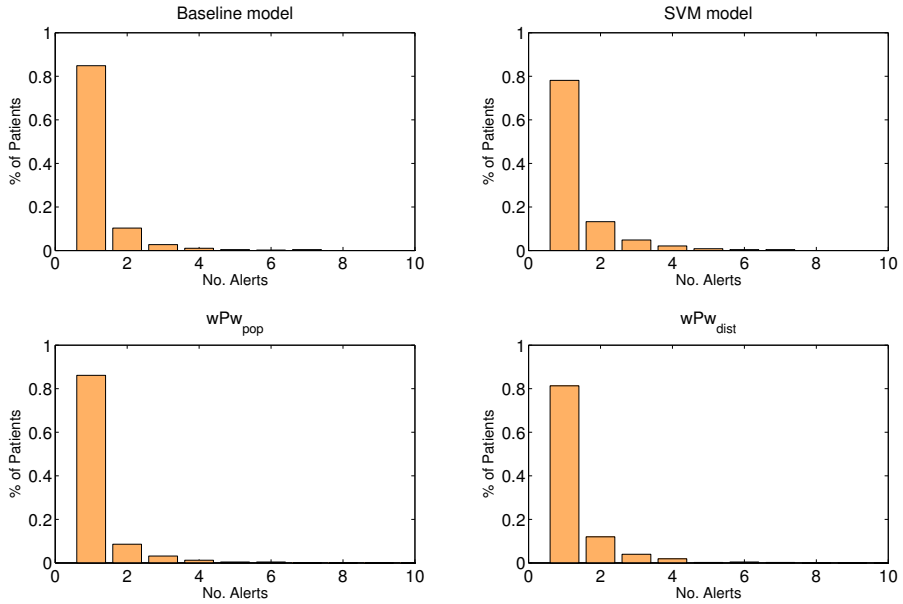

Figure 5.4.5.: Distributions of the number of alerts per patient for the four data fusion systems.

$$sens_{T\&T} = \frac{24}{24+5} = 82.7\% \quad spec_{T\&T} = \frac{49}{49+168} = 22.6\%$$
$$sens_{T\&T\,persist} = \frac{19}{19+10} = 65.5\% \quad spec_{T\&T\,persist} = \frac{74}{74+143} = 34.1\%$$

Alerts Per Patient

Figure 5.4.5 show the distribution of the number of alerts per patient, and Figure 5.4.6 shows the distribution of time spent in the alert state for each data fusion system. From these graphs, we can infer that the SVM model generates a greater number of alerts than the other three models, and also generates alerts for longer periods of time.

We can calculate the "alert rate" by dividing the total bed-time by the number of alerts (i.e. occasions on which the T&T score exceeds the alert thresholds for a particular patient group). For the entire study population, 1708.4 hours of data were recorded, and the baseline model generated 316 alerts. This gives an estimate of 0.18 alerts/hour per bed. Similar calculations were completed for each of the models, and the results are summarised in Table 5.8. The false alert rate was calculated using the method described in Section 3.2.2, and the results are also shown in Table 5.8.

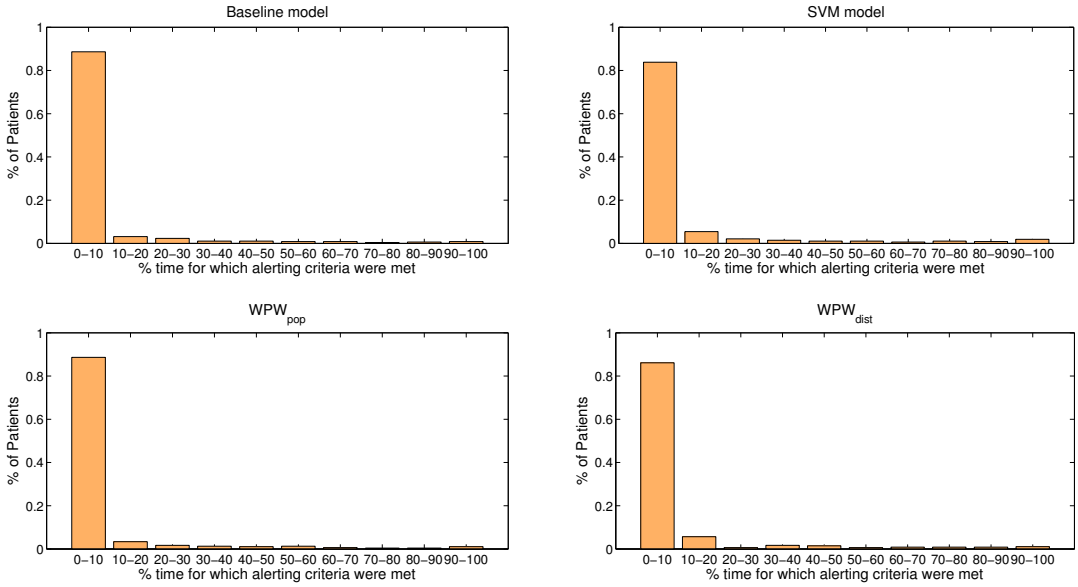

Figure 5.4.6.: Distributions over the duration of time spent in the alert state per patient for the four data fusion systems.

Model	Alert Rate	False Alert Rate
Baseline	0.18	0.055
wPw_{dist}	0.23	0.083
wPw_{pop}	0.18	0.057
SVM	0.28	0.125

Table 5.8.: Alert rate and false alert rate for the data fusion models

5.5. Discussion

In this chapter, we have described how four data fusion models: a baseline Parzen windows model, two weighted Parzen windows models wPw_{pop}, wPw_{dist} and an SVM model, were trained on previously acquired data and then tested on the continuous data recorded during the ED study. Results in Section 5.4.2 showed that the SVM model had the highest sensitivity and the lowest specificity of the four models tested, whereas the baseline model had the lowest sensitivity and highest specificity. Without prior knowledge of the relative importance of True positive and False positive events, it is impossible to make an overall comparison as to which of the models performed best. Instead, we can compare the results obtained with the data fusion models with the results generated using the continuous T&T system, using the sensitivity and specificity metrics.

In comparison to the continuous T&T system developed in Chapter 3, the data fusion models tested in this chapter had much higher specificities. This was confirmed by the low false alert rate, 0.05 alerts/hour per bed for the baseline model and 0.125 alerts/hour per bed for the SVM model, which correspond to 1 and 2.5 false alerts/hour respectively, on a typical 20-bed ward. In comparison, continuous T&T (even with the persistence criterion) produced 0.57 alerts/hour per bed, or 11 alerts/ward hour.

Whilst the data fusion models had a high specificity, the models' sensitivities were sub-optimal. The most sensitive data fusion model, the SVM, detected 8 fewer escalations than the most sensitive of all the models, continuous T&T without a persistence criterion. In Section 5.4.2, we showed that 5 of the 8 missed positive events were due to abnormal blood pressures.

The one other event that we would expect to detect was missed, due to the SVM model's insensitivity to low oxygen saturation. In the case of the SVM model, this was due to the method by which C events were generated (in the validation data set). The SpO_2 criterion for a C event was $SpO_2 \leq 80\%$ and consequently the SVM decision boundary along the SpO_2 direction in the input data space was close to 80%. Therefore, the SVM model could only detect very severe oxygen desaturation. The Parzen windows and wPw models also missed escalation events that were primarily due to low oxygen saturation, even though they did not use the C events for validation. In this case, the poor sensitivity

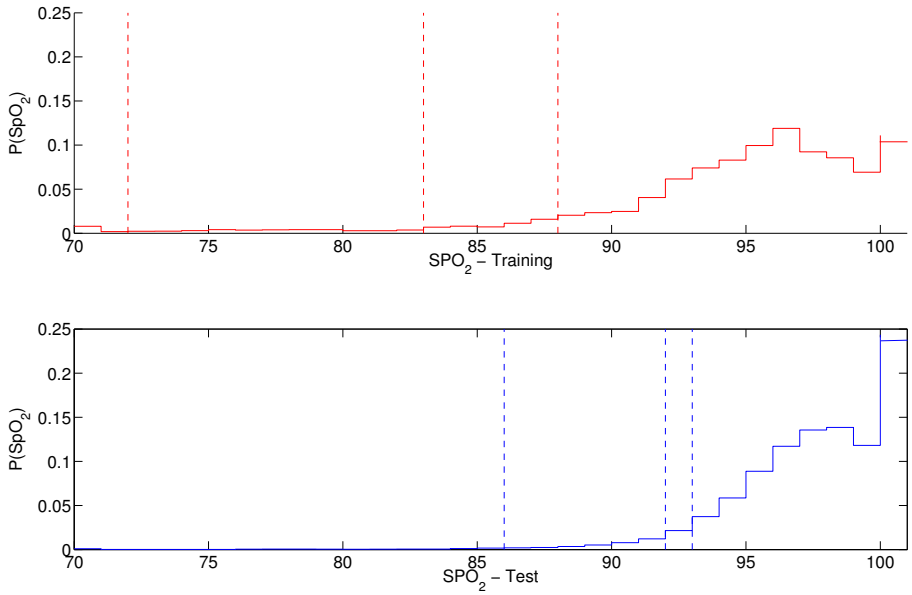

Figure 5.5.1.: The 1^{st}, 5^{th} and 10^{th} percentiles of SpO_2 for the training and test set distributions. The training set has a much wider tail, denoted by the fact that the centiles are situated at SpO_2 values of 72%, 83% and 88%. In comparison, the equivalent centiles in the test set distribution are 86%, 92% and 93%.

was due to differences between the training data and the test data in the tails of the SpO_2 distributions. This can be seen more clearly in Figure 5.5.1, which plots the 1^{st}, 5^{th} and 10^{th} percentiles for the training and test set distributions of SpO_2 values.

The sensitivity and specificity of the models also depends on the selection of the model parameters. In this chapter, the PSI alerting threshold for the baseline model and the wPw models were set using the method first described by Tarassenko et al. [110]. This method assumes that it is appropriate to generate an alert if a single vital sign parameter is close to ± 3 standard deviations from its mean value in the training set. While this is a sensible heuristic, it is unlikely that the resulting alert threshold is optimised for detecting escalation events. A more principled method would be to set the alert threshold based directly on the number of escalation events detected and false alerts generated, for instance, using Receiver-Operating Characteristic (ROC) analysis.

5.5.1. Model Retraining

In the previous section, we observed that the simple arithmetic mean of SBP and DBP was not appropriate, as it had no clinical or valid theoretical benefit. We now briefly consider the effect of modifying an SVM model to include SBP and DBP separately. We note that this approach will give a stronger preference towards blood pressure related conditions. A model using the Mean Arterial Pressure (MAP) may provide a better model, as the MAP is directly measured, whereas SBP and DBP are inferred heuristically. Unfortunately, MAP measurements were not saved in the training data set.

SBP and DBP were first normalised following the procedure described in Section 4.1.1. The SBP and DBP means were 125.38mmHg and 63.82mmHg, and the standard deviations were 21.39mmHg and 13.4mmHg respectively. A 5-D SVM was trained, using the same approach as that outlined in Section 5.3.3. The optimal values of the model parameters were $\nu = 0.1$ and $\gamma = 0.1$. The search space is visualised in Figure 5.5.2.

For the modified SVM model, the number of detected A2 events, and first A2 events per patient are shown in Figure 5.5.3 for window lengths between 1 and 60 minutes. The plots show that the modified model is worse at identifying escalation events than the original SVM model, detecting 13 A2 events, though we note that the result is not statistically significant. The modified model correctly detects 163/217 True Negatives, so the sensitivity and specificity can be calculated as 44.8% and 75.1% respectively.

The poorer performance of the retrained model could be due, in part, to the greater emphasis that the retrained model places on blood pressure parameters (two blood pressure measurements, SBP and DBP, instead of the one, SDA, used previously). More significantly, the escalation events were determined using the patient notes. These included T&T score data, which were based on SBP only. It is likely, therefore, that the addition of DBP does not provide much additional value to the model.

Other shortcomings of the data fusion models could be addressed by retraining on a data set that is more "similar" to the ED data set, so that any differences between the training and test vital sign distributions, as we previously saw for SpO_2 in Figure 5.5.1,

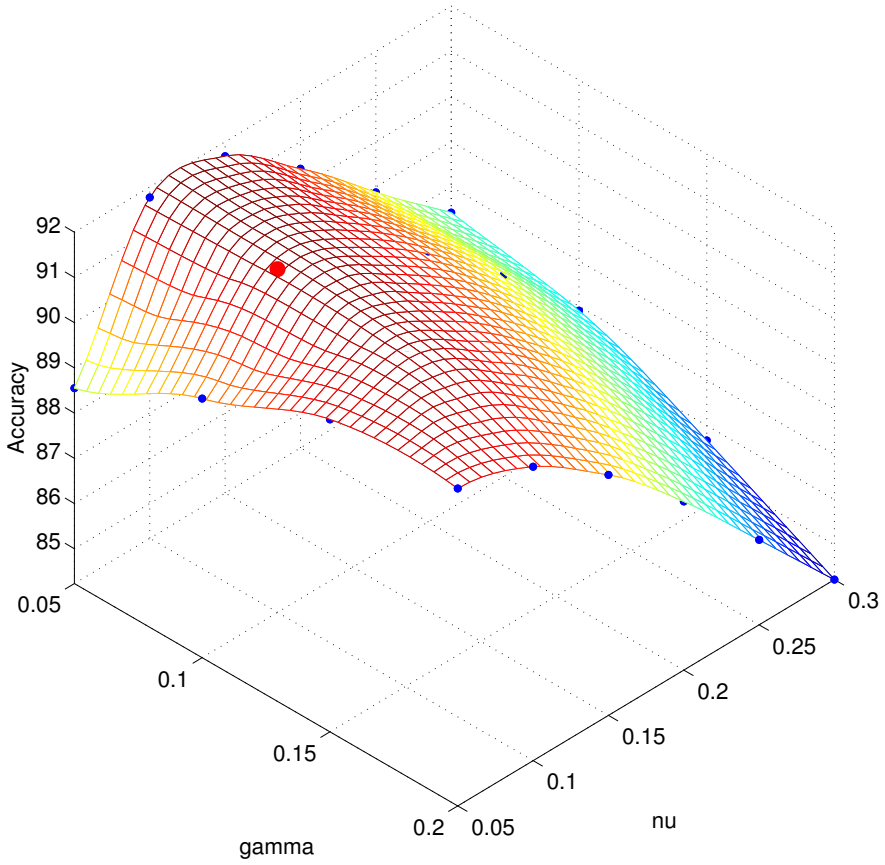

Figure 5.5.2.: A plot of the SVM accuracy for the retrained model, as calculated on the validation data subset. Grid search was performed at the points in blue. The optimum point (in red) was chosen.

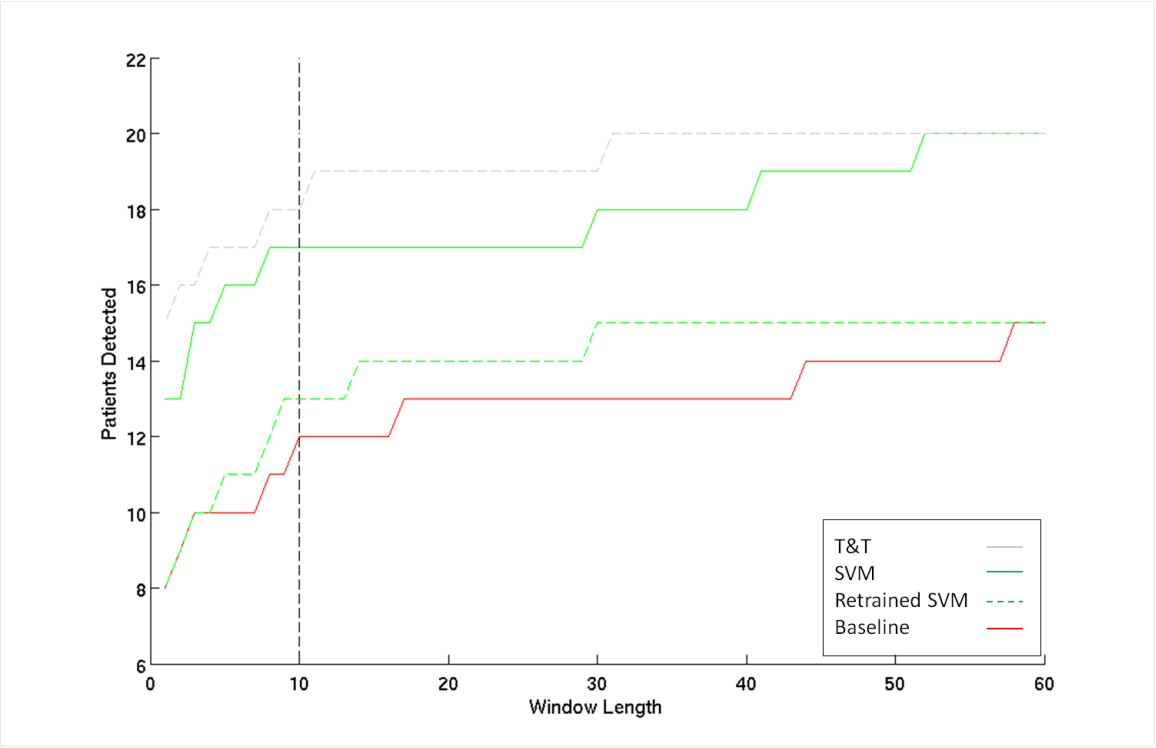

Figure 5.5.3.: The true positive rate for the baseline data fusion model, support vector machine model, and retrained support vector machine (using SBP and DBP). The true positive rate for continuous T&T is also shown in dashed grey.

will be minimised. During retraining, the alerting threshold would be set based on how well the model identifies escalation events, rather than on the statistical method used in Section 4.1.6.

Typically, this analysis would be accomplished using ROC analysis. With this approach, the alerting threshold is evaluated at a range of PSI values, and the resulting sensitivity and specificity, as determined on a subset of the training data, are recorded using an ROC plot. The resulting ROC curve then allows the optimal alerting threshold to be set, based on either a.) a minimum allowable sensitivity, b.) a minimum allowable specificity, or c.) the ratio between true positives and false positives, as shown in Figure 1.3.2. In this application, it is difficult to estimate the cost ratio (True Positives and False Positives), and so a.) or b.) must be used.

One method of ensuring similarity between the training, validation, and test sets is to generate the data sets directly from the ED data, using a cross-validation scheme to ensure that the data fusion models are not over-trained. We considered whether this would be feasible using two possible cross-validation schemes. In the first, *repeated random sub-*

Figure 5.5.4.: Pictorial representation of: a.) Repeated random sub-sampling validation, b.) 5-fold cross-validation. In each case, patients with A2 escalations are shown in red, patients with no escalations are shown in green, and all other patients are shown in orange. In a.) the A2 events are selected randomly for each retraining. In b.) the retraining is repeated 5 times, using the patient in F_n on the n^{th} retraining

sampling validation, a balanced test set is generated by randomly selecting half of the A2 patients and then randomly selecting the same number of patients with no escalations. Similarly, a balanced validation set is created using the remaining A2 patients, along with randomly selected patients with no escalations. The remaining patients are used as a training set. The scheme is shown pictorially in Figure 5.5.4. In order to attain a result that is representative of the whole data set, the process is repeated numerous times and the mean sensitivity and specificity on the test set is reported. The scheme is ineffective in this case, as the validation and test subsets contain only 14 or 15 events, so an operating point cannot be set with any degree of confidence.

The second method, *5-fold cross-validation*, creates a test set by splitting the A2 events into 5 equal sized groups, and choosing one group as the test subset, and the remainder for the validation subset. This leads to the same problems as the previous scheme, but to a greater degree, as only 5 A2 events are available for the test set (see Figure 5.5.4). Therefore, cross-validation is not feasible, and we conclude that effective retraining will require an independent data set with a greater number of patients.

Although retraining was not possible, improvements to the models may be generated by using a heuristic approach. The approach is presented here for the baseline model in

order to show that the changes to the models can lead to improved results, although no scientific justification for the changes is given here.

In the baseline model, input data were scaled by the training set population mean and variance. In order to make the model more sensitive to hypotensive events, the SDA values were re-scaled if the value was lower than the training set population mean. Rather than the scaling:

$$SDA_n = \frac{SDA - \mu_{SDA}}{\sigma_{SDA}} = \frac{SDA - 94.68}{16.54}$$

a new value of $\sigma_{SDA} = 12$ was chosen. This change has the effect of giving greater importance to low and high blood pressures, so that hypotension is more likely to be detected. Similarly, the SpO_2 values were scaled using the mean and variance from the ED data $\mu_{SpO_2} = 96.98$ and $\sigma_{SpO_2} = 3.13$, in preference to the baseline model values of $\mu_{SpO_2} = 95.20$ and $\sigma_{SpO_2} = 3.49$

5.5.2. Modified Model Results

The number of A2 events, and first A2 events per patient are shown in Figure 5.5.5 for window lengths between 1 and 60 minutes. The plots show that the modified baseline model (changes in $\sigma_{SDA}$from 16.54 to 12, changes in μ_{SpO_2} from 95.2 to 96.98 and changes in σ_{SpO_2} from 3.49 to 3.13) is better at identifying escalation events than the original SVM model. Previously, we highlighted six escalation events that were caused primarily by changes in blood pressure or SpO_2 and were previously missed by SVM. Five of these events were now identified using the modified baseline algorithm. The one event that remained undetected, for ED00320, was due to a borderline hypotensive event that resulted in a PSI score of 2.1, which did not exceed the alerting threshold. However, we note that this patient had sustained hypotension throughout their stay, which would have been identified by the system in the hour preceding the escalation event.

The five extra escalations were detected for three additional patients, so the number of True Positives is 20 and the sensitivity of the model is therefore:

$$sensitivity = \frac{20}{20 + 9} = 69.0\%$$

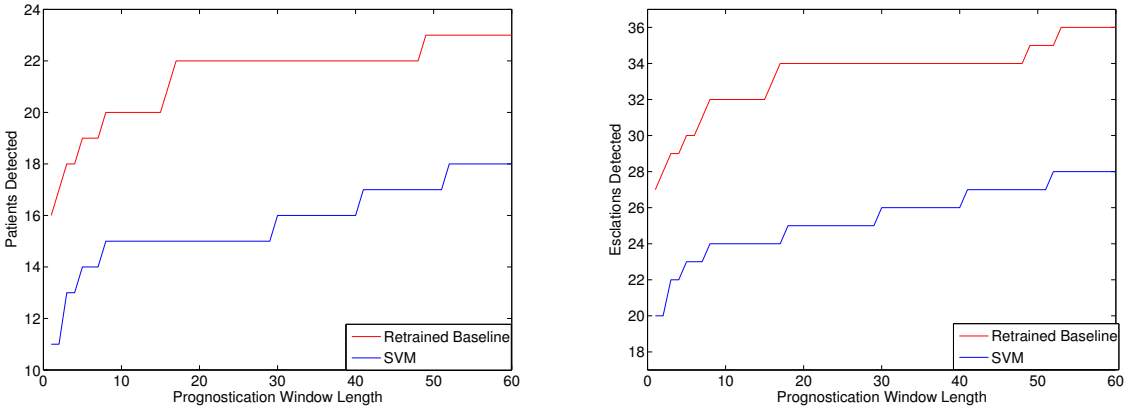

Figure 5.5.5.: Positive events detected by the modified baseline model and SVM model, on a per-event and a per-patient basis

There were a total of 151 true negatives and 66 false positives; 8 extra false positives in comparison to SVMs. This leads to a specificity of:

$$specificity = \frac{151}{151 + 66} = 69.6\%$$

From these results, we postulate that proper retraining of the data fusion models could lead to improvements in sensitivity, with little change in specificity.

5.5.3. Further Limitations

The blood pressure and oxygen saturation adaptations to the baseline model show that improvements in sensitivity and specificity are achievable. A larger data set would allow for model retraining and optimisation in a more structured way. However, even if the model was retrained properly, it can only ever be effective for *identifying* patient vital sign abnormality, not *predict* the onset of abnormality. This is because the baseline model (and each of the other models considered here) only uses the most recent vector of vital signs in its calculations, and discards any temporal information.

For example, consider the vital signs for the patient that was introduced at the start of Chapter 3 in Figure 3.0.1; we now show the result of applying each data fusion model to the vital signs in Figure 5.5.6. Each of the Parzen windows based models is able to detect deterioration in the patient when there are severe oxygen desaturations at 14:20, 16:10 and 17:45, at which stage the current vital sign vector is abnormal with respect to the model.

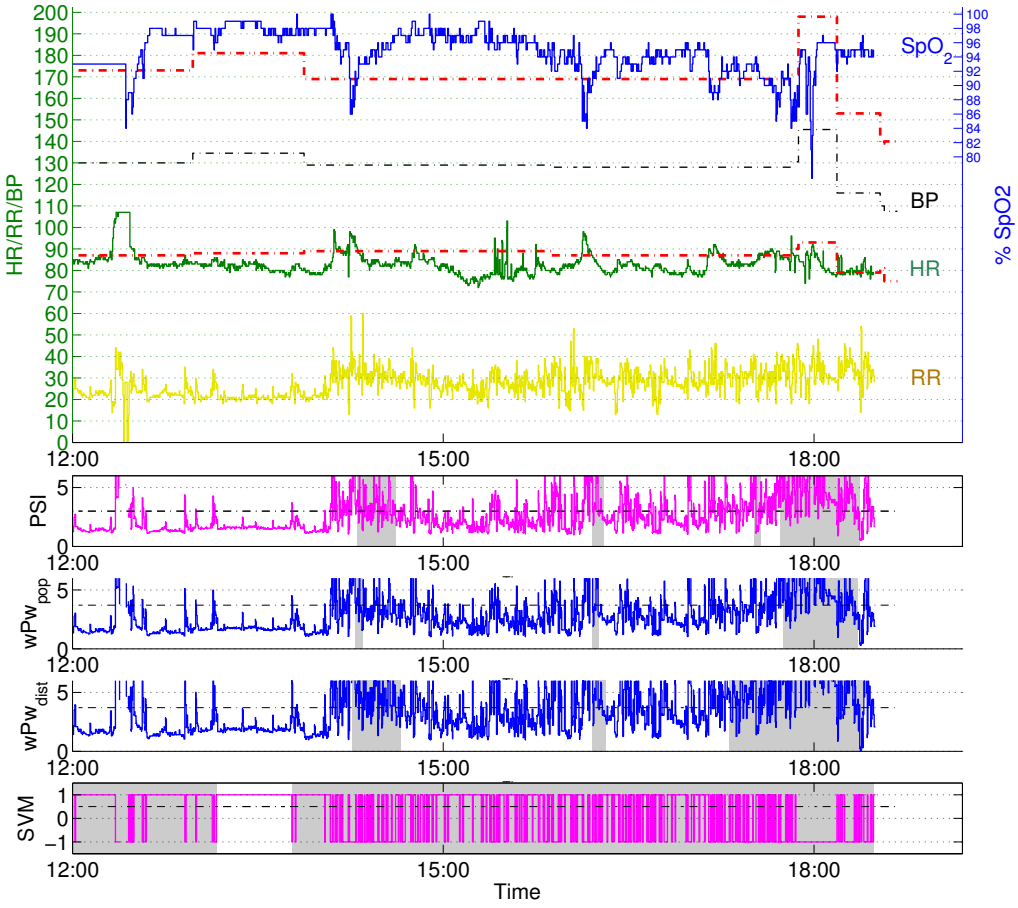

Figure 5.5.6.: Vital signs and data fusion models for the deteriorating patient previously seen in Figure 3.0.1. The data fusion models show that deterioration could have been detected at 14:20, approximately 4 hours before clinical intervention at the end of the record. The gradual increase in score also suggests that trend information may enhance the model.

However, the PSI values for all of the models increase over the course of the patient's stay, indicating a gradual deterioration. In some cases, there will be long-term trends in the vital signs that may be predictive of deterioration which may be detected using more sophisticated methods. In the next chapter, we consider how time-series information may be synthesized into a data fusion model through the use of Gaussian processes.

6. Trend Analysis Using Gaussian Processes

We have so far considered a selection of classification techniques for detecting vital sign abnormality. In the previous chapter, we optimised two of these techniques for the ED data set, which resulted in improvements in the number of physiological escalation events detected. In this chapter, we develop a novel method for vital sign data analysis, which overcomes two of the main limitations of current approaches: data drop-out and the lack of temporal information.

6.1. Remaining Issues With Current Methods of Vital Sign Data Analysis

6.1.1. Data Dropout

In Chapter 3, we noted that for the ED data set the instances for which a single channel of data was temporarily missing affected between 6% and 12% of the total continuous data. The baseline algorithm, as described in Section 4.1, has a short-term and long-term approach to this problem. In the short-term, if single-channel data are not received for a one-minute period, then a local median, calculated from the data for that channel in the previous 5 minutes, is used as a proxy for the missing value. If the data loss is long-term, persisting for more than 30 minutes, the mean of the training data for that channel is used instead.

These approaches were developed heuristically and the parameters have not been justified scientifically. For instance, it is unclear why 30 minutes was chosen as the time at

which to switch to using the population mean, and why the short-time median should use 5 minutes of data. The approach may also introduce artefactual step-changes in the Patient Status Index when switching to the short-term median, and then when switching from this median to the population mean.

An alternative long-term solution which may be used instead of the population mean was proposed by Hann [41]. He suggests that rather than assigning a specific value for the missing variable, a more principled method is to switch to a lower-dimensional model. In doing so, the missing variable is marginalised out, so that the channel contains no information. There are two main problems with this approach. Firstly, to account for drop-out in each of the four vital signs, a separate model needs to be trained for every combination of valid vital sign vectors. If a maximum of two vital sign channels are allowed to drop out, then a total of six ($C_3^4 + C_2^4$) models must be trained and stored.

Secondly, the PSI scores generated by the marginalised models are not necessarily equivalent. That is, if n vital signs produce a PSI of 3 in the 4D model, $(n-1)$ vital signs will not produce a score of 3 in a 3D model. In order to create a standardised score, we may convert the PSI scores generated by each model based on their cumulative distributions, as shown in Figure 6.1.1. In this scheme, any vectors of vital signs that have the same cumulative probability $P(PSI < X)$ are given the same score.

The validity of this method depends on how well the underlying vital sign distribution is estimated. In addition, the method described here does not address the issue of how long before the system switches to a model of lower dimensions, and the switch may still produce step-changes in the PSI that are unrelated to the underlying human physiology.

In summary, the methods currently used to deal with data drop-out are not based on evidence, and are prone to errors caused by switching between short-term and long-term solutions. A more complete solution would not require any such switching, and should provide a probabilistic estimate of the missing vital sign based on all of the previously seen data for that vital sign.

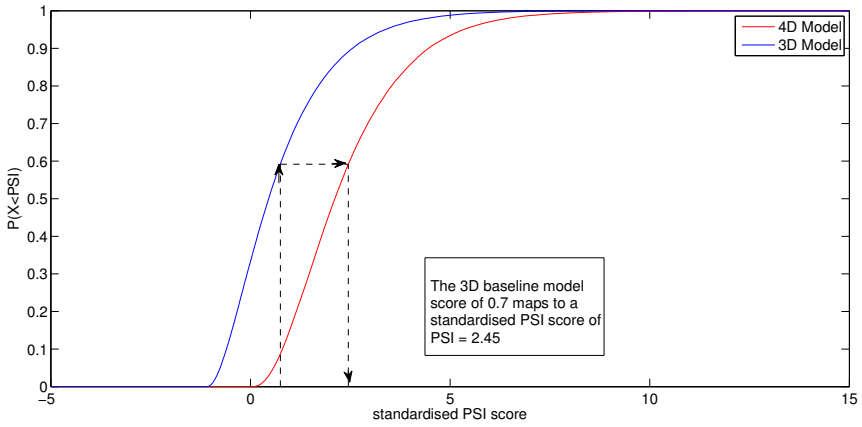

Figure 6.1.1.: Cumulative Distribution Functions for PSI values in the 4D baseline model and a 3D models. The mapping between scores is shown by the dotted line, such that a score of 2.45 in the 4D model is equivalent to a PSI value of 0.7 in the 3D model.

6.1.2. Lack of Temporal Information

One further drawback of the data fusion methods presented so far is their inability to assess the vital sign trends for a patient. Only the most recent vital sign vector is used to assess the patient's physiological status. This differs from standard clinical practice, in which rudimentary trend analysis is considered to be an important part of detecting deterioration, and Track-and-Trigger observation charts are designed so that the overall trajectory of a patient can be ascertained as easily as possible.

Trend analysis in automated systems has previously been attempted, most notably by Charbonnier and Gentil [20]. In their system, semi-quantitative trend features are calculated such that the output score is based not only on the latest values for the vital signs, but also on trend features. However, one problem with their approach is that the choice of trend features is arbitrary: an infinite number of trend features could be selected, and the research may be improved by principled feature selection.

6.1.3. Time Series Analysis

To deal with the drawbacks described in the previous section, we hypothesise that it may be possible to use information from a vital sign time series to infer missing or future vital sign values. The problem of data drop out could be potentially solved by predicting the missing values, while the prediction of future vital sign values may be used for trend

analysis.

For time-series techniques to be successful, we must first show that our vital sign data are not comprised of independent and identically distributed (i.i.d.) samples, but instead that there is some degree of correlation between vital sign values that are close together in time. We can check whether the vital signs are i.i.d. by analysing each channel for "records" [30]. A new record is set if the current value exceeds all previous values of that vital sign. The status of a vital sign can be denoted by the vector X_j, where $X_j = 1$ if the j^{th} value is a record, and 0 otherwise. If a data set is i.i.d., the probability that the j^{th} value is a record is simply $\frac{1}{j}$. For any number of data samples the expected number of records can be derived from simple probability theory and is given approximately by:

$$E\left(X_j\right) \approx ln(j) + \gamma \tag{6.1.1}$$

where γ is the Euler-Mascheroni constant:

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - ln(n)\right) = \int_1^{\infty}\left(\frac{1}{|x|} - \frac{1}{x}\right)dx \approx 0.577 \tag{6.1.2}$$

This number of vital sign records was calculated for the entirety of JR training data set from Chapter 4, by concatenating all of the vital sign vectors in chronological order, to form one long data stream for each vital sign. This is less accurate than calculating the number of records on a per-patient basis, as vital signs between patients are likely to be independent. However, this method has the advantage of allowing results to be presented in a single graph.

The results are shown in the upper graph in Figure 6.1.2. In each case, the shape of the distribution is significantly different to that given by $E(X_j)$, indicating that the vital signs are not i.i.d. We confirm that $E(X_j)$ is a good estimate by comparing it to a pseudo-independent vital sign data set that was generated by randomising the order of the vital signs. The lower graph in Figure 6.1.2 shows there is a good match between the distribution of $E(X_j)$ and the number of records for the independent set.

The sudden step changes for each of the vital signs were due to instances where a patient had abnormal, record-breaking, vital sign values to begin with, but then continued to deteriorate such that every subsequent recording was also a record. While HR, SpO_2 and

SBP contained hundreds of records, there were only 23 respiratory rate records in total, in comparison to 13 predicted by $E(X_j)$. The low total is due to the vital signs for the initial patients containing a very extreme value of respiratory rate (43 respirations/min), which is only exceeded 3 times thereafter. Therefore, the total number of records depends on the order of patients, and instead it is the different shapes of the distributions that indicate whether a data series is i.i.d.

6.2. I.I.D. Patient-Specific Model

In the next section, we will introduce a time-series model that will be trained on a patient-specific basis. In order to test its effectiveness, we will firstly compare it to the short-term median filter used by the baseline model as described in Chapter 4. In addition to this, we will compare the time series model to an i.i.d. model that we now describe. The purpose of this comparison is to determine whether time-series techniques have benefits over i.i.d. techniques.

A simple i.i.d. model can be generated by applying Parzen windows to the vital sign data from each patient. Parzen windows can be considered as i.i.d. as the time information for the training points is not included in the model, and so each training point has an equal influence on the prediction of future points. Figure 6.2.1 gives a pictorial representation of how a kernel is applied to each data point. These kernels are then summed to generate a posterior probability distribution depicted in red, which may be used to predict the next vital sign value. Note that the model described here is generated on a patient-specific basis; this will allow a fair comparison with the time-series model, which will also be patient-specific. The kernel width parameter can again be estimated using the standard metric (Equation 4.1.5), or else by maximising the likelihood of the data. The output of the model is always zeroth order, so that the posterior distribution is the same for all time. This is because the i.i.d. assumption implies that there is no difference between vital sign behaviour in the near future and distant future.

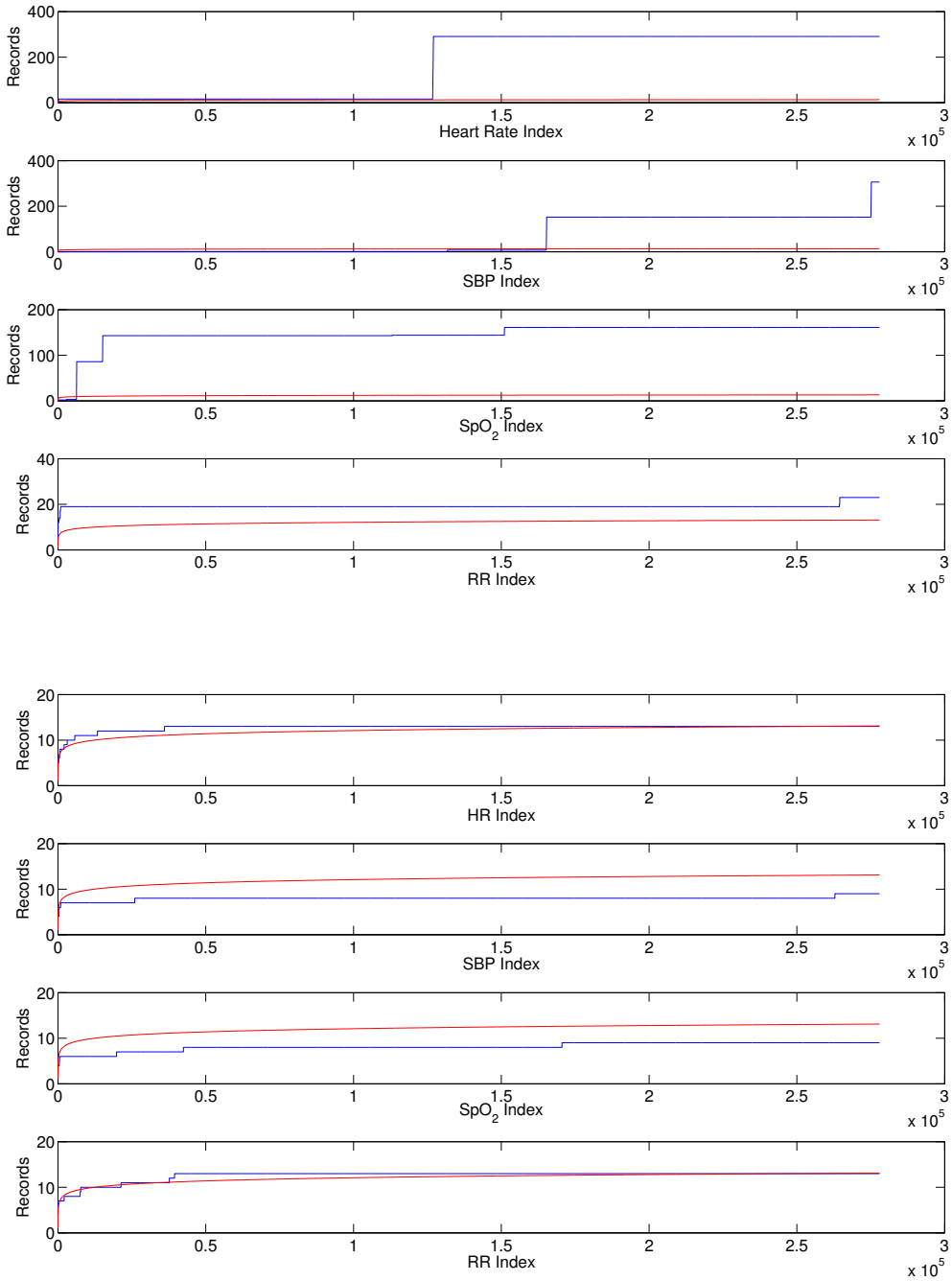

Figure 6.1.2.: Number of records for each of the vital signs in chronological order (upper figure), and in a random order (lower figure). The expected number of records for an i.i.d. data set is shown as a red line, while the actual number of records is shown in blue.

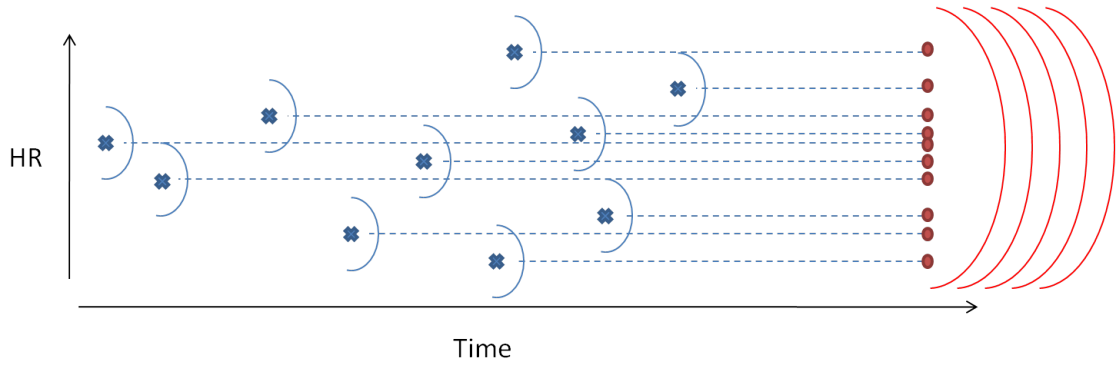

Figure 6.2.1.: Demonstration of a univariate patient-specific i.i.d. model for Heart Rate. Each point of previously seen data (shown in blue) contributes to a Parzen windows model (in red). Because the model is i.i.d., the Parzen windows estimate remains the same in time until a new data point is seen.

6.3. Gaussian Processes

In the remainder of this chapter, we consider how the time series analysis technique known as Gaussian processes may be used to deal with the problems highlighted in Sections 6.1.1 and 6.1.2 and then show how improvements to a data fusion model can be achieved within an integrated framework which incorporates the Patient Status Index.

Time series analysis is a wide-ranging field, with practical applications in economics [112] and meteorology [91], among others. A number of other time series methodologies are not considered here, but may result in solutions that are as effective as the Gaussian process in certain circumstances. In particular, Kalman filters, which recursively estimate the mean and variance of a variable with a Gaussian distribution can be shown to be equivalent to a special case of Gaussian processes [67], in which the linear state state space model used by the Kalman filter maps to a Gaussian process model with a particular covariance function. Gaussian process regression is a more general formulation than a Kalman filter, in that the training process involves choosing from a family of models rather than assuming a model to begin with.

For more complex multimodal behaviour, the particle filter approach may be more applicable [3]. With this methodology, random 'particles' with associated weights are used to represent the posterior density, which provides the ability to represent arbitrary densities. However, the method has high computational complexity, and the number of particles required increases exponentially with the number of dimensions in the problem

domain.

6.3.1. Gaussian Process Overview

A Gaussian process is defined as a stochastic process for which any finite combination of samples have a joint Gaussian distribution. Using this property, it is possible to regress onto previously unseen data points. We considered that the vital sign data met the criteria for a Gaussian process, firstly noting that under normal physiological conditions, on the order of minutes, one would expect vital sign values to be similar to previous vital sign observations. Furthermore, the probability that a new vital sign value deviates from previous values should decrease with the magnitude of the deviation . Therefore, we postulate that at short timescales, the posterior state of any of the vital sign parameters can be well-modelled as a Gaussian.

This is particularly useful, as the distribution can be fully parameterised by a mean and variance. In more complex applications, the posterior state may be multimodal, such that the posterior mean is unsuitable as it may fall between two distinct regimes. Secondly, Gaussian processes are able to deal naturally with asynchronous data, which is useful in the ED setting where we may expect temporary data loss as a result of movement artefacts.

The Gaussian process methodology for regression was used for time-series analysis as long ago as 1880 [62]. However, the first modern applications of Gaussian processes were developed in the 1970s in the field of geostatics where the method is known as Kriging [80]. Kriging was developed from the ideas of Krige [1], and was developed into a mathematical framework by Matheron in 1963 [70]. Typically, the method is used within geostatics to map physical surfaces from limited sample data.

The use of Gaussian processes in machine learning is much more recent and was derived independently from previous theory. Subsequently, it has been shown that the Gaussian process model has a close relationship to a Bayesian neural network [77, 119]; in particular, Neal showed that a neural network with one hidden layer converges to a Gaussian process as the number of hidden nodes goes to infiinity. However, for a finite Bayesian network, Rasmussen demonstrated (using two independent examples) that Gaussian process

methods performed best for small (up to 1000 point) training sets [92].

Gaussian process regression has since been used in a number of diverse applications including generating music playlists [87], visualisation [63], and has been considered for use in predicting patient condition in intensive care [40]. Gaussian processes can also be adapted for classification [117], though we only consider their use for regression in this chapter.

We can gain an intuitive understanding of the Gaussian process model by first considering the 2-D Gaussian in Figure 6.3.1, which has a mean of $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance of $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$.

Any point on the distribution (y_1, y_2) can be said to describe the joint probability of two samples. For instance, the red dot on the left-hand side plot represents the probability of the two points $y_1 = 1.5$, $y_2 = -0.2$.

Alternatively, the points y_1 and y_2 can also be depicted as a two-point time series, as shown in the right subplot. Note that the time at which y_1 and y_2 occur has not yet been specified; this will be considered later, but for now, let us suppose that these points occur at arbitrary times x_1 and x_2. As the time series grows with more points $y_3, y_4, ..., y_n$ it can continue to be modelled as a Gaussian process provided that the dimensionality of the Gaussian is increased to $3, 4, ..., n$. Thus an n-point stochastic time series can be fully described by an n-dimensional Gaussian.

The primary advantage of modelling a time series with this approach is that it allows us to infer a posterior probability distribution over the missing values of y, given the known data. The mean of the posterior then represents our best estimate of the missing data, and the variance provides a measure of confidence in the inference.

For instance, in the example shown previously, the outermost contours indicate that y_2 is likely to take values between -2 and 2. However, now consider the case for when the value of y_1 is known to be 1.5. The selection of y_1 constrains y_2 to a 1-D slice, denoted by the dotted red line in Figure 6.3.1. This is the conditional distribution $p(y_2|y_1 = 1.5)$. From the contours, we can see that y_2 is now likely to lie between -0.3 and 2. By using prior knowledge about y_1, we have reduced the uncertainty about the value of the unknown

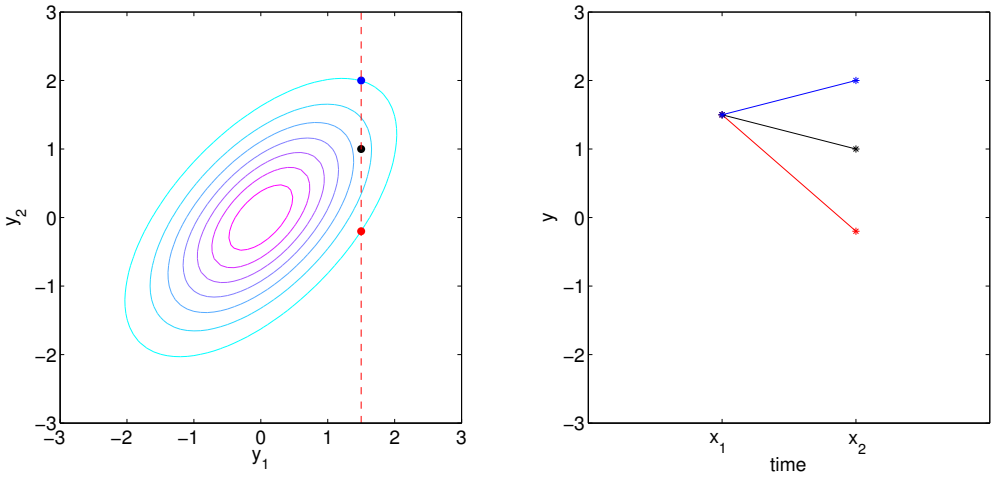

Figure 6.3.1.: Simple example of a Gaussian process for two points. The left figure shows the joint probabilities of all possible points as a bivariate Gaussian distribution. The right figure shows the time series plots for the three points on the Gaussian distribution highlighted in blue, black and red.

point, y_2.

6.3.2. Covariance Functions

The previous section showed how we could develop a time series, but without showing how the values on the time axis were set. Similarly, we introduced a covariance matrix governing the correlation between data points, but did not explain how the covariance matrix was derived. We now show that the two are linked through the use of a covariance function.

A covariance function is a function of the times (x_i, x_j) of any two data points (y_i, y_j). The form of the function is free, but the function must lead to the production of positive semidefinite covariance matrices. In most practical applications, we expect local behaviour to be highly correlated and the correlation to decrease as data samples become further apart in time. Furthermore, in many cases, we may also expect the function to be stationary, so that only the distance between x_i and x_j is important, and not their absolute values.

For this reason, one of the most common classes of covariance functions is the squared exponential, which takes the form:

$$Cov(y_i, y_j) = K(x_i, x_j) = \sigma_0^2 exp(-\frac{1}{2}\frac{|x_i - x_j|^2}{\lambda}) \tag{6.3.1}$$

Note that x_i and x_j may take any value, and therefore the generated time series does not require evenly spaced data. The covariance function in this case contains two free hyperparameters. The amplitude hyperparameter, σ_0, defines the maximum allowable variance, and is high for variables with a high dynamic range. The length-scale hyperparameter, λ, controls how long an observation will be correlated to future observations and thus has the effect of shaping the smoothness of the output.

The hyperparameters may be set using prior knowledge. For example, if we were modelling respiratory rate, we may want to set $\sigma_0 \approx 3$ to reflect the fact that the range of respiratory rates is approximately 20 ± 10 rpm (as $10 \approx 3$s.d. from the mean).

In the absence of strong prior knowledge, the hyperparameters can be learned from the data by maximising the likelihood for the hyperparameters. Given Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

and assuming that the prior $P(\mathbf{\Phi})$ is uninformative, the posterior over the hyperparameters can be written as:

$$P(\mathbf{\Phi}|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\mathbf{X}, \mathbf{\Phi})$$

The likelihood, $p(B|A)$, in this situation is simply:

$$P(\mathbf{y}|\mathbf{X}, \mathbf{\Phi}) = N(\mu, \mathbf{K}) \tag{6.3.2}$$

where K is the covariance matrix, $\mathbf{y}$ represents the N time series observations, $\mathbf{X}$ is the corresponding vector of input data (the time points), and $\mathbf{\Phi}$ are the set of covariance function hyperparameters. The log likelihood is:

$$L = \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log 2\pi \tag{6.3.3}$$

The derivative of the log likelihood with respect to Φ is then:

$$\frac{\partial L}{\partial \Phi_i} = -\frac{1}{2} trace(\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Phi_i}) + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Phi_i}\mathbf{K}^{-1}\mathbf{y} \tag{6.3.4}$$

By setting this to zero, the most likely value of Φ can be calculated. Alternatively, the posterior distribution of $\mathbf{\Phi}$ can be calculated using Monte Carlo sampling methods.

Where there is further structure in the data, other classes of covariance function may be more suitable. For instance, a periodic covariance function (Equation 6.3.5) can be used to enforce sinusoidal behavior.

$$k(x, x') = exp\left(-\frac{2sin^2\left(\frac{x-x'}{2}\right)}{\lambda^2}\right) \tag{6.3.5}$$

In addition, new covariance functions can be constructed by combining standard classes of covariance function by summing, convolution, or using the product of two known covariance functions. This allows for multi-scale behaviour to be modelled. An example of this is given by Stegle et al. [102], who attempted to infer missing heart rate data using Gaussian processes. They noticed that there were two types of behaviour - a short-term process which appeared smooth on a timescale of a few minutes, and a long-term periodicity due to circadian rhythm. In response, a covariance function consisting of the sum of a Matern covariance function and periodic function was used successfully. The Matern function takes the form:

$$k(x, x') = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(2\sqrt{\nu}\frac{(x-x')}{\rho}\right)^{\nu} K_\nu\left(2\sqrt{\nu}\frac{(x-x')}{\rho}\right)$$

where ρ, σ, and ν are the covariance function hyperparameters, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind.

6.3.3. Gaussian Process Regression

The possibility of inference using a Gaussian process model was highlighted in Section 6.3.1. In this section, we formally derive the mathematics for Gaussian processes in the general case, and then show how the equations may be applied to the simple 2D example described in Section 6.3.1.

Consider the case for which the values of a number of points, $\mathbf{y}$, are already known, and

we wish to estimate the mean and covariance of an additional unknown points, $\mathbf{y}*$, based on the known data. The vector of all the points, [$\mathbf{y}$, $\mathbf{y}*$], has a vector with corresponding times points [$\mathbf{x}$, $\mathbf{x}*$]. By using a covariance function, we can build a covariance matrix, $\mathbf{K}$, for all the points, so that $K_{i,j} = k(x_i, x_j)$. The elements of the covariance matrix describe all the correlations between each pair of points, and as a whole describe the joint distribution of all the points in the time series, $P(\mathbf{y}, \mathbf{y}*)$.

The covariance matrix can be divided up into the components that describe the correlations between the known points y, the correlations between the unknown points, $y*$, and the cross-terms. Let us label these components, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively, so that $\mathbf{K}$ can be expressed as:

$$\mathbf{K} = \begin{bmatrix} \mathbf{A} & \mathbf{C}^T \\ \mathbf{C} & \mathbf{B} \end{bmatrix} \tag{6.3.6}$$

The conditional distribution $P(\mathbf{y}_*|\mathbf{y})$, which describes the probability distribution of the new points, y_* when each of the other y points is fixed, may be derived using the joint distribution. We may consider the conditional distribution to be a slice through the joint distribution, and hence the conditional distribution is also Gaussian distributed, with mean and variance:

$$p(\mathbf{y}_*|\mathbf{y}) \sim N(\mathbf{C}^T\mathbf{B}^{-1}\mathbf{y},\ \mathbf{A} - \mathbf{C}^T\mathbf{B}^{-1}\mathbf{C}) \tag{6.3.7}$$

The full derivation of the conditional distribution can be found in Appendix C. In order to compute the conditional distribution, the elements of the covariance matrix must be inverted, a process which is of order $O(N^3)$ complexity. This means that there is a practical limit to the number of points in the time series of a few thousand points on a regular desktop computer. The upper limit can be further extended through the use of sparse matrix techniques. However, these limitations will not affect the use of Gaussian processes in our application, where we may expect a maximum data series of approximately 500 points during a patient's four-hour stay in the ED (assuming that the maximum sampling rate is approximately equal to 30 seconds).

We can confirm the result of Equation 6.3.7 using the example in Figure 6.3.1. In this

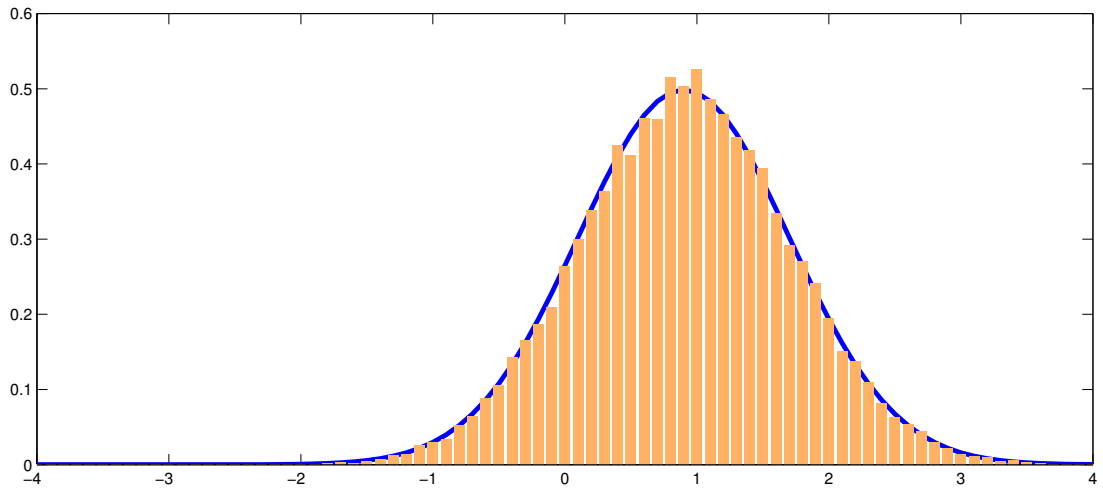

Figure 6.3.2.: Comparison of the calculated conditional distribution $P(x_2|x_1)$ and samples selected from the joint distribution $P(x_1, x_2)$ when samples are only accepted if $1.49 \leq x_1 \leq 1.51$.

example, the covariance matrix $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$ was used to generate the 2D Gaussian. Let us now assume that the value of y_1 has previously been observed as $y_1 = 1.5$, so that we want to infer the distribution of the slice along the dotted red line. Using Equation 6.3.7, the conditional distribution of the point x_2 is:

$$p(y_2|y_1 = 1.5) \sim N(0.6 \times 1 \times 1.5,\ 1 - 0.6 \times 1 \times 0.6) = N(0.9, 0.64) \tag{6.3.8}$$

We can test this result by randomly sampling from the joint distribution, and only accepting the values of y_2 when $1.49 \leq y_1 \leq 1.51$. These samples can then be used to generate a discrete probability distribution that can be compared to our calculated conditional distribution. The result is shown in Figure 6.3.2, which shows a very close match between the experimental and theoretical results.

6.3.4. Noise Processes

Up to this stage, the Gaussian process model that we have derived makes the assumption that the input data are noiseless. However, in most practical applications, we do not have access to the true underlying state, but only a noisy measurement of the state. This is certainly true in our application, for which we may expect noise to be generated by

external factors such as patient movement, as well as by noise processes in the measuring device itself.

We assume that the noise is additive, and that it has behaviour characterised by a Gaussian distribution with zero mean and variance σ_n^2, such that our observations can be modelled by:

$$y = f(x) + \epsilon \tag{6.3.9}$$

where ϵ is a noise process. This additional term can be incorporated into the Gaussian process model by adapting the covariance function so that:

$$Cov(y_i, y_j) = K(x_i, x_j) + \sigma_n^2 \delta_{ij} \tag{6.3.10}$$

where $\delta_{ij} = 1$ is the Kronecker delta function which takes a value of 1.0 if $p = q$, and is zero otherwise. We can then update the predictive equations in Equation 6.3.7, so that the mean and variance of the conditional distribution are:

$$\begin{aligned} mean &= \mathbf{C}^T(\mathbf{B} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \\ variance &= \mathbf{A} - \mathbf{C}^T(\mathbf{B} + \sigma_n^2\mathbf{I})^{-1}\mathbf{C}) \end{aligned} \tag{6.3.11}$$

where $\mathbf{I}$, the identity matrix, is necessarily the same size as $\mathbf{B}$. The parameter, σ_n^2, that controls the level of noise can be set explicitly using prior knowledge. Otherwise, we may treat it as another hyperparameter and use a maximum likelihood estimate. The effect on the Gaussian process regression output is that the mean of the Gaussian process now no longer has to pass through each of the observed data points. This allows a more general regression, and so the introduction of the noise process can be likened to that of the slack variable in the SVM method.

6.4. Univariate Gaussian Processes

Having established the Gaussian processes methodology, we now apply it to univariate data. Initially, we demonstrate the method on synthetic data, and then test it more fully on one of the vital signs, the heart rate, as a proof of concept. Heart rate was chosen

because there was little data loss for this parameter in the ED data set. However, the procedure outlined here is equally valid for any of the vital signs.

We use a squared exponential covariance function because of its simplicity and its ability to model a wide range of behaviours. Furthermore, the function matches our prior belief that data points close to each other in time are more strongly correlated than points further apart in time. The hyperparameters are selected using the *maximum a posteriori* estimates of the marginal log-likelihoods, as described is Section 6.3.2.

In general, the choice of covariance function is of critical importance, and in a more detailed study it would be appropriate to select the covariance function in a more principled manner. This may involve, for instance, searching through the data set and investigating temporal features such as the average rate of change for each variable, which may indicate a sensible range for the length-scale parameter, and whether there were any regular patterns such as circadian rhythms.

6.4.1. Synthetic Data Example

The Gaussian process model was tested on a synthetic data example using training data that consisted of 90 points generated from a sine wave signal given by $y = sin(x)$ with additive Gaussian white noise $Noise \sim N(0, 0.125)$. The points, y_{train}, were evaluated at evenly spaced intervals between $1 \leq x_{train} \leq 9$. Gaussian process regression was then used to estimate the values of the next ten data points, between $9 \leq x_{test} \leq 10$, using the *gpcovar.m* and *gpfwd.m* functions from the Netlab toolkit, which estimates the hyperparameters by maximising the likelihood over the hyperparameters. The results are shown in Figure 6.4.1, where the training data are shown as black markers, and the underlying sine wave is shown as a dashed black line. The mean of the regression, and ± 2 standard deviations are shown in red on the figure.

The 90 training points, x_{train}, were used to generate the matrix A described in Equation 6.3.6, and the Gaussian process regression estimates were evaluated simultaneously at all 10 values of x_{test}, such that the matrices B and C in Equation 6.3.6 are of dimensionality 10×10 and 1×10 respectively.

The regression estimate is calculated from Equation 6.3.7, and the solution is a 10-

dimensional Gaussian distribution. The mean and variance at a single point of interest can be derived by marginalising out all of the other dimensions. We note that this is exactly equivalent to evaluating each of the test points sequentially. Unlike other time-series analysis methods such as Autoregressive Moving Average (ARMA) models, the data points initially estimated are not used in estimation of any future next data points.

The upper plot in the figure shows that the Gaussian process models the signal well, continuing the short-term downward trend. In addition, the range between the ± 2 s.d. markers increases between $x = 9$ and $x = 10$, correctly indicating that the confidence in the prediction decreases as we get further away from the observations.

In comparison, Figure 6.4.1 also shows the results that would have been generated if an i.i.d. model had been used (lower plot), and if the median filter method employed by the baseline model was used (middle plot). The i.i.d. model was trained on the 90 data points used previously, using the Parzen windows algorithm as described in Section 6.2. The Parzen windows width was set in the standard way as described by Equation 4.1.5. The model was used to predict the 10 test points, and again, the data points initially estimated were not used to update the model. The estimate of the median and the 5^{th} and 95^{th} percentiles are shown as green dashed lines in the figure, and the i.i.d. model distribution is shown on a separate axis. Under the i.i.d. assumption, every training data point, regardless of distance from the test points, has an equal influence on the posterior distribution. As a consequence, the i.i.d. model is unable to capture short-term changes, and each data point estimate has a wide variance that reflects the sample variance of y_{train}.

In the baseline model, the unknown data are modelled by the short-term median of the most recent training data. For the vital sign data, "short-term" is defined as all data within the most recent 5-minute period. In this example, we calculate the median of the ten most-recent data points, which reflects the typical number of data points we would expect within a 5-minute period, when the sampling rate is 30 seconds.

The result of this is shown as a solid blue line in Figure 6.4.1, from which it is clear that the median does not model the characteristics of the time series well. In fact, one would expect the median to perform poorly in any case where the signal has large changes

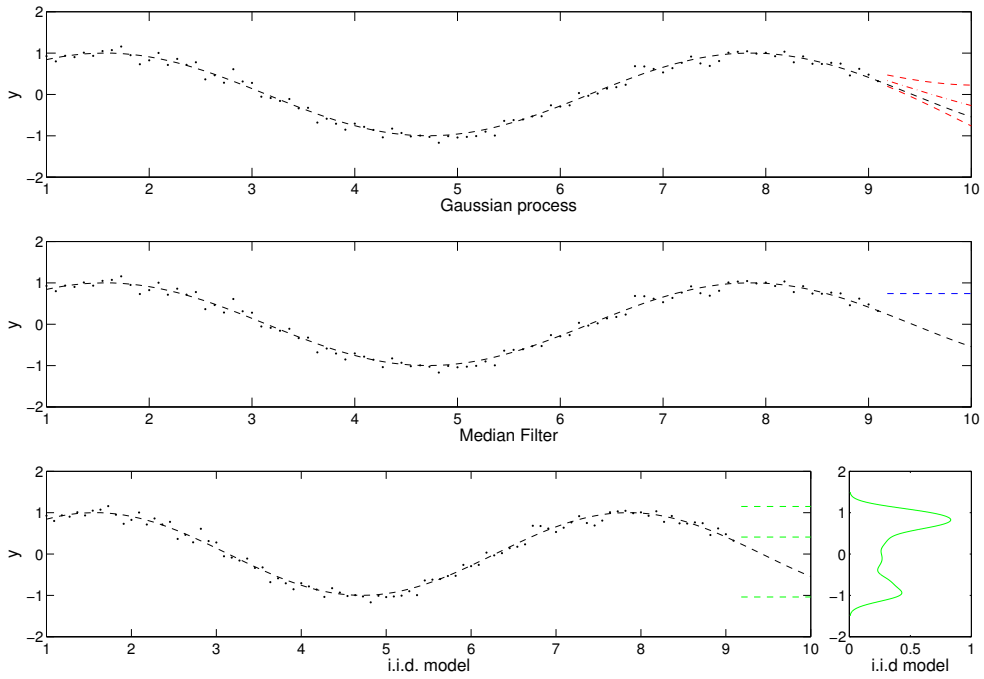

Figure 6.4.1.: Gaussian process regression estimate on noisy sine wave data. The underlying sine wave is shown in black, along with the data points with added Gaussian noise. The Gaussian process regression estimate is shown in the uppermost graph, where the central line shows the mean of the posterior distribution, and the two outer lines show the values of the mean ± two standard deviations. For comparison, the estimate of the missing data using a 10-point median filter that has a similar behaviour to the 5-minute median filter described for the baseline algorithm is shown in the middle plot, and the i.i.d. model estimate is shown in green in the lower plot.

within the median filter window. Unlike the Gaussian process, or the i.i.d model, the median filter does not provide a posterior distribution, and cannot therefore estimate the uncertainty in a prediction.

6.4.2. Patient Data Examples

Gaussian processes were then tested on two examples of real vital sign data. The Gaussian process regressions for heart rate data samples are shown in Figure 6.4.2. In the example in Figure 6.4.2(a), 15 minutes of data were selected at random from an arbitrary patient in the ED data set. This data were then split into an initial 10 minutes of training data, which are depicted as black circles, and 5 minutes of test data, which are shown in red. The Gaussian process model was implemented using the training data, and the mean of

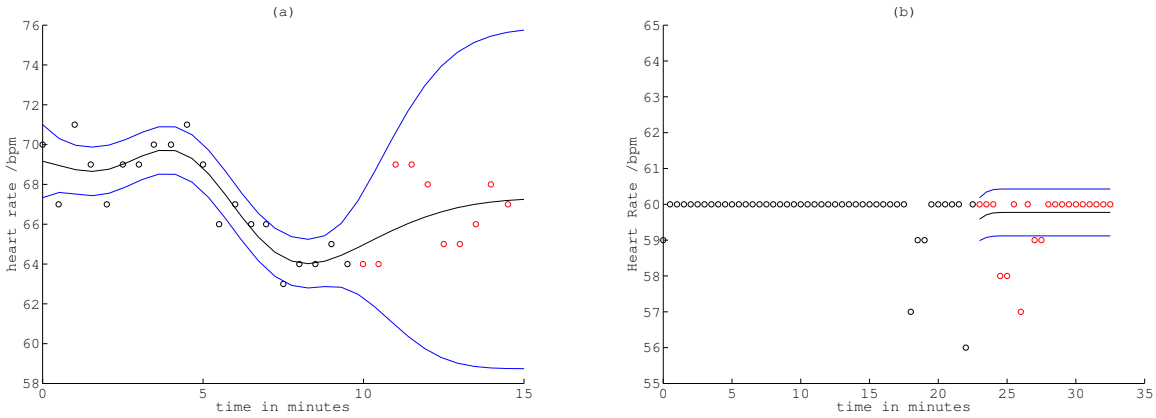

Figure 6.4.2.: Gaussian process regression examples for heart rate on (a) an arbitrarily chosen patient (b) a patient with a pacemaker. The black dots were used to train the model. The mean of the resulting model is shown as a black line, and ±2s.d.s are shown as blue lines. The red dots were unseen test data, and allow us to assess of the performance of the model.

the estimate is shown as a black line alongside the ±2s.d. confidence intervals, which are shown in blue. A visual inspection confirms that the regression appears to predict an overall upward trend in the heart rate values. In addition, we once again see that the estimate becomes less certain as time progresses and the points that we try to estimate become further away from the last observed data.

Figure 6.4.2(b) shows Patient ED00236, who had an activated pacemaker during his stay at the ED. The graph shows that the patient's heart rate remained at 60 bpm for most of the 23-minute segment of data that was used as the input to the model. The short-term deviations from the fixed heart rate are likely to have been caused by errors from the heart-rate monitor, rather than the pacemaker. The test data, again marked in red, shows that the heart rate remained constant for the following ten minutes, apart from two periods of short-term deviation.

The Gaussian process model correctly estimates that the heart rate remains constant throughout the ten-minute period, and also predicts that the heart rate is 59.8 ± 0.5bpm with 95% confidence (2sds). Compared to Figure 6.4.2(a), the 95% confidence interval is much smaller due to the low variance in the training data. We note that for this atypical example, both the i.i.d. model and the median filter would also provide the correct heart rate estimate.

6.5. Testing the Model

Up to this point, we have shown how Gaussian processes appear to provide a sensible estimate of missing data for two simple examples. The effectiveness of the model can be assessed more thoroughly using a larger set of samples from the ED data set.

We selected 1000 samples at random from the ED data set, each consisting of 30-minute segments of heart rate data. The randomisation process was as follows: firstly, a patient was selected using a random number generator. A data segment was extracted from the data for this patient by setting the start of the segment to be a randomly-selected time between the start and end of the patient's stay at the ED. The data segment was selected to be the 30 minutes of data immediately following the start time.

Each segment was checked for data completeness and was accepted if it contained at least 40 data points, representing a 2/3 completion rate, assuming that the data were sampled every 30 seconds. If the segment did not have enough data points, it was discarded, and another segment was selected at random. This process was repeated until 1000 segments had been selected. Each of the 30-minute segments were then divided into an initial 20-minute period of training data, and a 10-minute period of test data. The Gaussian processes were created using only the training data, and then tested by seeing how well the model predicted the values in the test data set. Similarly, the i.i.d model was generated from the 1000 segments and the baseline model median-filter was also applied to the final five minutes of each segment of training data.

The choice of 30 minutes for the length of the data segments provided a long enough training period so that medium-term trends in the data might be identified, but was also short enough for the models to be computed quickly. A more detailed study could examine the performance of Gaussian processes using different segment lengths for the training data, as the regression estimate should improve given more training data.

6.5.1. Quantifying the Gaussian Process Error

We can quantify the accuracy of the Gaussian process prediction by calculating the root-mean-squared error (RMSE) for each of the data points in the 10 minutes of test data. Typically, this will include up to 20 data points. If the test data, $[y_1 y_2 ... y_n]$, and the

Model	Gaussian Process	i.i.d.	Baseline Median
Average Time (/s)	4×10^{-1}	3.5×10^{-3}	4.7×10^{-5}

Table 6.1.: The average time taken, in seconds, to compute the Gaussian process, i.i.d., and median filter models using 1000 data segments.

Gaussian process estimate, $[y_{*1}y_{*2}...y_{*n}]$, evaluated at points $[x_1, x_2...x_n]$ are represented by:

$$Y_{test} = [y_1, y_2...y_n] \quad and \quad Y_* = [y_{*1}, y_{*2}...y_{*n}] \tag{6.5.1}$$

respectively, then the RMSE is given by the formula:

$$RMSE(Y_{test}, Y_*) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{*i})^2}{n}} \tag{6.5.2}$$

We can then examine the distribution of the errors, and calculate the mean error over all the 1000 segments to provide a metric which quantifies the performance of the Gaussian process model. Similarly, we can use the same method to evaluate how well the i.i.d model, and the median-filter baseline model perform.

6.5.2. Results

The Gaussian process, i.i.d., and median filter models were applied to the 1000 segments of training data. The average time to compute each model is shown in Table 6.1. Of the three models, Gaussian processes were the slowest, taking 100 times longer to train than the i.i.d. model. The training time may be improved to some extent by using a more efficient implementation, though it is still likely to be considerably longer than for the other two techniques. Most importantly, although the Gaussian process was slowest, with an average training time of 0.4 seconds, each of the models could be trained far more quickly than the data acquisition rate of one sample every 30 seconds, and such models could therefore be used in real-time applications.

The mean RMSE errors for the three models are summarised in Table 6.2. The average absolute errors at 30 seconds, 1, 2, 3, and 5 minutes from the start of the test data are also shown in the table. The mean RMSE for Gaussian processes is approximately the

Model	Gaussian Process	i.i.d	Baseline Median
Average Error (all times)	4.38	4.57	4.37
30 secs	2.51	3.48	2.85
1 min	3.14	3.78	3.19
2 min	3.60	3.81	3.52
3 min	3.87	3.93	3.66
5 min	4.01	4.28	4.10

Table 6.2.: The average RMSE error for the Gaussian process, i.i.d., and median filter models. The RMSE errors at various times from the start of each test data segment are also presented.

same as the RMSE for the 5-minute median method, while the mean i.i.d. error is worse.

The average absolute error for the first 30 seconds of test data is smallest for the Gaussian Process model, performing better than both the median and i.i.d. models. The i.i.d. model performs particularly poorly, with an average error 50% greater than for the Gaussian Process model. Each of the models shows an increase in mean absolute error over the duration of the test data segment. The increase is most significant for the median and Gaussian Process models, which had errors ranging from 2.85 to 4.10, and 2.51 to 4.06 respectively, in comparison to the i.i.d. model, which ranged from 3.48 to 4.28. The distributions of the RMS errors are shown in Figure 6.5.1. The figure shows that each of the models appear to have similar distributions but that the modal error for the i.i.d. and Gaussian Process methods is less than the modal error for the 5-minute median. Furthermore, the i.i.d. method has a greater number of instances for which the RMS error is greater than 20.

The results indicate that the the Gaussian Process regression performs at roughly the same level as the 5-minute median. We postulate that this is due to the fact that in many instances, the vital sign does not change significantly over the test (10-minute) period, so there are no significant trends to detect. To test this, we repeated the experiment on another randomly-selected set of 1000 samples. This time, we included an additional condition, that the training data had a variance of 10 or greater. An additional method for estimating missing data, sample-and-hold (in which the last observation is held through time), was also included.

The RMS errors for each of the methods are shown in Table 6.3. As before, each of the models shows an increase in error over the duration of the test. However, unlike

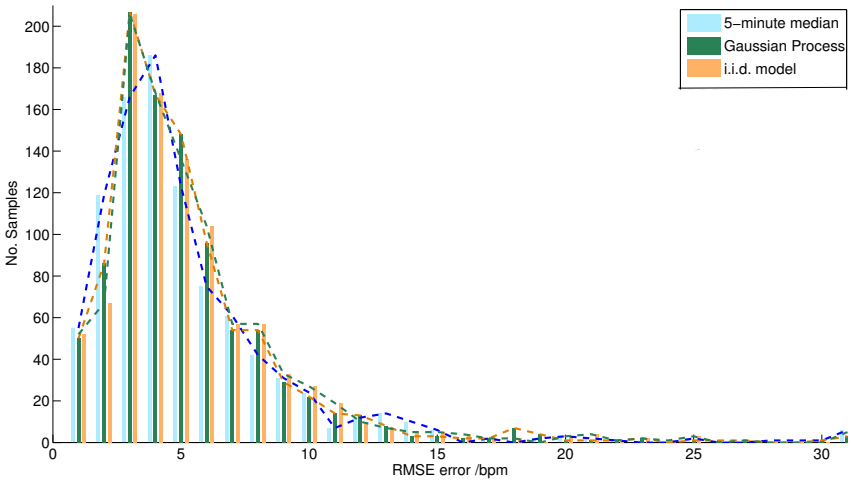

Figure 6.5.1.: Distribution of the RMS error between the 1000 test data segments estimated for the heart rate time series, and the corresponding estimates from the Gaussian process, i.i.d., and median filter models.

Model	Gaussian Process	i.i.d	Baseline Median	Sample-and-Hold
Average Error (all times)	5.62	6.30	5.85	6.22
30 secs	3.10	4.78	3.62	3.33
1 min	3.76	4.98	3.95	4.00
2 min	4.37	5.23	4.53	4.73
3 min	4.75	5.40	4.85	5.21
5 min	5.30	5.70	5.29	5.75

Table 6.3.: The average RMSE error for the Gaussian process, i.i.d., and median filter models. The RMSE errors at various times from the start of each test data segment are also presented.

the initial experiment, Gaussian process regression performs noticeably better than the baseline median method. The sample-and-hold method performed very well in the short-term, providing estimates comparable to those produced by Gaussian processes and the baseline median, but much worse at later times.

In Section 6.4.1, we showed that the median filter method fails in situations where there are large changes in the values of the time series. We now consider the instances in which the Gaussian process model performed particularly poorly on the ED data, by identifying the occasions for which there were large RMS errors. The plots for the test data samples that produced the ten largest RMS errors are shown in Figure 6.5.2. In nine of these cases, the regression estimate seems reasonable, and the poor RMS error is due to unexpected, and unpredictable step-changes in the data during the test period.

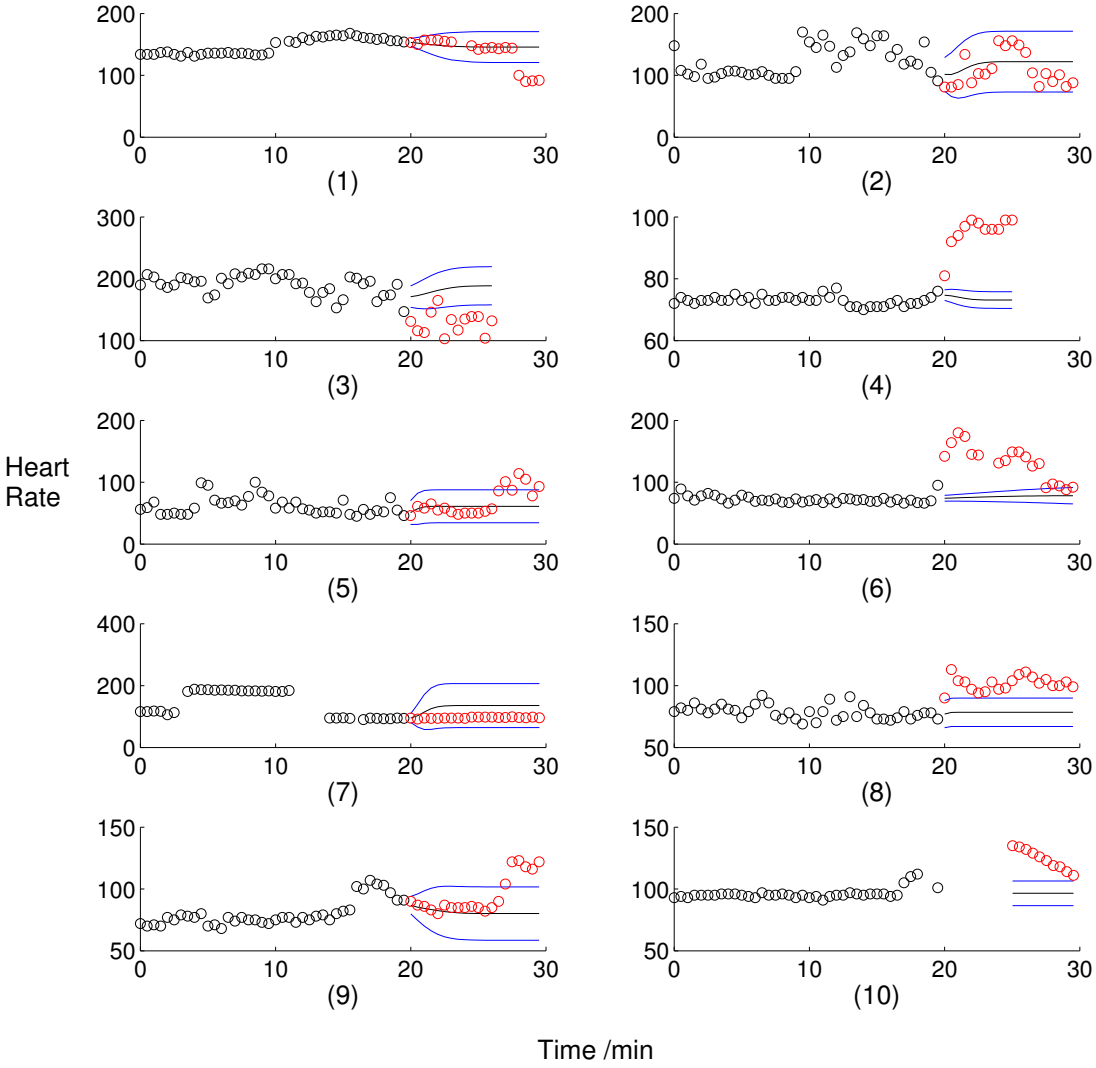

Figure 6.5.2.: The Gaussian process regression estimates for the ten instances with the highest RMS errors. The mean and ±2 s.d. of the posterior Gaussian distribution are shown in black and blue respectively.

For instance, in (1), the overall trend in heart rate data continues at approximately 150 bpm for 8 minutes, before the heart rate drops to 100 bpm, where the Gaussian Process estimate fails. For the remaining case, (7), the training data consisted of three distinct modes; an initial period of HR = 110, a middle period of HR=180, and a final period at HR=95. The regression estimate mean tended towards 135 bpm, and the ±2 s.d. range was between 65 and 205 bpm, indicating a high uncertainty in the result. In this case, the Gaussian process recognises that the training data are spread over a large dynamic range, but is unable to model each of the likely modes with a single Gaussian.

6.5.3. Discussion

In this test of 1-D Gaussian processes, we attempted to infer ten minutes of missing heart rate data, given the previous 20 minutes of data. The aim of this process was originally to provide a principled method of estimating the value of a single vital sign when data were lost due to probe disconnection. Accurate inference of the missing channel of data will then allow us to generate a PSI value without having to revert to a lower-dimensional model.

We assessed the performance of three methods for inferring the missing heart rate data: Gaussian Processes, i.i.d., and median filter methods. The effectiveness of each method was determined by comparing the RMS error between the model estimates and the test data. This allowed us to show that the median filter and Gaussian Process approaches outperform the i.i.d. model. In addition, we showed that the error over the duration of the record increases least for the i.i.d. model, which is expected given the time-independent nature of the model.

The advantage of using RMSE to quantify error is that it allowed a simple comparison between all three models. However, there are two potential problems with the method described in this chapter. Firstly, it is unclear whether RMSE is the most suitable error metric, as it heavily penalises outliers, and may therefore favour methods that may be sub-optimal. One way of minimising with this is to use a different metric such as the mean absolute error, which is defined as:

$$MAE = \frac{1}{n}\sum |y_i - y_{*i}| \tag{6.5.3}$$

or else using a combination of both error metrics through the Huber loss function, which is simply a piecewise loss function that is equivalent to the RMSE below a given threshold, and linear above the threshold [52].

Secondly, the way in which error is measured, as a distance between points generated by the model output and the test data points, does not fully describe the outputs of the i.i.d. and Gaussian process models. Both of these methods produce posterior distributions that must be reduced to a single point (using the distribution median and mean values respectively) for the error to be measured. In doing so, we forego the benefit of having a

distribution from which the confidence in our estimate can be captured.

A more appropriate way of assessing the i.i.d. and Gaussian models is to calculate the likelihood for the test data, using its posterior distribution(s). For the i.i.d. model, this can be calculated as:

$$L_{iid}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta) \tag{6.5.4}$$

where θ is the vector of parameters (i.e. the means and variances of the Parzen kernel centres) , and x are the test data. For the Gaussian processes model, the likelihood must take into account the changing mean and variance of the posterior Gaussian through time:

$$L_{gp}(\theta_1, \theta_2, \theta_n|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta_{\mathbf{i}}) \tag{6.5.5}$$

6.6. Dependent Gaussian Processes

So far, we have introduced univariate Gaussian processes and considered how they may be used to predict values in the future. We then tested the models, and showed that they were effective at making short-term predictions for missing heart rate data. The models that we introduced only considered one channel of vital sign data. We now hypothesise that information about all the vital signs may lead to a narrower posterior over the missing channel of data, as a result of the extra information present in the correlations between vital signs. The hypothesis is based on the knowledge that some of the vital signs are not independent. For instance, heart rate and breathing rate are related; as the heart rate increases, then the breathing rate usually increases (see Figure 6.6.1).

We now show how multiple channels of data can be incorporated into a Gaussian process model. In the first instance, we describe the difficulties in creating dependent Gaussian processes models. We then introduce Gaussian processes in terms of linear filters, which allows simple modelling of inter-channel dependencies. Finally, we construct a simple bivariate example to demonstrate how a single channel of missing data can be inferred using information from two vital sign channels. This will provide an overall framework that allows the use of Gaussian processes in conjunction with a data fusion algorithm

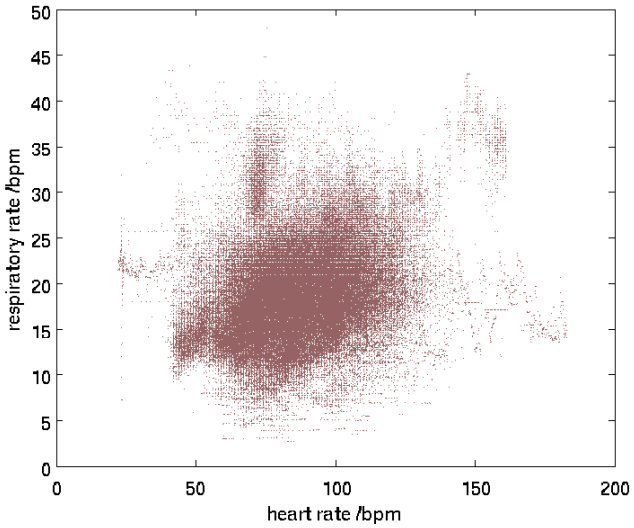

Figure 6.6.1.: Scatter plot of heart rate and breathing rate from the JR data set. It is clear that there is some positive correlation between the parameters. In particular, note that at extremely high heart rates (HR>120), the corresponding breathing rate is highly correlated, belonging to one of two regimes: very fast breathing (RR>35) or normal breathing (15<RR<20).

to output a probabilistic PSI score. We demonstrate how this may be done within the context of the Parzen windows baseline model.

6.6.1. 2D Dependent Gaussian Processes

In this section, we review the mathematics which underpin a dependent Gaussian process model, and then apply such a model to synthetic data. Let us first consider a 2-point example. In Section 6.3.1, we showed that a two-point time series could be represented by two univariate Gaussian distributions, so that the joint probability of the two points is a bivariate Gaussian. By extension, it follows that a 2D vector of points, $[HR_n, RR_n]^T$ can be represented by a 2D Gaussian, and that a two-point time-series can be represented by a 4-D Gaussian. The corresponding covariance matrix thus takes the form:

$$
\begin{aligned}
Cov &= \left[\begin{array}{cc|cc}
\sigma(\mathbf{HR_1},\mathbf{HR_1}) & \sigma(\mathbf{HR_1},\mathbf{HR_2}) & \sigma(HR_1,RR_1) & \sigma(HR_1,RR_2) \\
\sigma(\mathbf{HR_2},\mathbf{HR_1}) & \sigma(\mathbf{HR_2},\mathbf{HR_2}) & \sigma(HR_2,RR_1) & \sigma(HR_2,RR_2) \\
\hline
\sigma(RR_1,HR_1) & \sigma(RR_1,HR_2) & \sigma(\mathbf{RR_1},\mathbf{RR_1}) & \sigma(\mathbf{RR_1},\mathbf{RR_2}) \\
\sigma(RR_2,HR_1) & \sigma(RR_2,HR_2) & \sigma(\mathbf{RR_2},\mathbf{RR_1}) & \sigma(\mathbf{RR},\mathbf{BR_2})
\end{array}\right] \\
&= \left[\begin{array}{cc}
\mathbf{cov_{HR}} & cov_{HR,RR} \\
cov_{HR,RR} & \mathbf{cov_{RR}}
\end{array}\right]
\end{aligned}
\tag{6.6.1}
$$

where $\sigma(x, x')$ is the value of the covariance function between x and x', and the matrix has been labelled to show the dependencies of each of the elements. The elements in bold represent correlations within one channel of data, and can therefore be treated independently. Thus, the bold elements may be described by two independent covariance functions of the form shown in Section 6.3.2, one for HR (Cov_{HR}) and one for RR (Cov_{RR}). The hyperparameters may be estimated as before, treating each channel independently. If the remaining, non-bold, elements are zero, then we have simply collapsed two Gaussian processes into a single covariance matrix.

Unfortunately, there is no simple way to use standard covariance functions to generate the remaining cross-covariance terms in the matrix while still ensuring a positive-definite covariance matrix. An $n \times n$ real matrix, $\mathbf{M}$, is positive definite if $\mathbf{z}^T\mathbf{M}\mathbf{z} > 0$ for all non-zero vectors $\mathbf{z}$. Using covariance functions to compute $cov_{HR,RR}$ and $cov_{RR,HR}$ will result in a strictly positive matrix that is not necessarily positive-definite. For example, the covariance matrix $\left[\begin{array}{cc} \mathbf{a} & \mathbf{b} \\ \mathbf{b} & \mathbf{a} \end{array}\right]$ in which both $\mathbf{a} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right]$ and $\mathbf{b} = \left[\begin{array}{cc} 2 & 0 \\ 0 & 2 \end{array}\right]$ are derived from a squared exponential covariance function, is not positive-definite (e.g. if $\mathbf{z} = [-1\,0\,-1\,0]^T$).

One solution to this problem was proposed by Boyle and Frean [14], which firstly involves recasting Gaussian processes in terms of linear filters. Covariance functions are then formed in terms of the filter's impulse response and the input to the filter. The linear filter approach can then be extended to allow for multiple inputs and outputs, to

describe a fully-dependent model. It is possible to derive valid cross-covariance functions from the impulse response and filter inputs. We now review this approach in more detail.

6.6.2. Gaussian Processes and Linear Filters

First, consider that Gaussian white noise is a special case of the Gaussian process in which the covariance between two points x_i and x_j is σ^2 for $i = j$, and zero otherwise. It is known that if the input to a linear filter is a Gaussian process, then the output is also a Gaussian process [43]. Therefore, we may use the form:

$$y(x) = h(x) * w(x) = \int_{-\infty}^{\infty} h(x-\tau)w(\tau)d\tau \tag{6.6.2}$$

to describe a general Gaussian process, where the input, $w(x)$, is Gaussian white noise, and $h(x)$ is the impulse response of a linear filter. Using this construction, any Gaussian process is now fully described in terms of an impulse response and the noise variance on $w(x)$, and consequently the corresponding covariance function must also be parameterised in terms of these new variables. From first principles, we can define the variance between two points as:

$$Cov(y, y') = E\{yy'\} \tag{6.6.3}$$

and hence the covariance function in terms of $h(x)$ and $w(x)$ is:

$$\begin{aligned} Cov(y, y') &= E\{yy'\} \\ &= E\left\{\int_{-\infty}^{\infty} h(\tau)w(x-\tau)d\tau \int_{-\infty}^{\infty} h(\lambda)w(x'-\lambda)d\lambda\right\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(\lambda)E\{w(x-\tau)w(x'-\lambda)d\tau d\lambda \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(\lambda)\delta(\lambda-(x'-x+\tau))d\tau d\lambda \\ &= \int_{-\infty}^{\infty} h(\tau)h(x'-x+\tau)d\tau \end{aligned} \tag{6.6.4}$$

where δ is the Dirac delta function. In general, Equation 6.6.4 will not have an analytic solution. However, if we choose $h(x)$ to be a Gaussian filter:

$$h(\mathbf{x}) = v\, exp(-\frac{(\mathbf{x}-\mu)^T\mathbf{P}^{-1}(\mathbf{x}-\mu)}{2}) \tag{6.6.5}$$

in which the free parameters ν,$\mathbf{P}$, and μ are the amplitude scale, covariance, and offset respectively, then the covariance function may be calculated analytically, resulting in a squared exponential. The full derivation of this result can be found in [13]. The solution involves the convolution of two Gaussians:

$$c(\mathbf{s}) = Cov(\mathbf{y}, \mathbf{y}') = \frac{v^2(2\pi)}{\sqrt{2\mathbf{P}^{-1}}} exp\left(-\frac{1}{2}\mathbf{s}^T\left(\frac{\mathbf{P}^{-1}}{2}\right)\mathbf{s}\right) \tag{6.6.6}$$

where the filter is time invariant, so the covariance function is fully described in terms of $\mathbf{s} = x' - x$. The advantage of the linear filter approach is that it may be generalised for M Gaussian noise inputs, and N Gaussian processes outputs, so that the N filter outputs, y_n are given by the sum:

$$y_n(t) = \sum_{m=1}^{M} \int_{-\infty}^{\infty} h_{mn}(\tau) w_m(t-\tau) d\tau \tag{6.6.7}$$

Each output is a weighted sum of filtered inputs, with the weights incorporated within h_{mn}. In this case, the output Gaussian processes are also dependent, as they rely on a common set of M inputs.

6.6.3. 2D Example

Using the linear filter approach developed in the previous section, we now consider an example with two channels of univariate data, for which we would like to infer values for one of the channels based on information from previous data samples (as in standard Gaussian processes), and also on information from the correlations between the channels.

Boyle and Frean proposed a 2D dependent Gaussian process model such that the Gaussian process output for each of the two channels was composed of an independent Gaussian process and another Gaussian process common to both channels. In addition to this, each of the channels is subject to measurement noise, which can be modelled with additional Gaussian white noise sources. We adopt this approach, which is represented pictorially in Figure 6.6.2.

The noiseless outputs, (z_1, z_2) can be thought of as arising from a 3-input, 2-output model of the form described in Section 6.6.2 where h_{mn} can be found from the elements

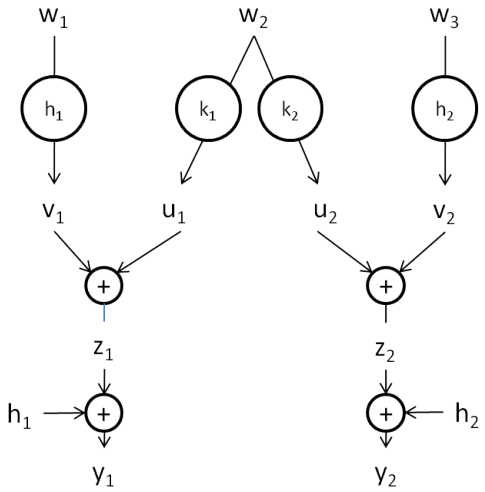

Figure 6.6.2.: Model for two dependent Gaussian Processes y_1 and y_2, adapted from Boyle and Frean [14]. The clean outputs, z_1 and z_2 are the sum of an independent Gaussian process and a common Gaussian process, and can be thought of as a 3-input, 2-output model of the form given in Equation 6.6.8. The clean outputs are then subjected to additive Gaussian noise, which represents measurement error.

of the 3×2 impulse response matrix:

$$\mathbf{h}(x) = \begin{bmatrix} h_1 & 0 \\ k_1 & k_2 \\ 0 & h_2 \end{bmatrix} \tag{6.6.8}$$

For this example, we choose each filter to be Gaussian so that:

$$h_i(x) = w_i \exp(-\frac{x^2}{2Q_i^{-1}}) \qquad where\, i = 1, 2 \tag{6.6.9}$$

$$k_1(x) = v_1 \exp(-\frac{x^2}{2P^{-1}}) \tag{6.6.10}$$

$$k_2(x) = v_2 \exp(-\frac{(x-\mu)^2}{2P^{-1}}) \tag{6.6.11}$$

The Gaussian filters k_1 and k_2 model the dependent Gaussian processes common to both y_1 and y_2. The term μ in Equation 6.6.11 merely allows the dependency between y_1 and y_2 to be phase shifted by an amount μ, and the term could equally have been included in $k_1(x)$ instead of $k_2(x)$. The free parameters in the filters are described by an

amplitude matrix, v, and variance matrix, A:

$$\mathbf{v} = \begin{bmatrix} w_1 & 0 \\ v_1 & v_2 \\ 0 & w_2 \end{bmatrix} \quad , \quad \mathbf{A} = \begin{bmatrix} Q_1 & 0 \\ P & P \\ 0 & Q_2 \end{bmatrix} \tag{6.6.12}$$

Now that the model has been fully parameterised, the remaining task is to construct the covariance matrix.

We attempt to find $cov_{ij}^{z}(x_i, x_j)$, by generalising Equation 6.6.4 using the property that the sum of covariance functions is itself a covariance function, so that cov_{ij}^{z} can be written as:

$$cov_{ij}^{z}(\mathbf{d}) = \sum \int h_{mj}(\tau) h_{mi}(\tau + \mathbf{d}) d\tau \tag{6.6.13}$$

where $z \in (v_1, v_2, u_1, u_2)$ and $\mathbf{d} = |\mathbf{x}_j - \mathbf{x}_i|$

By setting h to be a Gaussian filter, it is possible to derive the general covariance function, which is a squared exponential:

$$cov_{ij}^{u}(d) = \sum_{m=1}^{M} \frac{(2\pi)^{1/2} v_{mi} v_{mj}}{\sqrt{|\mathbf{P}_{mj} + \mathbf{P}_{mi}|}} exp(-\frac{1}{2}(\mathbf{d} - [\mu_{mi} - \mu_{mj}])^T \mathbf{\Sigma} (\mathbf{d} - [\mu_{mi} - \mu_{mj}])^T \tag{6.6.14}$$

with $\mathbf{\Sigma} = \mathbf{A}_{mi}(\mathbf{A}_{mi} + \mathbf{A}_{mj})^{-1}\mathbf{A}_{mj}$. In our problem, we consider only one input dimension, so that $\mathbf{d}$, $\mathbf{x}_n$ and $\mathbf{P}$ are all scalar quantities. Then the covariance matrix for the noisy model is simply:

$$cov_{ij}^{y}(d) = cov_{ij}^{z} + \delta\sigma_1^2 \tag{6.6.15}$$

Noting that there is no linear filter $H_{1,2} = H_{3,1} = 0$, then $v_{mi}v_{mj}$ is only non-zero when $m = 2$, which gives:

$$cov_{11}^{y}(d) = cov_{11}^{u} + cov_{11}^{v} + \delta_{ab}\sigma_1^2 \tag{6.6.16}$$

$$cov_{22}^{y}(d) = cov_{22}^{u} + cov_{22}^{v} + \delta_{ab}\sigma_2^2 \tag{6.6.17}$$

$$cov_{21}^{y}(d) = cov_{21}^{u} \tag{6.6.18}$$

$$cov_{12}^{y}(d) = cov_{12}^{u} \tag{6.6.19}$$

where the distance between two points is $d = x_a - x_b$, and δ_{ab} is the Kronecker delta function. Using these 4 equations, the full covariance matrix can be assembled. In total, there are now nine Gaussian process hyperparameters: $v_1, v_2, w_1, w_2, P_1, P_2, Q_1, Q_2, \mu$ and two further unknowns which represent the measurement noise, σ_1^2, σ_2^2. We can estimate values for the hyperparameters in the same way as for the 1-D case, by minimising the negative log likelihood. The negative log likelihood is exactly analogous to Equation 6.3.3:

$$L = \frac{1}{2} log|\mathbf{K}| + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{n_1 + n_2}{2} log 2\pi \tag{6.6.20}$$

where $\mathbf{K}$ is the full covariance matrix for the dependent Gaussian processes.

Now that a dependent Gaussian process model has been created, regression can again be calculated as a conditional slice through the joint distribution described by the covariance matrix. If we want to regress onto a new point $x*$, then the regression equations are:

$$\begin{aligned} mean(x*) &= \mathbf{C}^T\mathbf{B}^{-1}\mathbf{y} \\ variance(x*) &= \kappa - \mathbf{C}^T\mathbf{B}^{-1}\mathbf{C} \end{aligned} \tag{6.6.21}$$

where $\mathbf{C}$ and $\mathbf{B}$ are the sub-matrices defined in Equation 6.3.6, and $\kappa = v_i^2 + w_i^2 + \sigma_i^2$.

The model was tested on the synthetic data set shown in Figure 6.6.3. The first channel of data is composed of equally-spaced points from the sine wave $Y_1 = sin(5x)$ with additive Gaussian noise (with variance $\sigma_1^2 = 0.125$). The second channel of data was also composed of a noisy sine wave, but had been phase shifted by $x = 0.5$, so $Y_2 = sin(5x - 0.5)$. The dependent Gaussian process model described in Figure 6.6.2 was then used to estimate values for the second channel of data between $0.6 < x < 1.5$. The hyperparameters were selected using gradient descent methods. The result of the Gaussian process regression is shown by the black and blue lines, which take the same meanings as before. In comparison, the univariate regression is also shown on the same figure in light red.

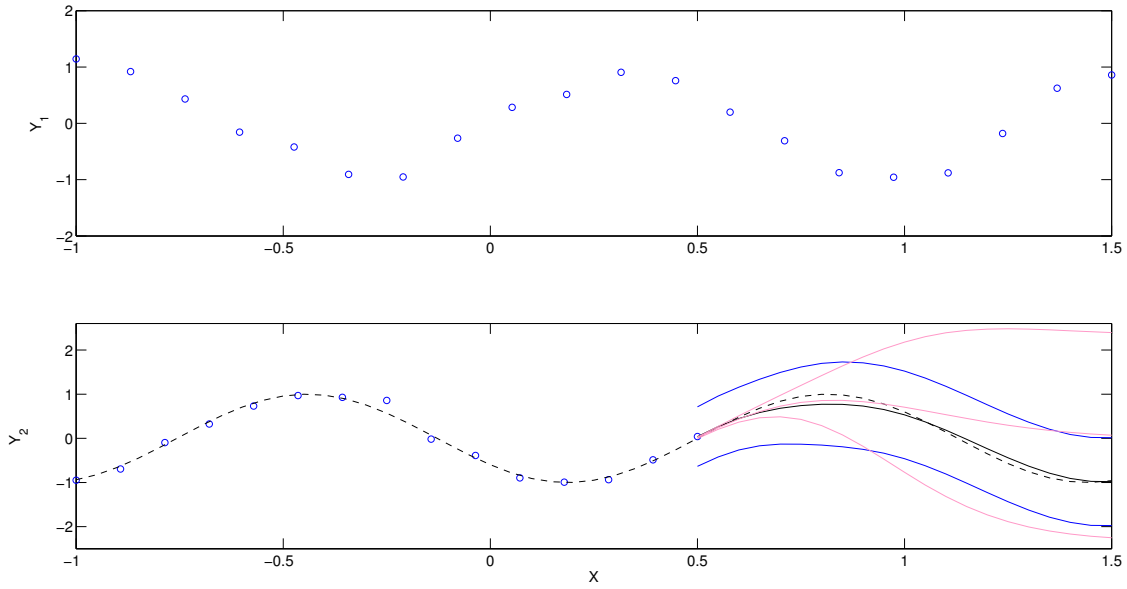

Figure 6.6.3.: An example of a bivariate Gaussian process model for two single channels of data, Y_1 and Y_2. In this example, Y_1 is known for the duration of the time series, and we attempt to predict the missing values of Y_2. The dependent Gaussian process model performs better than the univariate case, as additional information regarding correlations between the vital signs is included in the model.

The dependent Gaussian process was able to predict the underlying sinusoidal shape due to the common input process w_0. In this case, the Y_1 and Y_2 are entirely dependent, and as such, the *maximum a posteriori* estimate for the hyperparameters that model the magnitude of the independent part of the Gaussian processes were $w_1 = w_2 = 0$. In contrast, the univariate Gaussian process is able to correctly predict the short-term increase in Y_2, based on the previous few values of data, but unable to identify the longer-term sinusoidal trend.

6.6.4. Summary

We have demonstrated how the principles for Gaussian processes can be extended from the 1-D to the 2-D case and shown how dependencies between two Gaussian processes can be modelled. In the 2-D case, the covariance functions used previously were no longer adequate for providing a positive-definite covariance matrix in the case of dependent Gaussian processes. Instead, we considered Gaussian processes as linear filters with a Gaussian white noise as input. This makes it possible to generate covariance functions

and cross-covariance functions, parameterised in terms of the filter impulse response, and the variance of the input noise.

The resulting 2-D model was tested on synthetic data, from which it was clear that modelling the dependencies between the two channels allowed for a more accurate estimate of missing data. Although we have only demonstrated the case for 2 dependent processes, the extension to N dependent processes is straightforward, requiring one additional Gaussian filter to model the independent component of each additional output, and two additional Gaussian filters for each dependency between the outputs that are analogous with filters k_1 and k_2 in Figure 6.6.2.

6.7. Gaussian Processes for Data Loss

Until now, we have used Gaussian processes to infer missing vital sign data, but we have not been concerned with how this may be used within a patient monitoring system. We now present a full framework for dealing with data loss within the baseline model. The baseline model, introduced in Chapter 4, uses the Parzen windows algorithm to generate a Patient Status Index (PSI). Although we describe the framework for one particular case, it is easily generalisable to the other data fusion methods described in Chapter 4. The description below is accompanied by Figure 6.7.1, which provides a 2D visualisation of each stage.

1. *Infer the missing vital sign data* - Infer the posterior distribution for the missing vital sign using dependent Gaussian processes, following the procedure outlined in Section 6.6, for four dimensions. The Gaussian Process estimate will be dependent on both the previous values of the missing channel, and correlations with the other vital signs, and the posterior distribution will be a *N-D* Gaussian if N channels of data are missing. In the 2D example in Figure 6.7.1(a), one vital sign is missing, which produces a univariate Gaussian posterior.

2. *Calculate the PSI score distribution* - The output from the Gaussian processes model can now be interpreted within the context of our data fusion model of choice. For now, we consider only the baseline Parzen windows model, but the method de-

scribed here can be generalised to other data fusion models. In the Parzen windows model, a 4D distribution is used to estimate the probability of a vital sign vector, $[V_1, V_2, V_3, V_4]^T$. If all of the vital signs are known exactly, then a single point on the distribution can be defined. The resulting probability, $P(V_N)$ can then be converted into a PSI value through a negative log transformation, as described in Equation 4.1.7.

Let us now consider the case in which only three vital signs, V_1, V_2, V_3, are known to have the values v_1, v_2, v_3 at one particular instant in time. The missing vital sign, V_4, takes a value according to a normal distribution $\mathcal{P}(v_4) \sim \mathcal{N}(\mu, \sigma^2)$, in which the mean, μ, and variance, σ^2, are determined by the output of a Gaussian process regression, as described in step 1.

If we interpret the vector $[V_1, V_2, V_3, N(\mu, \sigma^2)]^T$ using the Parzen windows probability distribution, we observe that the value of the distribution in the V_1, V_2 and V_3 dimensions are fixed, while the remaining dimension in the V_4 direction is constrained by the normal distribution $\mathcal{N} \sim (\mu, \sigma^2)$. An analogous 2D example, showing V_1 and V_4, is shown in Figure 6.7.1(b). Here, the probability $P(V_N)$ is no longer a single value, but a 1D distribution that can be described in terms of conditional probability:

$$P(V_N) = P(V_1, V_2, V_3, V_4 | v_1, v_2, v_3, \mu, \sigma^2) \tag{6.7.1}$$

where the joint distribution $P(V_1, V_2, V_3, V_4)$ is merely the Parzen windows model. More generally, if n of the vital signs are not known exactly, then the output of the Parzen windows model will be an N-D distribution over a subspace of the space enclosed by $P(V_1, V_2, V_3, V_4)$. A closed-form solution for the conditional probability, $P(V_N)$, is not possible as, in the limit, the Parzen windows model that generates $P(V_1, V_2, V_3, V_4)$ can model any possible distribution [28]. Instead, we can generate a set of samples from $P(V_N)$, by sampling points from the joint distribution that

meet the constraints on each of the vital signs, $V_1 = v_1, V_2 = v_2, V_3 = v_3$. The set of samples can then be converted into a discrete probability distribution that approximates $P(V_N)$. By using a negative log transform, the $P(V_N)$ distribution can be converted into an 1-D cumulative probability distribution over PSI scores, $P(PSI)$ by:

$$P(PSI < x) = \int_{e^{-x}}^{\infty} P(V_N) dV_N \tag{6.7.2}$$

3. *Determine whether an alert should be generated* - In steps 1 and 2, we have described a novel method of dealing with uncertainty in the vital sign measurements by propagating the uncertainty through a data fusion system, providing a probabilistic output. We now consider how this probabilistic output may be used to generate alerts. In the original baseline model, we decided to alert if 4 out of the previous 5 minutes of data, that is, 80% of the data, were above a pre-determined alerting threshold. For each vital sign vector input to the baseline model, a single PSI score is generated, which lies either above or below the alerting threshold (see Figure 6.7.1(c)).

   In our novel method, we allow the vital signs to be distributions rather than single values, and hence the PSI for each vital sign vector can now be described as a distribution, $P(PSI)$ (see step 2).

   Consider first the 1D case, which is shown in Figure 6.7.1(c). In this case $P(PSI)$ is a 1D distribution over PSI scores which may not lie entirely above or below the alerting threshold, but may contribute some probability mass to both sides, as shown in the Figure. In keeping with precedent, we still alert if 80% of the probability mass of the PSI scores is above the threshold. Mathematically, an alert should be generated if:

$$\frac{\sum_{T=0}^{T \leq M} P(PSI > \zeta)_T}{\sum_{T=0}^{T \leq M} P(PSI < \infty)_T} = \frac{\sum_{T=0}^{T<M} \int_{e^{-\zeta}}^{\infty} P(V_N)_T dV_N}{M} \geq 0.8 \tag{6.7.3}$$

where ζ is the alerting threshold, and there are a total of M vital sign input vectors within a 5 minute period. In the case that all of the vital sign variables are known and the PSI is a single value, then $P(PSI < \zeta)$ is simply a Dirac delta function centred on the PSI score. For the N-dimensional case, the above method is entirely applicable as the cumulative distribution function $P(PSI < x)$ is 1-D in all cases. So far, we have considered the use of Gaussian processes in cases for which there is a single channel of missing data. However, there is no reason why we cannot infer more than one variable at a time using the same methodology.

6.7.1. Trend Analysis

In the limit, it is possible to estimate posterior distributions for all of the vital sign parameters at a future point in time. Having predicted the future distribution of each of the vital signs, we can also predict a corresponding PSI value distribution.

We can use the framework outlined in the previous section to generate the PSI score distribution. This time, unlike the 2D example in Figure 6.7.1, we sample from the Parzen windows model in a region defined by a 4D Gaussian, which represents uncertainty in all four vital signs, rather than a 1D slice. The output of this process is also a 4-D distribution of $P(PSI)$. By sampling from this distribution, and remembering that $PSI \sim -logP(PSI)$, we reduce the 4-D distribution to a 1-D distribution of PSI values. Alerts can then be generated as we described previously.

The effectiveness of the proposed framework has not yet been tested, and is left as an opportunity for future work. Figure 6.7.2 shows one preliminary example, in which a simple (non-dependent) Gaussian process model was used to estimate the posterior distribution of the missing heart rate data at times between 37 and 70 minutes into the record. 1000 samples from each posterior distribution were then taken in order to calculate a PSI distribution. The mean of the distribution is indicated by the magenta line, and its 5^{th} and 95^{th} percentiles are shown in blue. Unlike local median, or the population mean methods of dealing with missing data, the PSI distribution does not introduce artificial jumps in the PSI score. This can be seen at 37 minutes into the record, when breathing rate data is first unavailable. In addition to producing a consistent estimate,

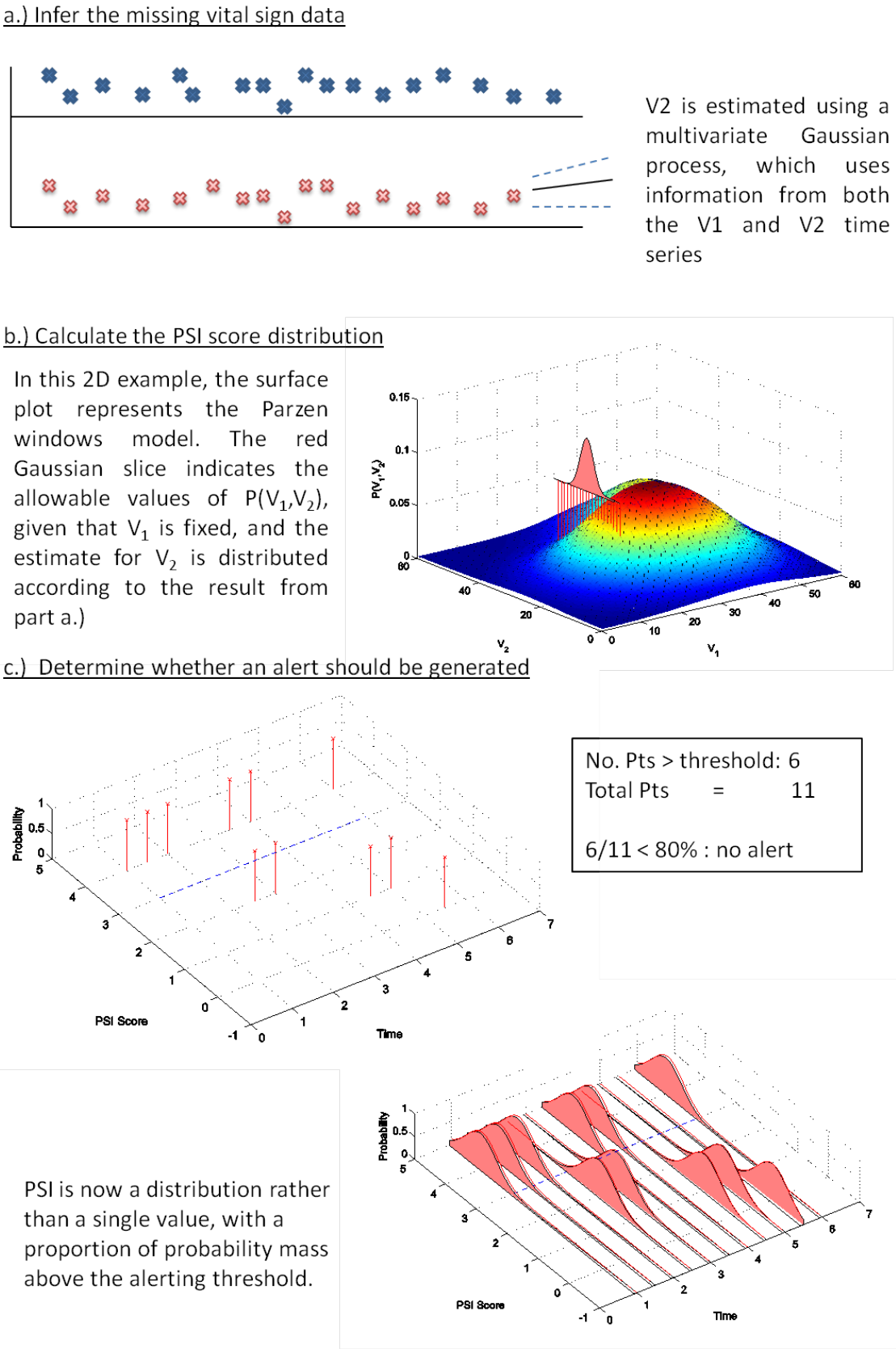

Figure 6.7.1.: Pictorial description of the Gaussian Process framework for generating missing data within an alerting-system framework.

the Gaussian process-assisted PSI is able to generate appropriate alerts when data is missing. In Figure 6.7.2, we see that the mean of the PSI distribution, and the 5th and 95th percentiles, exceed the PSI alerting threshold of 3 at 50 minutes into the record, when the SDA drops to 63 mmHg and would therefore generate an alert.

In contrast, the alternative approach of collapsing the missing data dimensions or using the population mean would be less likely to produce an alert. Both of these methods attempt to make the missing data channel uninformative and have no way of interpreting the historical heart rate trend of bradycardia. For instance, by setting the Heart Rate to the population mean of 83.7 bpm, the patient's PSI will appear to much more normal than their true underlying state.

It is likely that in many instances, the single channel posterior becomes largely uninformative within a few minutes (for instance, see the examples in Figure 6.5.2). This will in turn lead to a wide distribution of PSI values, so that a PSI alert based on future data is unlikely to occur. However, a PSI alert may be generated in some instances when there are clear trends in the data (for instance, see the example in Figure 6.4.2). On these occasions we can expect the Gaussian process to have a tighter posterior distribution, denoting a higher confidence in the estimate, which results in a narrower distribution of PSI values.

6.8. Discussion

In this chapter, we have highlighted how Gaussian processes may be used to infer missing single channels of data, and also how we may extrapolate to estimate a whole vector of vital sign data. The Gaussian process model provides a posterior distribution so that the level of certainty in the inferred data can be estimated, a task which has not been previously attempted in this context. We compared the Gaussian process model to two other techniques; a short-term median filter; and an i.i.d. model, which does not use time-series information, but does provide a posterior distribution.

By calculating the RMS error between the model predictions and the test data, we showed that Gaussian processes were superior to the i.i.d. model, and comparable to the median-filter method. The Gaussian process model predicted the test data most accu-

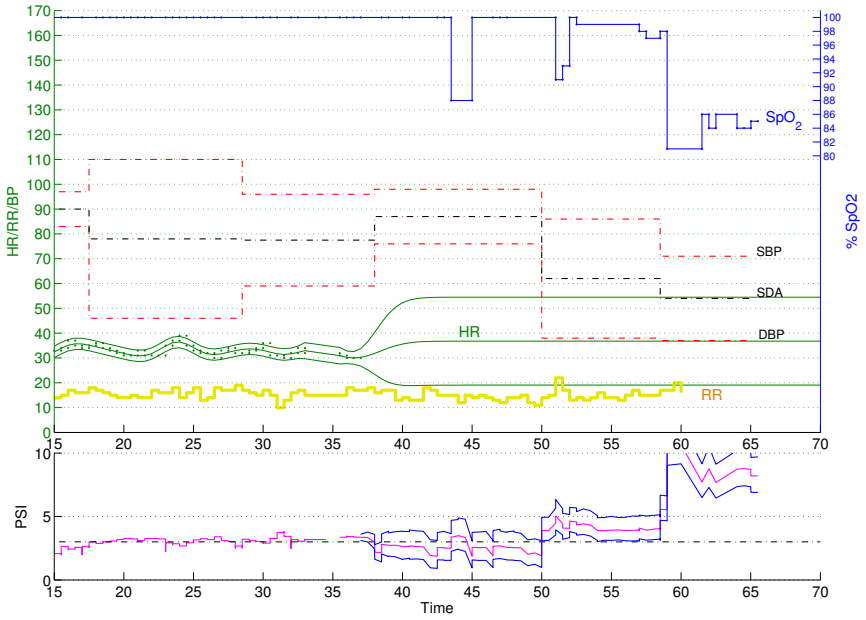

Figure 6.7.2.: Preliminary results using the Gaussian process model framework to generate PSI distributions as a means of dealing with missing channels of vital sign data. The outer green lines show the mean of the estimated HR posterior distribution±2 s.d. The uncertainty in the HR estimate is carried over into the PSI calculation, where the outer blue lines indicate the 5^{th} and 95^{th}percentiles of the PSI distribution.

rately (along with the median filter method), and also provided a posterior distribution of vital sign values. The median filter and i.i.d. approaches were shown to have poor predictive capabilities wherever there are rapid changes in the time series, whereas the Gaussian process model failed when there were two separate regimes underlying the data generation, as shown for example in Figure 6.5.2. A suitable choice of covariance function may be able to correct for this problem, though a non-Gaussian approach may be more appropriate. We also showed in this chapter how an estimate of a single vital sign may be improved by modelling dependencies from other vital signs within a multiple output Gaussian process model.

In Section 6.7, we outlined a framework for applying Gaussian processes within the baseline data fusion model. The value of this framework is twofold. Firstly, it allows the PSI to be calculated, and alerts to be generated during periods of missing data, avoiding the problems associated with switching to a lower-dimensional data fusion model. Secondly, it provides a method of estimating PSI values in the future. If the PSI is above

the alerting threshold at a future time, then we may alert with some degree of early warning. This is an improvement on the methodologies considered in the rest of this thesis, which only examine the most current vital sign vector.

In conclusion, we have shown that the Gaussian process framework provides a way of dealing with missing channels of data and offers the possibility of providing early warning of deterioration by using a PSI distribution based on future estimates of the vital sign data. The work presented in this chapter has a number of limitations, however; in particular, the choice of covariance function was generic to many types of time series, and better performance may be gained by tailoring the covariance function to match prior knowledge about the vital signs, such as the presence of circadian rhythms.

7. Conclusion

7.1. Summary of Results

Physiological observations in the Emergency Department (ED) are a required part of patient care, and are used to monitor the condition of patients in the Department. Manual observations by nursing staff are routinely recorded at approximately hourly intervals, and clinical decisions based on the observations are facilitated by a Track and Trigger (T&T) system. Under this system, observations are converted into a T&T score, in which a higher score reflects greater physiological abnormality. Once the score reaches a critical threshold, a medical intervention is triggered.

The effectiveness of T&T in the ED has been investigated only once previously to the best of our knowledge [61]. In Chapter 2, we described a study of the (T&T) system within the ED that was carried out at the John Radcliffe hospital, Oxford during 2009. 500 patients attending the Majors, Resus, and CDU areas of the department were recruited to the study. The aim of the study was to investigate how well integrated the use of T&T was within the ED and to quantify how effective the system was at identifying escalation events. Escalation events were defined as any documented instances requiring intervention by clinical staff. In addition, we aimed to identify some possible errors made in the use of the observation charts for recording vital signs.

The study showed that T&T completion was poor, with only 34.3% of overall scores calculated. In the instances where T&T scoring had been completed, 20% had been calculated incorrectly. Incorrect T&T scores primarily occurred when a vital sign was assigned an incorrect T&T score. In a small number of cases, 14 out of 202, the total T&T score was not added up correctly. We also showed instances of illegible observation charts and demonstrated that errors may also be caused by an over-reliance on bedside

monitoring equipment. The T&T system, as used by nursing staff in the Department, had a sensitivity of 0.47 and specificity of 0.87 for identifying escalation events. Subsequent to the end of the study, the observation charts were redesigned to minimise human error and to improve their legibility.

To investigate the effect of improved T&T completion and calculation, we computed a retrospective T&T score that was calculated without error from the manually-recorded observations. Using this score, the errors due to incorrect assignment and incorrect addition were eliminated. The sensitivity and specificity of this retrospective system were 0.94 and 0.70 respectively, a very clear improvement.

In addition to nurse observations, higher acuity patients such as those in the Major and Resus sections of the Department also have their vital signs monitored continuously by a bedside monitor. The benefit of continuous monitoring is that a patient's condition can be monitored in real time, and in principle, deterioration can be detected as soon as it occurs, even if this is between nurse observations. Single-channel alerts from the monitors bring nurses to the bedside when the given vital sign is outside of normal limits. However, a high percentage of these alerts are false (up to 86%) [115], and consequently the alerts are routinely ignored.

In Chapter 3, we initially investigated a method for monitoring continuous vital signs based on the T&T scoring system. The T&T scoring criteria were applied to the continuous data to simulate the effect of observing them at one-minute intervals. Using this system, we detected most of the physiological escalations, but in doing so, also generated a very large number of false alerts. We estimated that a continuous T&T system would generate one alert every 5 minutes in a 20-bed Emergency Department. A partial solution is the introduction of a persistence criterion to limit the effect of transient observations.

In Chapter 4, a baseline data fusion system was introduced, which showed how a threshold for physiological abnormality can be derived from the integration of vital sign data using a probabilistic data fusion model, rather than from a heuristic set of criteria (such as the T&T criteria). Two improved data fusion methods, weighted Parzen windows and Support Vector Machines, were then reviewed. The best data fusion models had a sensitivity of 58.6% and a specificity of 83.4% on the ED test set. It was not possible to

select the optimal model, due to the difficulty in assessing the relative importance of False Positives and False Negatives on the initial data set.

The sensitivity and specificity values were lower than those reported for the computer-assisted, intermittent, T&T system. However, the sensitivity was biased towards the T&T method as a result of the method for the determination of escalation events. Furthermore, the specificity for the data fusion systems tended to be relatively low, as false alerts are more likely to be generated when data are monitored continuously. It is also possible that there are periods of patient instability that resolve by the time of the next observation.

In addition, it was clear from inspection of the vital sign records that the models performed sub-optimally as a result of using a training data set that was not optimal for vital sign monitoring within an ED environment, and an inappropriate choice of the Systolic-Diastolic Average blood pressure feature. Heuristic changes to the pre-processing of the blood pressure and oxygen saturation data improved the baseline model, leading to a sensitivity and specificity of 69.0% and 69.6% respectively, on the ED test set.

Two further limitations of the data fusion models were highlighted in Chapter 6. Firstly, we noted that the method of dealing with missing data in the baseline model leads to sudden changes in the Patient Status Index that are unrelated to patient physiology. Secondly, we observed that the existing data fusion models do not make any use of temporal information.

One method of incorporating historical data into a data fusion model, Gaussian processes, was considered in Section 6.3.1. With this method, previous data are used to derive a Gaussian posterior probability estimate of a missing data point. The mean of the Gaussian can be considered as the *maximum a posteriori* estimate, and the variance of the Gaussian provides an indication of the confidence in the estimate. Using this method on single-channel heart rate data, it was shown that the Gaussian process method performs at least as well as a short-term median filter, when assessed with a root-mean-squared error metric.

We then described how a Gaussian process-based model can be improved by including dependencies between vital signs. We showed how a method of achieving dependencies, devised by Boyle and Frean [14], could be applied to two channels of input data. Finally,

we showed how Gaussian processes could be used within a data fusion system so that Patient Status Indices could be represented as a distribution when one or more vital signs are unknown. In a similar way, alerts may be generated using posterior distributions.

In conclusion, we have shown that manual vital sign observations are limited by human factors, and showed how computer systems could be used to improve the completion rate and documentation accuracy of observations. Continuous monitoring systems can be used to detect deterioration in real time, which provides early-warning with respect to the nurse observations. The false alert rate of each data fusion systems tested in this thesis was much lower than continuous T&T, but a modified model was shown to match the sensitivity of continuous T&T. However, the true performance of data fusion systems cannot be properly assessed on the 402 patients in the ED test set, since only 29 patients had physiological escalations post-arrival.

Time-series techniques could provide further improvements to the data fusion models, especially with respect to the *prediction* of patient deterioration, rather than simply detecting it.

7.2. Future Work

The future direction of the work described here should be split into two themes. Firstly, the conclusions regarding the effectiveness of T&T and continuous vital sign monitoring shall be further evaluated in an intervention study, rather than the retrospective study described in this thesis. Secondly, outstanding theoretical concerns also need to be addressed. We discuss these themes in more detail in the next two sections.

7.2.1. Design of an ED Intervention Study

The study described in Chapter 2 was limited by three main factors. Firstly, all of the analysis was performed retrospectively; a more powerful study design would involve a clinical intervention that could be assessed prospectively. Secondly, the outcome measure in the ED study, escalations, was usually triggered by the T&T alerting thresholds being met, and were therefore biased towards T&T scores. Finally, the study was conducted on a relatively small sample size of 500 patients. From these, only a small number of patients,

29, had physiological escalations that occurred after arrival, meaning that results are subject to wide confidence bounds.

A new study is currently being planned with the following aims: (i) to determine whether the rate of T&T completion limits the system's effectiveness at identifying patient deterioration; (ii) to determine whether continuous monitoring leads to earlier interventions, and (iii) to quantify the added benefit of a continuous monitoring system.

The new study will include all patients attending the Majors section of the ED over the course of 6 months. Internal figures give an estimate of 400 patients entering the Majors section of the ED each week, which leads to a conservative estimate of 9000 patients to be included in the 6-month study.

Each phase of the study will last for two months, with a two week training period in between each phase. In the first phase, all patients will continue with standard care. Observations and T&T scores will be recorded manually on the new T&T charts.

In the second phase, an electronic T&T system, known as VitalPAC [89], will be introduced on the ward, and paper observation charts will be withdrawn. VitalPAC is a system designed to facilitate the completion of accurate and prompt observations. It uses a central station to record the locations of each of the patients on the ward, and indicates when the next set of observations are due. Nurses will use handheld devices (an Apple iTouch) to record the vital sign observations electronically, and the T&T score will automatically be calculated by the iTouch. If the T&T score exceeds the alerting threshold, a visual alert is displayed on a central station which shows the status of all patients in the Department. The introduction of this system to the ED is an attempt to eliminate some of the human errors associated with T&T scores, and to increase the promptness and completion of observations.

In the third phase of the study, nurse observations will continue to be recorded electronically. In addition, the modified baseline algorithm will process the continuous vital sign data from the bedside monitors, and audible alerts will be generated when it detects vital sign abnormality. From a scientific viewpoint, the order of the three study phases should be randomised. However, practically, it was deemed impractical to return to paper-based observations after the introduction of an electronic system.

Outcome Measures

In the clinical study described in this thesis, we attempted to match vital sign data to "escalation" events. As well as being biased, the process was extremely laborious, requiring independent, retrospective interpretation of each set of patient notes by two members of the senior clinical team. In the new study, the outcome measure will be the change in 30-day mortality, 24-hour mortality, and the number of cardiac arrests within the ED before and after the interventions.

An additional outcome measure will aim to determine whether there is a reduction in the frequency and duration of periods of physiological abnormality after each intervention. Physiological abnormality will be deemed to occur whenever the values of the continuously-measured HR, RR or SpO_2 and intermittently-measured blood pressure give rise to a single-channel T&T score of 3 or greater. We hypothesise that real-time alerts triggered by a data fusion system will result in prompter intervention. This in turn could lead to faster treatment and a reduction in the duration of patient physiological abnormality.

7.2.2. Improvements to Data Fusion Models

Selection of Optimal Vital Sign Variables

In Chapter 5, we adapted the baseline model to be more sensitive to hypotension events by adjusting the Systolic-Diastolic Average (SDA) scaling factor. However, in Section 5.4.2, we observed that under certain circumstances, the SDA blood pressure cannot accurately model the change in a patient's condition. Furthermore, we noted that the SDA has no precedent in physiological vital sign monitoring.

We postulate that one further improvement to the model would be the use of a clinically validated blood pressure parameter in the place of SDA. As well as the Systolic and Diastolic blood pressures, there are at least two blood pressure variables that combine the Systolic and Diastolic blood pressure in a clinically relevant way. These are the pulse pressure, $BP_{PP} = SBP - DBP$, and the Mean Arterial Pressure, $BP_{MAP} \approx \frac{1}{3}(SBP - DBP) + DBP$. In particular, the Mean Arterial Pressure (MAP) may be a useful feature as it is commonly considered as the average blood pressure in an individual

[19] and automated blood pressure monitors measure the MAP directly.

Gaussian Processes

In Chapter 6.3, we introduced Gaussian process regression as a tool for modelling vital sign data. Unlike the other models considered in this thesis, the Gaussian processes approach takes into account correlations in time between successive vital sign values, and we showed that such an approach was more accurate than making an i.i.d. assumption. The Gaussian process methodology also allows us, in principle, to predict the patient's state in the future, and allows us to deal with missing data.

The quality of the output of the statistical methods shown here depends on the quality of the data input to the model. Despite the inclusion of rudimentary algorithms for determining data loss within the Philips monitors themselves, a good quality of data cannot be assumed. For instance, Figure 6.7.2 shows physiologically suspicious drops and recovery in SpO_2 at 45 and 50 minutes into the record. At the very least, some sort of signal quality index would provide confidence over whether short-term trends were artefactual or genuine. A rudimentary signal quality index could be built by determining common temporal trends for probe disattachment (for instance, see Hann [41], for one method for estimating thermistor probe disconnection). This could be further enhanced by using information from multiple data channels. The most effective signal quality indices can be achieved only with access to the underlying waveform data; signal quality algorithms using raw waveforms is already an area of active research (see, for example, [68]).

Even if good data quality can be ascertained, it is unclear whether the sampling rate used in Chapter 6, 30 s, is sufficient for estimating short-term trends. In this respect, we were limited by the data output by the bedside monitors. However, a principled lower-bound on the sampling rate can be simply derived using a data set that contains the raw waveform data, such as the MIMIC waveform database [38, 74]. By using the analysis presented in Section 6.5, in which a test segment of data was estimated from previous training data, for a range of sampling frequencies, it should be possible to determine at what point the performance of the Gaussian Process estimate diminishes by using an appropriate metric (such as the Huber metric) to quantify the quality of the estimate.

Gaussian process regression also requires a covariance function which determines the temporal relationship between vital sign measurements. The choice of the covariance function is critical, as it represents our prior knowledge of the vital sign trends. In the initial investigations in Section 6.5, we chose to use the squared exponential covariance function, which makes the valid, but weak assumption that observations closer together in time are more highly correlated than those far apart in time.

Future work should consider whether the squared exponential covariance function is appropriate. Stegle et al. [102] used visual inspection to determine that there were both short-term and long-term interactions within their heart rate data set, and consequently chose to model the vital sign record with a sum of standard covariance functions. While their heuristic approach led to positive results, an alternative approach which may lead to better results would be to run a number of gaussian process models in parallel for a set of covariance functions. An objective estimate of the best model can then be selected by maximising the marginal likelihood over the covariance function set and their respective hyperparameters for each vital sign record. The most selected covariance function over all vital sign records can then be considered as the optimal function.

It is possible that a single covariance function may not be optimal in all cases, particularly for a heterogeneous patient population, when the vital signs may be dependent on an underlying physiological condition. An alternative method of determining a covariance function would be to cluster "similar" groups of vital sign records. Similarity may either be defined with respect to the clinical diagnosis, or else using metric based directly on the data, such as cross-correlation. An optimal covariance function could then be derived for each group using the methods described above. For each patient to be monitored, the covariance function could be selected by assigning the vital sign record to the most similar group.

Gaussian processes may also be used to develop a personalised early warning system, by using the posterior distribution to assess whether new measurements are abnormal. For instance, one simple way to define abnormality would be to assess whether the new measurements are more than three standard deviations away from the posterior mean, over a given length of time. We can imagine how this may work in practice by considering

Figure 6.5.2(4). In this example, the heart rate starts at 80 bpm and jumps to 100 bpm after 20 minutes. Although the heart rate is slightly elevated, a single-channel measurement of 100 bpm would be unlikely to cause clinical concern by itself, based on the population statistics. However, the step-change in heart rate indicates an unexpected change in the patient state that may be of clinical interest. The corresponding Gaussian process model was able to determine that the change in heart rate was unexpected, as indicated by the fact that the 100 bpm measurements were well above the dashed line denoting the Gaussian process mean + 2 standard deviations.

A. Updated Track and Trigger Chart

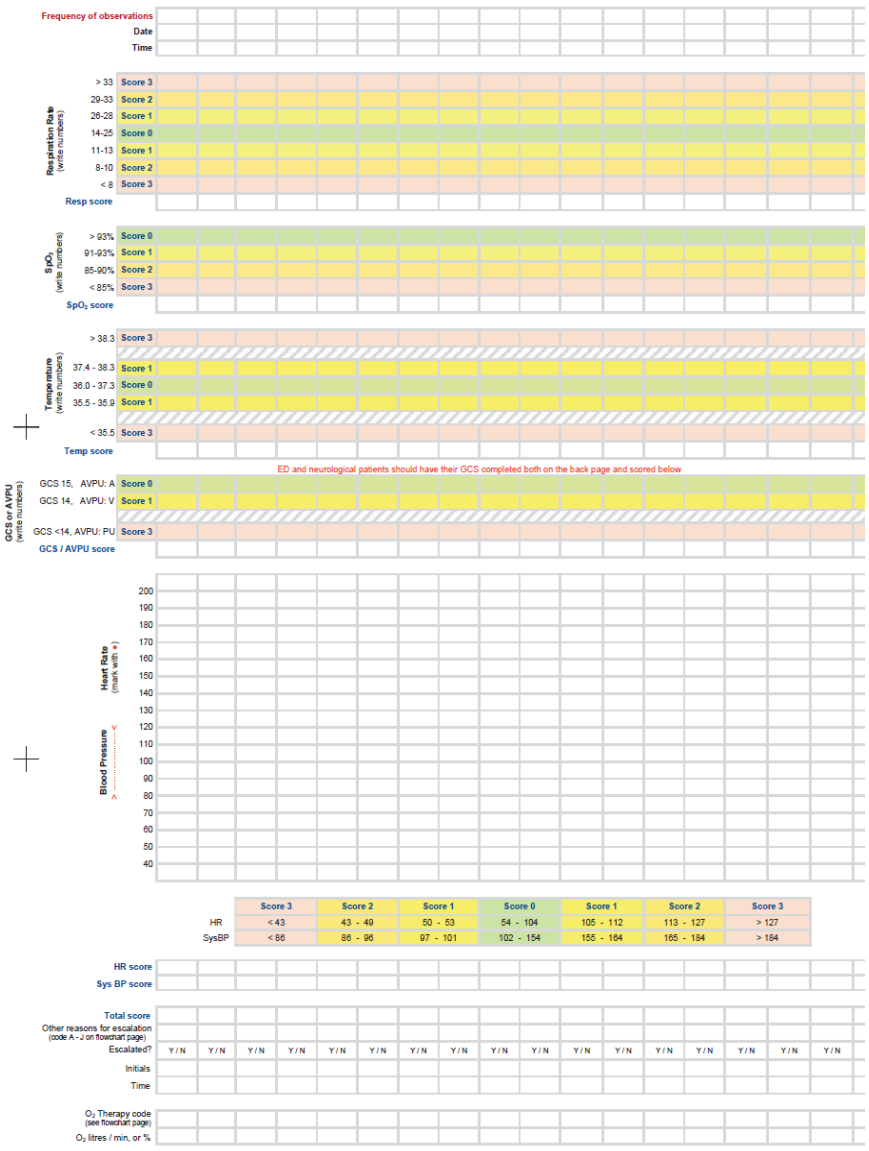

Figure A.0.1.: T&T chart used the John Radcliffe Hospital, Oxford from June 2011

B. Visualisation of High-Dimensional Data for Very Large Data Sets

Visualisation of High-Dimensional Data for Very Large Data Sets

David Wong WONG@ROBOTS.OX.AC.UK
Institute of Biomedical Engineering, Headington, Oxford, OX3 7LF, UK

Iain Strachan IAIN.STRACHAN@OXFORD-BIOSIGNALS.COM
Oxford Biosignals Ltd., Brook House, 174 Milton Park, Abingdon, Oxfordshire, OX14 4SE UK

Lionel Tarassenko LIONEL@ROBOTS.OX.AC.UK
Institute of Biomedical Engineering, Headington, Oxford, OX3 7LF, UK

Abstract

This paper proposes a modification on the Sammon map algorithm for data visualisation. The modification, known as the Sparse Approximated Sammon Stress(SASS), allows mappings to be produced for very large data sets of the order of 10^6 points. While the technique may be useful in a variety of applications, the results presented here will demonstrate its usefulness for visualising patient deterioration in vital sign data collected from step-down unit hospital patients. A final result demonstrates an application of the SASS visualisation for drug safety analysis.

1. Background

In the field of patient monitoring in critical care, researchers are often overwhelmed with large quantities of high-dimensional data. The data typically consist of simultaneous readings of vital signs such as breathing rate, blood pressure, temperature and arterial-oxygen saturation. The new generation of automatic patient monitors and hospital IT systems enable data to be collected quickly and efficiently, so it is no longer unusual for researchers to deal with data sets containing millions of data points.

Initial exploration and analysis of such high-dimensional data is a difficult task. Any analytic tools or algorithms must deal with the data in a coherent and intuitive manner in order to provide useful insight, but must also be usable with large volumes of data.

Appearing in *Proceedings of the Workshop on Machine Learning for Health Care Applications, 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

One important aspect of high-dimensional data analysis is visualisation. This involves transforming the original data to a visualisation space with fewer dimensions. Typically, two or three dimensions are chosen so that the results can be plotted for visual inspection. The transformation is chosen in such a way as to maintain key aspects of the data distribution; for example, topology may be preserved between the dimensions.

A variety of visualisation algorithms have been proposed, including Kohonen's (1997) Self Organising Maps (SOMs) and kernel Principal Component Analysis (PCA) (Schoelkopf et al., 1997). SOMs use a neural network to map data onto a 2D grid such that similar data (i.e. data close to each other in the original high-dimensional space) are grouped together on the grid. This provides insight into the spatial relations within the data. In kernel PCA, the appropriate choice of kernel allows the data to firstly be mapped to a higher dimensional space so that a standard PCA in kernel space has the effect of producing a non-linear mapping between the original data space and visualisation space.

One popular alternative to these methods is the Sammon Map algorithm (Sammon, 1969). This produces a mapping which attempts to keep the Euclidean distances between all pairs of data points in the 2-D visualisation space as close as possible to those in the high-dimensional data space. Mathematically, this is equivalent to minimising the so-called Sammon STRESS objective function for N data samples:

$$STRESS = \frac{1}{\sum_{i=1}^{N}\sum_{j>i}^{N} d_{ij}^{*}} \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(d_{ij} - d_{ij}^{*})^2}{d_{ij}^{*}}$$

where the Euclidean distances between patterns i and j in the data space are denoted by d_{ij}^{*}, and the corre-

sponding distances in visualisation space are denoted by d_{ij}. The objective function is minimised by a gradient descent technique that adjusts the position of the points in visualisation space.

Unfortunately, there are two major drawbacks to the method. Firstly, the process of creating a Sammon Map is intractable for large data sets, as the STRESS calculation involves order $O(N^2)$ point comparisons. On a typical desktop PC, a few thousand data vectors is the practical limit. Secondly, the Sammon Map cannot accommodate new data, and must be retrained each time.

A number of authors have attempted to circumvent these problems. For instance, the Neuroscale algorithm developed by Lowe and Tipping (1997) uses a neural network trained on the data to derive an explicit non-linear transformation between data space and visualisation space that allows new points to be visualised using the interpolation properties of the trained neural network. However, this method also suffers from the same drawback of being unsuitable for large data sets, necessitating either a sub-sampling of the data used for training, or pre-clustering to a smaller set of exemplar vectors using a clustering algorithm such as k-means. At present, the authors are unaware of any method described in the literature that creates a true Sammon map for large ($> 10^4$ point) data sets in reasonable time.

2. Method

We propose a novel alternative to the original Sammon Map algorithm which we have named the Sparse Approximated Sammon STRESS(SASS). SASS reduces the problem to one of order $O(N)$ by sub-sampling from the complete set of inter-point distance pairs to approximate the Sammon STRESS. In practice, it has been discovered that many of the inter-point distances can be removed from the STRESS calculation, with little effect on the Sammon Map output. The method used to sub-sample is critical for obtaining an accurate mapping and is discussed further in the following section. Formally, if we define S to be a sparse subset of the index pairs (i, j) for which the Euclidean distance is calculated, then the modified STRESS objective function to minimise is:

$$SASS = \frac{1}{\sum_{i,j \in S} d^*_{ij}} \sum_{i,j \in S} \frac{(d_{ij} - d^*_{ij})^2}{d^*_{ij}}$$

For very large data sets consisting of at least $N = 10^6$ points, a sparse distance matrix with an average of 50 distance comparisons for each point has been tested and shown to work successfully. In this case, only one distance comparison is computed using SASS for every 20,000 comparisons calculated for the original STRESS. By reducing the computational complexity in this way, the initial problem of large data sets is overcome. Furthermore, data storage is reduced by using memory saving techniques for sparse matrices. Further increases in speed are made by using an efficient optimisation algorithm, scaled conjugate gradients, in preference to gradient descent.

2.1. Initialisation of d_{ij} in Visualisation Space

In the preliminary tests, points in the visualisation space, d_{ij}, were initialised with random values, following the precedent set in Sammon's original paper. During these tests, it was clear that as the size of the data set increases and the STRESS calculation increases accordingly, it becomes likely that the STRESS optimisation procedure will get stuck in a local minimum.

SASS can be initialised in a more principled manner by using a two-stage approach. Firstly, SASS is applied to a subset of the data to produce a preliminary mapping. In this pre-mapping, the points in the visualisation space are initialised randomly. The Sammon map generated by this process creates a sparse outline, or a skeleton, of the data and so the second stage of the initialisation is to approximately map the remaining points into visualisation space using the skeleton. In this case, the distance mapping technique introduced by Pekalska et. al. (1999) was used, which creates an explicit linear transformation between the data and visualisation spaces. This provides an approximation to the transformation created by the Sammon mapping, which is generally non-linear. The result of this process is that all vectors in the data set are initialised to the correct region of the visualisation space.

In preliminary tests on a data set with with 10^6 points, a 4800 point skeleton was created to initialise d_{ij}. The SASS algorithm was then run using the new initialisation values for d_{ij}. In general, it was found that the final SASS error was smaller than for random initialisation of the d_{ij} values, and that the optimisation stage converged in fewer iterations.

2.2. Initialisation of Subset S

The SASS method can fail when a subset of the data, by chance, only possesses inter-point comparisons within the subset. A pictorial representation of this problem is presented in Figure 1. It is unsurprising that such an initialisation results in an incorrect visualisation, as the algorithm will treat the subsets as

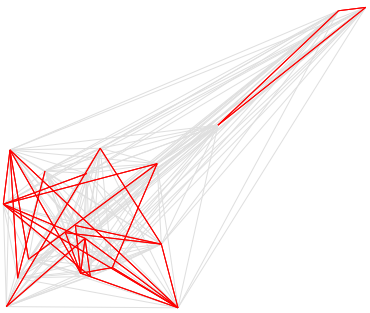

Figure 1. The graph shows an example of the connectivity between data points for the original Sammon algorithm (grey) and for the SASS algorithm (red). Each node represents a data point, and each edge represents an inter-point distance. In this example, the data has formed two unconnected subsets and SASS will fail to produce a correct mapping.

two separate data sets.

Fortunately, the probability of such an event occurring is very small. For instance, the probability of two subsets forming, where one of the subsets contains only one vector (which is equivalent to one point having no connections to any other point in the data set), is given by:

$$P(\textit{one point disconnected}) = N(1 - \frac{2}{N})^{\frac{\lambda}{2}N}$$

where λ is the average number of connections per point such that $\frac{\lambda}{2}N$ is the number of elements in set S, and N is the number of data points in the whole data set, as before. For a data set with over 10^6 points and and average of 50 connections per point, the probability of one point being disconnected is of the order of 10^{-16}. To prevent this problem from occurring at all, we ensure that the connections within the data set form a minimum spanning tree. The simplest way to do this is to initially connect each data vector to its neighbours, so that the n^{th} data vector in the set of N data vectors has distance comparisons to the $n - 1^{th}$ and $n + 1^{th}$ vectors.

SASS can be further enhanced by considering the manner in which the subset S of inter-point connections is chosen. In order to test the effectiveness of alternative choices of S, a unit-cube synthetic data set was created. This consisted of a 3D unit-cube with normally distributed data at each of the corners, so that there were 20×10^4 3D vectors in total. Furthermore, the $(1, 1, 1)$ data vector was added twice to the set as two distinct data points to test whether data are mapped consistently.

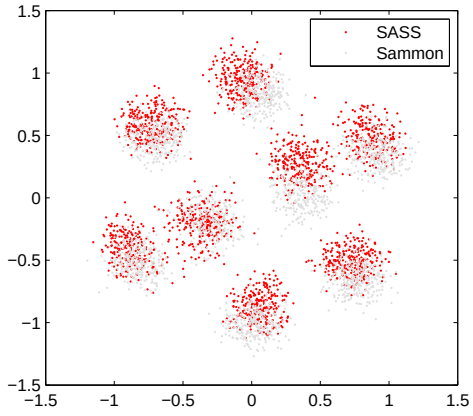

Figure 2. A Sammon Map for the unit-cube data set, containing 2×10^4 points. The SASS Sammon map is shown in red, and the output from the original method is shown in grey. In both instances, the separate data clusters are clearly visualised

In the initial tests, elements in S were chosen by selecting two data vectors at random. Figure 2 shows the result from SASS on the cube data set in red compared to results created directly from Sammon's algorithm in grey. The eight clusters corresponding to the corners of the cube are correctly mapped, and it is clear that SASS works satisfactorily. Although the results are acceptable, in order to maintain accurate local and global structure, the proportion of local and distant inter-point connections is of critical importance.

One natural way to do this is to force each data point to have an equal number of connections to both near and far points in the data set. Local and distant points can be defined for any data set as follows. Firstly, the data set is clustered using a technique such as K-means. Once the points have been grouped, half of the total inter-point connections that form set S are selected such that the two connected points are within the same cluster. These are defined as 'local' connections. The remaining inter-point connections are chosen so that any two connected points are from different clusters. Alternatively, for time series data where vari-

ation is slow compared to the data collection rate, one would expect consecutive samples to appear locally in visualisation space. Therefore, local connections can also be defined by appropriately partitioning a time series data set.

The unit-cube data set was retested using this method to define local and distant connections. Again, Figure 3 shows that the global structure was adequately captured. The duplicate points are highlighted in red, and visual inspection shows that they were mapped consistently. To quantify the accuracy of the mapping, the dataset was visualised 200 times for both a randomly initialised set S, and for the alternative method described above. In each of the 200 Sammon maps, the Euclidean distance between the mapped duplicate points was recorded, and the mean of these was calculated. For the randomly initialised set, the mean distance was 0.05, while the mean distance in the alternative method was 0.02. This indicates that it is important to ensure a sufficiently high proportion of local connections, and that selecting S at random is sub-optimal.

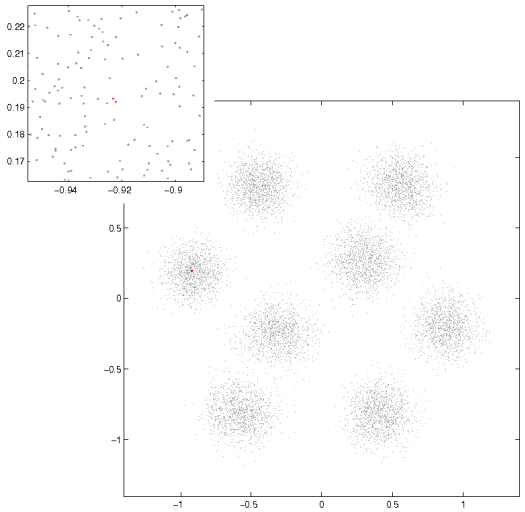

Figure 3. A Sammon Map for the unit-cube data set, containing 2×10^4 points. The SASS Sammon map is shown in grey. In the left-most cluster, the visualisation of the duplicate points at [1,1,1] are highlighted in red. The subfigure shows the left-most cluster in greater detail so that the duplicate points can be distinguished easily.

3. Results

We have used the SASS method as a tool for initially exploring extremely large data sets. Results so far have been encouraging, and have provided insight into ways of improving data fusion models for patient monitoring. The data set used to generate Figures 4 and 5 is taken from a clinical trial on a hospital step-down unit at the University of Pittsburgh Medical Centre(UPMC), and contains vital sign recordings taken over an eight week period for a total of 300 patients (Hravnak et al., 2008b). For each patient, four vital signs, the heart rate, breathing rate, arterial-oxygen saturation and blood pressure, were recorded simultaneously in a 4D data vector. In total, 961,031 vital sign vectors were recorded which corresponds to 28,782 hours of data collection.

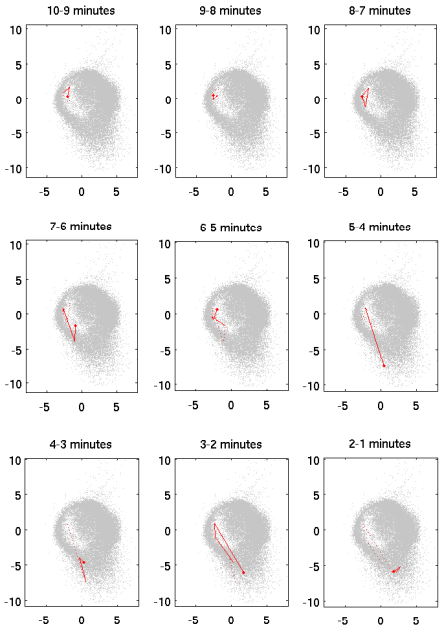

Figure 4. A time-lapse Sammon map showing the deterioration in health of patient C during the last 10 minutes of the patient's vital sign record. The points in light grey depict the vital sign distribution from the entire data set, while the points in white show the vital signs for the patient's entire stay on the ward. The lines in red mark the progression of the patient's vital signs over a one minute period.

One application of SASS allows one to see deterioration in patient health as time progresses. For the UPMC data, a time-lapse SASS map of patient C was created to depict the final ten minutes of the patient's record (Figure 4). The vital sign record for patient C is coloured in white for reference, and the entire record of vital signs recorded during the trial are plotted in light grey. Each point in the figure is a 2D representation of a 4D vital sign vector, and the vital signs recorded over one minute intervals are highlighted in red. The maps clearly show how the patient begins with relatively normal readings, which lie towards the centre-left of the population's distribution. As time progresses, the patient's vital signs become increasingly erratic as the blood-oxygen saturation readings become dangerously low. The bottom row of plots correspond to the last three minutes of the patient record where it can be seen that a number of abnormal vital signs are recorded, denoted by the points towards the edge of the grey (whole population) vital sign cluster and far away from the white (single patient) cluster, and it can be seen that there is a general trend away from normality. The fact that deterioration in patient health can be detected so clearly suggests that it is possible to use trends in time to improve patient monitoring devices.

Another SASS example is given in Figure 5. This Sammon Map depicts the vital signs for patient A and patient B from the same study in red and blue respectively. It is noticeable that the vital signs for each patient are confined to small regions of the whole distribution, indicating that there is considerable patient-to-patient variation within the bounds of vital sign normality. This is not an entirely unexpected result, as external factors such as patient age, physical fitness and reason for admission will have an effect on vital signs. However, given that in the Figure the patients' recordings do not overlap, the Sammon map provides important qualitative evidence that vital sign variation is significant enough to motivate the design of personalised data fusion models for vital sign monitoring.

A final result is presented in Figure 6, and shows the application of the SASS visualisation technique to an application in safety analysis of new drug compounds. This requires 12-lead electrocardiograms (ECGs) to be recorded from human volunteers, from which the effect of the drug on the timing of intra-beat intervals of waveform morphology are assessed. Each point on the plot represents the visualisation of the wavelet coefficients from single-beat ECG waveforms (Strachan et al., 2008; Hravnak et al., 2008a), sampled from the first eight hours of a 24 hour recording during a clinical study of the drug D-sotalol (Sarapa et al., 2004).The

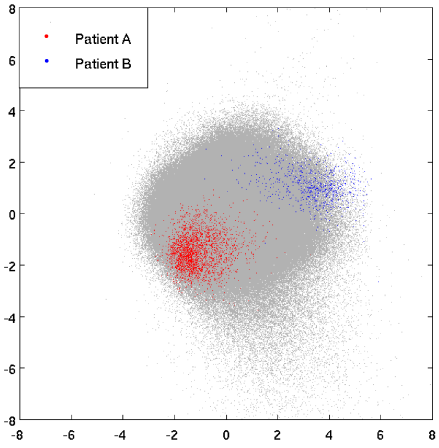

Figure 5. A Sammon Map for UPMC vital sign data. The whole data set, consisting of 961,031 4D vectors is visualised in grey. Points corresponding to vital signs for Patient A and patient B are plotted in red(left) and blue(right) respectively

blue points represent the 'baseline day' where no drug was administered, and the red points represent the drug dosage day for the same subject.

The SASS visualisation was constructed from a set of 8867 beats, roughly half of which were from each day. The distance calculation for the effective Euclidean distance between two beats is more time consuming for this application because the heart rate varies, so the beats are of different lengths. Hence, before a distance calculation can be made, the beats are stretched using Dynamic Time Warping, so they lie on a common axis that minimises the Euclidean distance of the two time sequences.

The visualisation clearly shows a big effect from the drug. It is known that D-sotalol produces large changes in the morphology of the ECG wave, particular in the region of Ventricular repolarisation (T-wave). This would give rise to large differences in the Dynamic Time Warping distance measure. As can be seen, the blue (baseline) cluster is relatively compact, whereas the red (drug) points show two distinct clus-

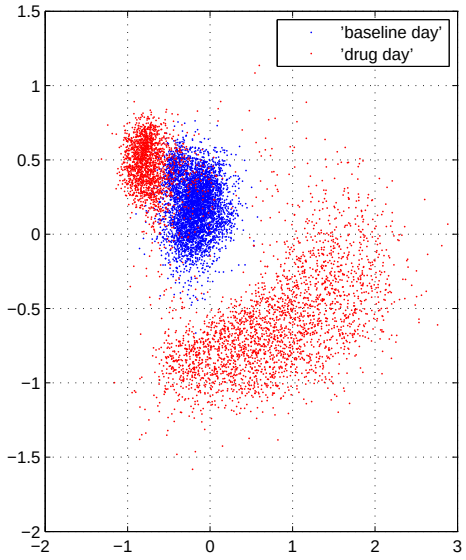

Figure 6. A Sammon map showing the effect of the drug D-sotalol on ECG waveform morphology. The 'baseline' day when no drug was taken is mapped in blue, and the red points represent readings recorded following the administration of the drug.

ters; one which is similarly compact to the baseline, and the other which is widely spread out, indicating a morphology change effect some time after administration of the dose has taken place.

The two more compact clusters are slightly displaced from each other. This is to be expected, as placements of the ECG leads can vary slightly from day to day, and this would be reflected in a small change to morphology.

4. Conclusions and Future Work

The SASS visualisation technique successfully deals with the problem of large data sets. Results in the preceding section show that sets with up to 10^6 points can be accommodated on a standard desktop PC, compared to around 10^4 points that can be mapped using standard Sammon mapping.

Visualisation of the unit-cube data set also confirms that for a medium sized data set, the SASS metric appears to be as accurate as the standard Sammon Map. The similarity between the plots in Figure 2 is encouraging, and one would expect the SASS mapping to also be accurate for larger data sets. This is not directly testable due to the Sammon algorithm limitations discussed previously.

While SASS overcomes the issue of large datasets, it continues to possess some of the other drawbacks of Sammon Maps. In particular, incorporating new data remains a problem. This is an area of current research, and we are investigating the effectiveness of methods in the literature including triangulation (Lee et al., 1977) and the distance mapping technique used previously (Pekalska et al., 1999). One promising idea is a modification to distance mapping which only assumes that local regions in data space can be accurately mapped by a linear transformation. In this way, each new data to point to be incorporated can be mapped according to its own unique, local distance map.

In the patient monitoring context, the results using the SASS technique have been especially useful for facilitating the design of 'smart' patient monitors by allowing us to compare a single patient's vital sign data to vital signs from a whole population (in a trial). Previously, such a large visualisation was computationally infeasible. Two examples have been presented. Firstly, Figure 4 showed that in some cases, deterioration of patient health through time can be clearly seen with respect to the vital sign readings of the trial population. This confirms that effective monitoring, such as the methods developed by Tarassenko et. al.(2006), can be used to provide early warning for certain adverse events and motivates the use of temporal information to improve the monitoring scheme. The second result (Figure 5), highlights the fact that, under certain circumstances, patient-specific models of vital sign data may be more appropriate than a global model of normality.

Acknowledgements

The authors gratefully acknowledge the support of the EPSRC. They are also very thankful to Oxford Biosignals Ltd. for allowing the use of the anonymised vital sign and ECG data during testing.

References

Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M. A., & Pinsky, M. R. (2008a). Defining the incidence of cardirespiratory instability in patients in step-down units using an electronic integrated monitoring system. *Arch Intern Med.*, *168(12)*, 1300–1308.

Hravnak, M., Edwards, L., Clontz, A., Valenta, C., DeVita, M. A., & Pinsky, M. R. (2008b). Impact of an electronic monitoring system upon the incidence and duration of patient instability on a step down unit. *Proc. 4th International Symposium on Rapid Response Systems and Medical Emergency Teams.*

Kohonen, T. (1997). *Self organising maps.* Springer, Berlin.

Lee, R. C. T., Slagle, J. R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from n-space to two-space. *IEEE trans. on computers.*

Lowe, D., & Tipping, M. (1997). Neuroscale: Novel topographic feature extraction with radial basis function networks. *Advances in Neural Information Processing Systems 9* (pp. 543–549).

Pekalska, E., de Ridder, D., Duin, R. P. W., & Kraaijveld, M. A. (1999). A new method of generalizing sammon mapping with application to algorithm speed-up. *ASCI'99 Proc. 5th Annual Conference of the Advanced School for Computing and Image* (pp. 221–228).

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, *18*, 401.

Sarapa, N., Morganroth, J., Couderc, J., Francom, S. F., Darpo, B., Fleishaker, J., McEnroe, J. D., Chen, W. T., Zareba, W., & Moss, A. J. (2004). Electrocardiographic identification of drug-induced qt prolongation: Assessment by different recording and measurement methods. *Annals of noninvasive electrocardiology*, *9*, 48–57.

Schoelkopf, B., Smola, A. J., & Mueller, K. R. (1997). A nonlinear mapping for data structure analysis. *Lecture notes in computer science*, *1327*, 583–588.

Strachan, I. G. D., Hughes, N. P., Poonawala, M., Mason, J. W., & Tarassenko, L. (2008). Automated qt analysis that learns from cardiologist annotations. *Annals of Noninvasive Electrocardiology (to be published).*

Tarassenko, L., Hann, A., & Young, D. (2006). Integrated monitoring and analysis for early warning of patient deterioration. *Br. J. of Anaesth*, *97*, 64–68.

C. Derivation of the Conditional Gaussian Distribution

We seek the conditional distribution $P(b|a)$ when we know that the relevant joint distribution is:

$$\begin{bmatrix} a \\ b \end{bmatrix} \sim \mathcal{N}(0, K)$$

The covariance K can be split up into the form:

$$\begin{bmatrix} A & C^T \\ C & B \end{bmatrix}$$

so that A contains the covariance for the unknown data, a, B is the covariance for the known data, b, and C contains the cross-terms. Thus, the form of the joint distribution is:

$$p(y, y_*) \propto exp\left(-\frac{1}{2}\begin{bmatrix} a \\ b \end{bmatrix}^T \begin{bmatrix} A & C^T \\ C & B \end{bmatrix}^{-1} \begin{bmatrix} a \\ b \end{bmatrix}\right)$$

The inverse covariance matrix can be decomposed into the Schur complement, which may be derived using Gaussian elimination to provide:

$$\begin{bmatrix} A & C^T \\ C & B \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -B^{-1}C & I \end{bmatrix} \begin{bmatrix} (A - C^T B^{-1} C)^{-1} & 0 \\ 0 & B^{-1} \end{bmatrix}^{-1} \begin{bmatrix} I & -C^T B^{-1} \\ 0 & I \end{bmatrix}^{-1}$$

by substituting this into the general form of a multivariate Gaussian and expanding the exponential term by term, the joint distribution can be described as the product of two Gaussians:

$$p(a,b) \propto exp(-\frac{1}{2}(a - C^T B^{-1} b)^T (A - C^T B^{-1} C)^{-1} (a - C^T B^{-1} b)) exp(-\frac{1}{2} b^T B^{-1} b)$$

Note that if there is no cross-variance (i.e. if C is set to zero), then the joint distribution becomes the product of two independent Gaussians. To determine the distribution conditioned on b, we assume that b is a known constant. In this case, the second exponential term is also just a constant. Therefore, the conditional distribution has the form:

$$p(a|b) \sim N(C^T B^{-1} b,\ A - C^T B^{-1} C)$$

Bibliography

[1] *A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand*, 1951.

[2] D.P. Andrulis, A. Kellermann, E.A. Hintz, B.B. Hackman, and V.B. Weslowski. Emergency departments and crowding in United States teaching hospitals. *Annals of Emergency Medicine*, 20(9):980–986, 1991.

[3] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2002.

[4] B.M. Asl, S.K. Setarehdan, and M. Mohebbi. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artificial Intelligence in Medicine*, 44(1):51–64, 2008.

[5] GA Babich and OI Camps. Weighted Parzen windows for pattern classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(5):567–570, 2002.

[6] C. Ball, M. Kirkby, and S. Williams. Effect of the critical care outreach team on patient survival to discharge from hospital and readmission to critical care: non-randomised population based study. *British Medical Journal*, 327(7422):1014, 2003.

[7] D.P. Becker, J.D. Miller, J.D. Ward, R.P. Greenberg, H.F. Young, and R. Sakalas. The outcome from severe head injury with early diagnosis and intensive management. *Journal of Neurosurgery: Pediatrics*, 112, 2010.

[8] M.B. Bell, D. Konrad, F. Granath, A. Ekbom, and C.R. Martling. Prevalence and sensitivity of MET-criteria in a Scandinavian University Hospital. *Resuscitation*, 70(1):66–73, 2006.

[9] R. Bellomo, D. Goldsmith, S. Uchino, J. Buckmaster, G.K. Hart, H. Opdam, W. Silvester, L. Doolan, and G. Gutteridge. A prospective before-and-after trial of a medical emergency team. *Medical Journal of Australia*, 179(6):283–288, 2003.

[10] D. Bennett and J. Bion. ABC of intensive care: organisation of intensive care. *British Medical Journal*, 318(7196):1468, 1999.

[11] C.M. Bishop. Novelty detection and neural network validation. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 141, pages 217–222. IET, 2002.

[12] C.M. Bishop and SpringerLink (Online service). *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[13] P. Boyle. *Gaussian processes for regression and optimisation*. PhD thesis, University of Auckland, 2007.

[14] P. Boyle and M. Frean. Dependent gaussian processes. In *Advances in neural information processing systems 17: proceedings of the 2004 conference*, volume 17, page 217. The MIT Press, 2005.

[15] M. Buist, J. Harrison, E. Abaloz, and S.V. Dyke. Six year audit of cardiac arrests and medical emergency team calls in an Australian outer metropolitan teaching hospital. *British Medical Journal*, 335(7631):1210, 2007.

[16] M.D. Buist, G.E. Moore, S.A. Bernard, B.P. Waxman, J.N. Anderson, and T.V. Nguyen. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *British Medical Journal*, 324(7334):387, 2002.

[17] V.C. Burch, G. Tarr, and C. Morroni. Modified early warning score predicts the need for hospital admission and inhospital mortality. *British Medical Journal*, 25(10):674, 2008.

[18] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines. 2001.

[19] D.W. Chang. *Respiratory care calculations*. Cengage Learning, 1999.

[20] S. Charbonnier and S. Gentil. A trend-based alarm system to improve patient monitoring in intensive care units. *Control Engineering Practice*, 15(9):1039–1050, 2007.

[21] D.A. Clifton, S. Hugueny, and L. Tarassenko. Novelty Detection with Multivariate Extreme Value Statistics. *Journal of Signal Processing Systems*, pages 1–19.

[22] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[23] M. Cretikos, J. Chen, K. Hillman, R. Bellomo, S. Finfer, and A. Flabouris. The objective medical emergency team activation criteria: a case-control study. *Resuscitation*, 73(1):62–72, 2007.

[24] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000.

[25] B.H. Cuthbertson and G.B. Smith. A warning on early-warning scores! *British Journal of Anaesthesia*, 98(6):704, 2007.

[26] Wong D. Optimisation of a multi-parameter monitor for early warning of patient deterioration. Technical report, University of Oxford, UK, 2007.

[27] R.W. Duckitt, R. Buxton-Thomas, J. Walker, E. Cheek, V. Bewick, R. Venn, and LG Forni. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study. *British Journal of Anaesthesia*, 98(6):769, 2007.

[28] R.O. Duda and P.E. Hart. Pattern classification and scene analysis. *A Wiley-Interscience Publication, New York: Wiley, 1973*, 1, 1973.

[29] M. Edwards, H. McKay, C. Van Leuvan, and I. Mitchell. Modified Early Warning Scores: inaccurate summation or inaccurate assignment of score? *Critical Care*, 14(Suppl 1):P257, 2010.

[30] Kluppelberg C Mikosch T. Embrechts, P. *Modelling Extremal Events for Insurance and Finance*, volume 33. Springer Verlag, 1997.

[31] Centre for Clinical Practice. NICE clinical guideline 50: Acutely ill patients in hospital. Technical report, NICE, 2007.

[32] NHS Institute for Innovation and Improvement. Length of Stay: Reducing Length of Stay, 2008.

[33] J. Fordyce, F.S.J. Blank, P. Pekow, H.A. Smithline, G. Ritter, S. Gehlbach, E. Benjamin, and P.L. Henneman. Errors in a busy emergency department. *Annals of Emergency Medicine*, 42(3):324–333, 2003.

[34] H. Gao, A. McDonnell, D.A. Harrison, T. Moore, S. Adam, K. Daly, L. Esmonde, D.R. Goldhill, G.J. Parry, A. Rashidian, et al. 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.

[35] G. Garcea, S. Thomasset, L. McClelland, A. Leslie, and DP Berry. Impact of a critical care outreach team on critical care readmissions and mortality. *Acta Anaesthesiologica Scandinavica*, 48(9):1096–1100, 2004.

[36] A.B. Gardner, A.M. Krieger, G. Vachtsevanos, and B. Litt. One-class novelty detection for seizure analysis from intracranial EEG. *The Journal of Machine Learning Research*, 7:1025–1044, 2006.

[37] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling. The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. *Annals of The Royal College of Surgeons of England*, 88(6):571, 2006.

[38] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C-K Peng, and Stanley H.E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), 2000.

[39] DR Goldhill. The critically ill: following your MEWS. *Quarterly Journal of Medicine*, 94(10):507, 2001.

[40] F. Guiza, J. Ramon, and H. Blockeel. Gaussian processes for prediction in intensive care. In *Proceedings of the Gaussian Processes in Practice Workshop, Bletchley Park, UK*, 2006.

[41] A. Hann. *Multi-parameter Monitoring for Early Warning of Patient Deterioration*. PhD thesis, University of Oxford, 2008.

[42] P. Harber. Report on patient satisfaction survey. Technical report, Worthing Hospital, 2006.

[43] S. Haykin. *Communication Systems*. Wiley-India, 2008.

[44] P. Hayton, B. Scholkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. *Advances in Neural Information Processing Systems*, pages 946–952, 2001.

[45] K. Hillman, J. Chen, M. Cretikos, R. Bellomo, D. Brown, G. Doig, S. Finfer, and A. Flabouris. MERIT study investigators (2005) Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet*, 365:2091–2097.

[46] K. Hillman, M. Parr, A. Flabouris, G. Bishop, and A. Stewart. Redefining in-hospital resuscitation: the concept of the medical emergency team. *Resuscitation*, 48(2):105–110, 2001.

[47] SW Hoare and PCW Beatty. Automatic artifact identification in anaesthesia patient record keeping: a comparison of techniques. *Medical Engineering & Physics*, 22(8):547–553, 2000.

[48] F. Hourihan, G. Bishop, KM Hillman, and K. Daffurn. The medical emergency team: a new strategy to identify and intervene in high-risk patients. *Clinical Intensive Care*, 6(6):269–272, 1995.

[49] M. Hravnak, L. Edwards, A. Clontz, C. Valenta, M.A. DeVita, and M.R. Pinsky. Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system. *Archives of Internal Medicine*, 168(12):1300, 2008.

[50] M. Hravnak, L. Edwards, M. Foster-Heasley, A. Clontz, C. Valenta, M. DeVita, and M. Pinsky. Abstract 75: Electronic Integrated Monitoring of Medical Emergency Team Calls to a Step Down Unit. *Circulation*, 116(16 Supplement), 2007.

[51] D.T. Huang. Clinical review: Impact of emergency department care on intensive care unit costs. *Critical Care*, 8(6):498, 2004.

[52] P.J. Huber. Robust statistics: a review. *Annals of Mathematical Statistics*, 43(3):1041–1067, 1972.

[53] T. Jacques, G.A. Harrison, M.L. McLaws, and G. Kilborn. Signs of critical conditions and emergency responses (SOCCER): A model for predicting adverse events in the inpatient setting. *Resuscitation*, 69(2):175–183, 2006.

[54] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.

[55] D. Jones, R. Bellomo, S. Bates, S. Warrillow, D. Goldsmith, G. Hart, H. Opdam, and G. Gutteridge. Long term effect of a medical emergency team on cardiac arrests in a teaching hospital. *Critical Care*, 9(6):R808–R815, 2005.

[56] Woods J.R., Pohlman T, and J Jackson. Predictive value of an automated neural-network-based early warning system to track the physiological status of hospitalized patients. Technical report, Clarian Health Partners, 2009.

[57] J. Kause, G. Smith, D. Prytherch, M. Parr, A. Flabouris, K. Hillman, et al. A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom–the ACADEMIA study. *Resuscitation*, 62(3):275–282, 2004.

[58] B.R. Kirkwood and J.A.C. Sterne. *Essential Medical Statistics*. Wiley-Blackwell, 2003.

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

[60] B. Kozier. *Fundamentals of nursing: concepts, process and practice*. Prentice Hall, 2008.

[61] T.S. Lam, P.S.K. Mak, W.S. Siu, M.Y. Lam, T.F. Cheung, and T.H. Rainer. Validation of a Modified Early Warning Score (MEWS) in emergency department observation ward patients. *Hong Kong Journal of Emergency Medicine*, 13(1):24–30, 2006.

[62] S.L. Lauritzen. Time series analysis in 1880: a discussion of contributions made by T.N. thiele. *International Statistical Review*, pages 319–331, 1981.

[63] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

[64] M. Lawson, A. Stone, D. King, and A. Davison. The use of a track and trigger system on general medical wards. *Critical Care*, 11(Suppl 2):P443, 2007.

[65] J.R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270(24):2957, 1993.

[66] T. Leary and S. Ridley. Impact of an Outreach team on re-admissions to a critical care unit. *Anaesthesia*, 58(4):328–332, 2003.

[67] D.J. Leith, M. Heidl, and J.V. Ringwood. Gaussian process prior models for electrical load forecasting. In *Probabilistic Methods Applied to Power Systems, 2004 International Conference on*, pages 112–117. IEEE, 2004.

[68] Q. Li, R.G. Mark, and G.D. Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiological measurement*, 29:15, 2008.

[69] L.M. Manevitz and M. Yousef. One-class svms for document classification. *The Journal of Machine Learning Research*, 2:154, 2002.

[70] G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246, 1963.

[71] J. McGaughey, F. Alderdice, R. Fowler, A. Kapila, A. Mayhew, and M. Moutray. Outreach and early warning systems (ews) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev*, 3, 2007.

[72] F.L. Meleney. Hemclytic streptococcus gangrene: importance of early diagnosis and early operation. *JAMA*, 92(24):2009, 1929.

[73] C.E. Metz. Basic principles of roc analysis+. In *Seminars in Nuclear Medicine*, volume 8, pages 283–298. Elsevier, 1978.

[74] G.B. Moody and Mark R.G. A database to support development and evaluation of intelligent care monitoring. *Computers in Cardiology*, 23:657–660, 1996.

[75] R.J.M. Morgan, F. Williams, and MM Wright. An early warning scoring system for detecting developing critical illness. *Clinical Intensive Care*, 8(2):100, 1997.

[76] I. Nabney and C. Bishop. Netlab neural network software. *URL: http://www.ncrg.aston.ac.uk/netlab/index.php*, 1996.

[77] R.M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical report, Department of Statistics, University of Toronto, 1997.

[78] C. Oberli, J. Urzua, C. Saez, M. Guarini, A. Cipriano, B. Garayar, G. Lema, R. Canessa, C. Sacco, and M. Irarrazaval. An expert system for monitor alarm integration. *Journal of Clinical Monitoring and Computing*, 15(1):29–35, 1999.

[79] P.J. Offner, J. Heit, and R. Roberts. Implementation of a rapid response team decreases cardiac arrest outside of the intensive care unit. *The Journal of Trauma*, 62(5):1223, 2007.

[80] M. Oliver. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4(3):313–332, 1990.

[81] E. Osuna, R. Freund, and F. Girosit. Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 130–136. IEEE, 2002.

[82] A.F. Pacela. Impedance pneumography: a survey of instrumentation techniques. *Medical and Biological Engineering and Computing*, 4(1):1–15, 1966.

[83] E.S. Patterson and R.L. Wears. Patient handoffs: Standardized and reliable measurement tools remain elusive. *Joint Commission Journal on Quality and Patient Safety*, 36(2):52–61, 2010.

[84] J. Pennington, J. Laurenson, C. Lebus, S. Sihota, and P. Smith. Evaluation of Early Warning Systems on a medical admissions unit. *Journal of the Intensive Care Society*, 6(2):19, 2005.

[85] D. Perloff, C. Grim, J. Flack, ED Frohlich, M. Hill, M. McDonald, and BZ Morgenstern. Human blood pressure determination by sphygmomanometry. *Circulation*, 88(5):2460, 1993.

[86] A.J. Pittard. Out of our reach? Assessing the impact of introducing a critical care outreach service. *Anaesthesia*, 58(9):882–885, 2003.

[87] J.C. Platt, C.J.C. Burges, S. Swenson, C. Weare, and A. Zheng. Learning a gaussian process prior for automatically generating music playlists. *Advances in Neural Information Processing Systems*, 2:1425–1432, 2002.

[88] G. Priestley, W. Watson, A. Rashidian, C. Mozley, D. Russell, J. Wilson, J. Cope, D. Hart, D. Kay, K. Cowley, et al. Introducing Critical Care Outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Medicine*, 30(7):1398–1404, 2004.

[89] D.R. Prytherch, G.B. Smith, P. Schmidt, P.I. Featherstone, K. Stewart, D. Knight, and B. Higgins. Calculating early warning scores: a classroom comparison of pen and paper and hand-held computer methods. *Resuscitation*, 70(2):173–178, 2006.

[90] D.R. Prytherch, G.B. Smith, P.E. Schmidt, and P.I. Featherstone. ViEWS: Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[91] Priestley M.B. Lessi O. Rao, T.S. *Applications of time series analysis in astronomy and meteorology*. Chapman and Hall, 1997.

[92] C.E. Rasmussen. *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto, 1996.

[93] J.M. Rothschild, E. Gandara, S. Woolf, D.H. Williams, and D.W. Bates. Single-Parameter Early Warning Criteria to Predict Life-Threatening Adverse Events. *Journal of Patient Safety*, 6(2):97, 2010.

[94] R.M. Schein, N. Hazday, M. Pena, B.H. Ruben, and C.L. Sprung. Clinical antecedents to in-hospital cardiopulmonary arrest. *Chest*, 98(6):1388, 1990.

[95] B. Schoelkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[96] R. Schoenberg, DZ Sands, and C. Safran. Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. In *Proceedings of the AMIA Symposium*, page 379. American Medical Informatics Association, 1999.

[97] J.T. Sharpley and J.C. Holden. Introducing an early warning scoring system in a district general hospital. *Nursing in Critical Care*, 9(3):98–103, 2004.

[98] A.F. Smith and R.J. Oakey. Incidence and significance of errors in a patient'track and trigger'system during an epidemic of Legionnaires' disease: retrospective casenote analysis. *Anaesthesia*, 61(3):222–228, 2006.

[99] G.B. Smith, J. Nolan, A. King, P. Pockney, M. Nielsen, M. Coombes, I. Bailey, M. Clancy, M. Buist, G. Moore, et al. Medical emergency teams and cardiac arrests in hospital. *British Medical Journal*, 324(7347):1215, 2002.

[100] G.B. Smith, D.R. Prytherch, P.E. Schmidt, and P.I. Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–179, 2008.

[101] G.B. Smith, D.R. Prytherch, P.E. Schmidt, P.I. Featherstone, and B. Higgins. A review, and performance evaluation, of single-parameter track and trigger systems. *Resuscitation*, 79(1):11–21, 2008.

[102] O. Stegle, S.V. Fallert, DJ MacKay, and S. Brage. Gaussian process robust regression for noisy heart rate data. *Biomedical Engineering, IEEE Transactions on*, 55(9):2143–2151, 2008.

[103] C. Stenhouse, S. Coates, M. Tivey, P. Allsop, and T. Parker. Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward. *British Journal of Anaesthesia*, 84(5):663, 2000.

[104] C.P. Subbe, R.G. Davies, E. Williams, P. Rutherford, and L. Gemmell. 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.

[105] C.P. Subbe, H. Gao, and D.A. Harrison. Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward. *Intensive Care Medicine*, 33(4):619–624, 2007.

[106] C.P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified Early Warning Score in medical admissions. *Quarterly Journal of Medicine*, 94(10):521, 2001.

[107] C.P. Subbe, A. Slater, D. Menon, and L. Gemmell. Validation of physiological scoring systems in the accident and emergency department. *Emergency Medicine Journal*, 23(11):841, 2006.

[108] O. Svenson and A. Edland. Change of preferences under time pressure: choices and judgements. *Scandinavian Journal of Psychology*, 28(4):322–330, 1987.

[109] L. Tarassenko, D.A. Clifton, M.R. Pinsky, M.T. Hravnak, J.R. Woods, and P.J. Watkinson. Centile-based Early Warning Scores derived from statistical distributions of vital signs. *Resuscitation*, 82(8):1013–1018, 2011.

[110] L. Tarassenko, A. Hann, A. Patterson, E. Braithwaite, K. Davidson, V. Barber, and D. Young. Biosign: Multi-parameter monitoring for early warning of patient deterioration. In *Medical Applications of Signal Processing, 2005. The 3rd IEE International Seminar on (Ref. No. 2005-1119)*, pages 71–76. IET, 2005.

[111] L. Tarassenko, A. Hann, and D. Young. Integrated monitoring and analysis for early warning of patient deterioration. *British Journal of Anaesthesia*, 97(1):64, 2006.

[112] S.J. Taylor. *Modelling financial time series*. World Scientific Pub Co Inc, 2008.

[113] G. Teasdale, G. Murray, L. Parker, and B. Jennett. Adding up the glasgow coma score. *Acta neurochirurgica. Supplementum*, 28(1):13, 1979.

[114] M.S. Thaler. *The only EKG book you'll ever need*. Lippincott Williams & Wilkins, 2009.

[115] C.L. Tsien and J.C. Fackler. Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine*, 25(4):614, 1997.

[116] C. Vorwerk. MEWS: predicts hospital admission and mortality in emergency department patients. *Emergency Medicine Journal*, 26(6):466, 2009.

[117] C.K.I. Williams and D. Barber. Bayesian classification with Gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1342–1351, 1998.

[118] C.K.I. Williams, J. Quinn, and N. McIntosh. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. *Advances in Neural Information Processing Systems*, 18:1513, 2006.

[119] C.K.I. Williams and C.E. Rasmussen. Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8*. Citeseer, 1996.

[120] S. Wilson, D Wong, D.A. Clifton, R Pullinger, R. Way, and L. Tarassenko. Track and Trigger in an Emergency Department: an observational evaluation study. *Emergency Medicine Journal*. In Review.

[121] Y. Zhang. Real-time development of patient-specific alarm algorithms for critical care. In *Engineering in Medicine and Biology Society, 2007. 29th Annual International Conference of the IEEE*, pages 4351–4354. IEEE, 2007.