Performance of early warning scoring systems to detect patient deterioration in the Emergency Department

Mauro Santos\textsuperscript{1,2},
Supervised by Dr. David Clifton\textsuperscript{2}, and Prof. Lionel Tarassenko\textsuperscript{2}

Abstract—Acute hospitals in the UK are required to use “track-and-trigger” (T&T) systems in which vital-sign data is collected periodically from patients and then scored according to their abnormality [1]. If the total score exceeds a pre-defined threshold, the care of the patient is escalated. The workload of the Emergency Department (ED), in which clinical observations are taken more frequently (less than 1 hour) than in other wards, affects the completion rate of T&T charts [2]. This has the consequence that patient deterioration may be missed between observations, as it has been show in other studies [3] [4]. In a recent survey to access the use of MEWS in UK EDs 80% of 145 departments were found to be using it, and despite the lack of strong evidence of the effectiveness of these systems in the ED, 90% of the respondents supported its use in the ED [5]. In this paper the performance of 6 early EWS systems in detecting patient deterioration is analysed retrospectively in 347 patients from the an ED trial conducted in the John Radcliffe Hospital in Oxford. Within the multi-parameter systems OR-EWS showed the best performance with 82% and 76% sensitivity and specificity, respectively, while VI-EWS showed the worst, in detecting patients who required escalation at and after arrival in the ED.

Index Terms—Vital-signs Monitoring, Patient Deterioration, Emergency Department, Early Warning Scoring Systems, Quantisation.

I. INTRODUCTION

Failure in recognizing that a patient’s condition is deteriorating in the hospital setting can cause serious adverse events which lead to life-threatening conditions, prolongation of hospitalization and significant disability or incapacity. In response to national guidance [1] many UK hospitals are now using objective methods to identify patient deterioration and ensure timely escalation of care, namely the track and trigger systems (T&T). These systems were introduced in many hospital wards including the high acuity areas of the Emergency Department (ED) to provide a standardized system of observation and continuity of patient care between wards [6]. Events that result in patient deterioration occur with very low frequency, and it is important to identify patients who deteriorate after arrival in the ED, and initiate corrective treatment at the earliest opportunity.

T&T systems use early warning scoring (EWS) systems to track the patients’ health status and although a wide variety of EWS systems have been proposed, there is no clear evidence for their validity, reliability, utility and performance [7], [8]. This paper presents the results of a retrospective analysis of the performance of six EWS systems to detect patients from the ED that require escalation to avoid further physiological deterioration. Continuous monitoring systems are also present in some areas of the ED and can identify, during the intervals between the intermittent nurse observations, patients who will require escalation of care in the ED. This data is also used in this study to quantify the length of time of abnormality each one of these systems is scoring while the patient is connected to the sensors of the bedside monitors. One of the systems is a fine-grained version of a recently proposed evidence based EWS system, QEWS [9], and it is introduced here to study the effect of EWS systems’ quantisation in the detection of patient deterioration.

II. EWS SYSTEMS

A variety of EWS systems exist with different criteria. In this paper six EWS systems are investigated: one single-parameter early warning scoring system, called Medical Emergency Team calling criteria (MET) [10], and five multi-parameter EWS systems (also called MEWS - Modified EWS): 1) a hospital-wide EWS system developed in the Oxford Radcliffe Hospitals (ORH) NHS Trust [2], named, OR-EWS in this paper, used in the John Radcliffe ED till 2010; 2) QEWS [9]; 3) VI-EWS [11]; 5) a version of QEWS currently used in the hospital setting, named JR-QEWS in this paper [12]; and a 6) fine-grained version of QEWS, named, FG-QEWS in this paper (Fig. 3).

These systems were selected because they represent the different research strategies used to define EWS systems that best recognize clinically relevant patient deterioration in the hospital setting. All systems used the following vital signs to assess the patient’s physiology: heart rate (HR), respiration rate (RR), temperature (Temp), Oxygen Saturation (SpO\textsubscript{2}) and systolic blood pressure (SysBP). The systems also use the Glasgow Coma Scale (GCS) or AVPU\textsuperscript{1} scale as a measure of

\textsuperscript{1}The initials are for Alert, Voice, Pain, Unresponsive
consciousness. MET is used in intensive care units (ICUs) to call a Medical Emergency Team when there are gross changes in a single vital sign or a sudden fall in the level of consciousness. It was implemented with the assumption that the physiological processes underlying catastrophic deterioration, such as cardiopulmonary arrest, are identifiable and treatable 6 to 8 hours before the condition occurs [10]. This system has the advantage of being easy for clinical staff to follow.

The multi-parameter systems attempt to assess a patients condition based on the combination of all the vital signs and level of consciousness measurements to produce an aggregated score. The thresholds defined for OR-EWS are based on the consensus of expert doctors from the ORH Trust Hospitals. Abnormal vital-signs and level of consciousness are scored as 0, 1, 2 or 3 according to their level of abnormality. If one of the individual scores is 3 or if the aggregated score is equal to or greater than 4 a clinical review will be requested for a particular patient according to the T&T system in use in the hospital. QEWS uses the same alerting thresholds but it differs in the way abnormality sub-scores are assigned to the vital-signs (explained in the methods section IV-A). At the time of this paper, a version of QEWS was being used in the John Radcliffe hospital. In this system the scores for low heart rate where selected from a distribution of older patients to fit better the older population that is admitted more frequently in hospital wards. This version of QEWS is named JR-QEWS in this paper. The FG-QEWS is a fine grained version of QEWS, that does not quantise the scores to integer values (explained in section IV-A). ViEWS also uses an evidence-based approach to track patient deterioration but it assumes that the criteria should be defined by the system with best AUROC ² to predict patient mortality at 24 hours.

With the use of the electronic T&T system, it is easy to present a fine-grain aggregated score, where all combinations of vital signs are evaluated, rather than a coarse score. QEWS is the first EWS system with a theoretical framework that enables a fine-grained abnormal score. In this paper the quantisation of QEWS is also studied.

A. Evaluation of EWS systems for the ED

In a recent survey to access the use of MEWS in UK EDs 80% of 145 departments were found to be using it, and in 71%, high EWS results in senior ED review. The article concludes by stating that despite the lack of strong evidence of the effectiveness of these systems in the ED, 90% of the respondents supported the use of EWS in the ED [5]. The evaluation of EWS systems for hospital general wards has been discussed by Hann [14] and specifically for the ED by Wong [12]. Early warning scores can be evaluated as diagnostic tools for patient deterioration 1) retrospectively, by relating a clinical outcome to the EWS score, 2) in Before-and-After studies, when the effectiveness of a newly introduced EWS system is tested on a specific hospital ward, 3) or in Randomised Controlled trials which reliably tests the EWS system by randomly allocating patients in the “control” group and the “intervention” group and comparing the outcome of both groups. This paper is focused in a retrospective analysis and only one other study, conducted by Lam et al. (2006) [15] has analyzed the use of MEWS, to detect deteriorating patients in the ED. In this study it was shown that MEWS scores 4 were associated with increased risk of death, ICU admission and hospital admission.

Due to the difficulty of correlating detection of patient deterioration in the ED with a patient hard outcome, or maybe because the patient time in the ED is short (usually 6 hours or less) when compared with time in other hospital wards, most studies focus on retrospectively evaluating EWS to improve the ED triage system. Burch et al. [16] showed that the proportion of patients admitted and those who died in hospital increased significantly as the MEW score increased. Christensen et al. [17], created a multi-parameter EWS where a score >5 at triage is associated with an increased risk of ICU admission within 48 hours of arrival and death within 48 hours of arrival in the ED. In contrast, a similar study by Subbe et al. [18] showed that the introduction of a physiological scoring system, such as MEWS, would have identified only a small number of additional patients as critically ill and added little to the triage system currently in use.

In this paper the patients’ clinical escalations during their stay in the ED are used has outcome variables to evaluate the ability of the EWS to identify patients deterioration in the ED. The advantage of this framework, which was already used in [12], [14], over the previous one (Lam et al.) is the fact that the escalation is likely to be related with the patient deterioration that occurred in the interval \( t \) before it (this interval is studied in section IV-G).

III. CLINICAL STUDY AND DATASET

This paper uses the dataset from the first ED clinical trial validated and reported by Wong [12]. In this observational study vital-sign data were acquired from 472 adult patients during their stay in the ED of the John Radcliffe (JR) hospital, Oxford, using existing hospital bedside monitors (Phillips Intellivue). These monitors provide measurements of HR, RR, SpO₂ at a sampling interval of approximately 30 seconds, and measurements of Blood Pressure (SysBP and diastolic - DiaBP) whenever the patients blood pressure cuff is inflated (approximately every 15 minutes).

Patients are observed by nursing staff every hour in the Majors area if the vital signs are normal, and at 30 and 15 minute intervals when the T&T score is greater or equal to 2, which happens more often in the "Resus" area. The nurse vital sign observation data (HR, RR, SpO₂, blood pressure and Temp) were collected from each participant during their stay in the ED, with the GCS score being used as the measure of consciousness. T&T scores were calculated by nursing staff during their periodic observations, in keeping with standard procedures using the scoring system in Table TABLE ???. 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.

Patients were admitted to the study on a random basis between January 2009 and January 2010, and patient consent was gained in accordance with approval from the Medical Research Ethics Committee (MREC). The clinical observations, T&T scores, continuous monitoring data, clinical escalations and the notes from the clinical records, of a total of 472 patients were combined by into one database. The clinical observations were transcribed from paper by two different teams of clinicians from the JR hospital research staff. The two databases were then reconciled into a final database, used in the investigation described in this paper. From a total of 472 patients, with an average age of 61, 152 (32%) were discharged home at the end of their stay in the ED, 2 patients died in the hospital, and the remainder were transferred to other wards in the hospital.

IV. METHODS

The methodology to study the performance of different EWS systems in identifying deterioration from vital-sign time-series data from ED patients is explained here. The effect of EWS systems quantisation is also studied and it is explained next how the FQ-QEWS was built.

A. FG-QEWS

QEWS uses the same alerting thresholds of MEWS but if differs in the way abnormality sub-scores are assigned to the vital-signs. It uses the cumulative distribution function computed from 64,622 h of vital sign data acquired from 863 patients from medical and surgical wards and it makes the assumption that scores 1, 2 and 3 correspond to the 90th, 95th and 99th centiles respectively, for HR, RR, SysBP and Temp Fig. 1. For SpO2, which has a one-sided distribution, these scores correspond to the 20th, 10th and 1st centiles. To build a fine-grained version of QEWS, first each value from the centile scale is matched to its correspondent vital sign value in the cumulative distribution function (CDF) of the latter. As in the case of QEWS score 1, 2 and 3 are first matched to the 90th, 95th and 99th centiles. Then for each interval between traditional EWS systems scores [1,2] and [2,3] a refined scale is interpolated linearly from the centile values calculated for the intervals [90,95] and [95,99] for HR, RR, Temp and SysBP and [20,10] and [10,1] for SpO2. The results are shown in Fig. 2 and in Fig. 3.

B. Periods of data collection

To study the performance of the EWS systems in the ED ward, the dataset was limited to the subset of patients with a length of stay equal to the mean length of stay ± one standard deviation (SD). 347 patients with total length of stay in the ED within 6:37 ± 4:58 hours were analyzed. For each patient the periods of data collection of each vital-sign were determined from the continuous data captured by the bed-side monitors. A tolerance of 30 min between two BP measurements (SysBP and DiaBP values) and 40 s between two measurements of HR, RR, Temp and SpO2 were considered, to hold the previous value when a data point was missing. The total length of stay and time of data collection are calculated by the patient’s starting ED clinical area: "Resus", Majors area and CDU.

C. Periods of time of abnormality

The periods of abnormality can only be calculated from the continuous data. The vital-sign values outside possible human physiological ranges [14] were considered as artefacts and therefore removed. A 3-point median filter of size 3 (1.5 min window) was used to filter other artefacts from the HR, RR, Temp and SpO2 time-series. Periods of abnormality were determined for each period of data collection calculated in section IV-A, using the six EWS systems.

D. Frequency of physiological abnormalities

The frequency of abnormalities was computed for each of the 347 patient both for the continuous and clinical observations data, using the six EWS systems. In the continuous data case, an abnormality corresponds to each abnormal period scored by each system. For the clinical observations each observation set is either normal or abnormal using the EWS systems’ criteria.

E. Effect of quantisation in EWS systems

QEWS quantisation is studied by comparing the length of time and number of abnormalities with its fine-grain version, for the continuous data and clinical observations, respectively.

F. Duration and Frequency of physiological abnormality per patient escalation group

The distributions of the number and duration of abnormalities scored by the 6 EWS systems are compared between the most normal and abnormal patients. The most abnormal patients are defined as patients with clinically relevant events, while the most normal patients are the ones with no events. The "event" patients were labelled according with the type of clinical treatment or escalation they required. An escalation represents a point in the time-series of the patient continuous data where they received clinical care due to their abnormal state. These were labeled by expert doctors who reviewed both the patients’ vital-signs time series and clinical notes. The escalations were categorized in type A, B and C. Type A are caused due to physiological abnormalities, B are caused due to neurological abnormalities and C are caused due to clinical concern. In the case where different types of escalations existed for the same patient the A type has priority over the B type and B has priority over C. These groups were further subdivided into the patients who escalated at arrival (A1, B1, C1) and patients that escalated after arrival (A2, B2, C2).

The number of patients per escalation group for the subset of patients defined in section IV-B is shown in TABLE I.

| Number of patients per escalation group for 347 patients dataset. |
|--------------------------|---|---|---|---|---|---|
| Total                    | 51 | 14 | 14 | 2 | 65 | 12 |
| Discharged Home          | 6  | 2  | 2  | 0 | 10 | 1  |

Normal patients: are defined as the ones with no escalations that were discharged home (NED - Non-escalated and
Discharged). From 189 patients with no escalations, 88 (47%) were discharged home.

**Abnormal patients**: patients with escalations due to physiological abnormality escalated at and after arrival. A total of 51 A1 patients, and 14 A2 patients were indentified. The B-type and C-type patients are not included in this analysis. In the case of the B-type patients, the neurological conditions are better discriminated with the GCS, which is not present in the continuous data. The C-type patients are escalated due to clinical concern which may not be related with physiological abnormality, which is what needs to be quantified in this analysis.

**G. Performance of EWS systems in the ED dataset**

ROC analysis is a classic methodology from signal detection theory used to depict the tradeoff between hit rates and false alarm rates of classifiers [19]. ROC graphs have also been commonly used on medical diagnosis for visualizing and analyzing the behavior of diagnostic systems. This paper follows the patient-based analysis formulated in [12]. In this analysis "Normal" patients comprise patients with no clinical escalations, corresponding to 189 (55%) of the 347 patients used in this study. "Event" patients include patients containing at least one event (escalation) due to physiological abnormality during their stay in the ED, including neurological abnormalities (B1 and B2 patients). These correspond to a total of 81 (23%) of 347 patients. For this analysis we removed the patients whose escalations were not due to physiological abnormalities (C1 and C2). A total of 270 patients are used.

True Positives (TP) are defined as the "event" patients for whom the first clinical escalation was successfully detected by an abnormal period scored by the EWS systems. False Negatives (FN) are "event" patients that were not detected because no abnormal periods were generated from an EWS system. In the case of TP, the alert (abnormal period) is considered if it is generated within an interval \( t \) before the first escalation and 10 min after the escalation, because the escalation times are not precise. For this study, the sensitivity of the systems is evaluated when the interval \( t \) is varied from 10 minutes to 1 hour. Only the first escalation is considered to evaluate the performance of EWS systems on identifying deterioration that is independent of treatment that might have occurred between two escalations. A True Negative (TN) is considered to be a normal patient for which there were no alerts generated and finally a False Positive (FP) is a normal patient for which at least one alert was generated. One of the effects of this methodology is the fact that the sensitivity will be the same in all cases since the TN and FP do not depend on time \( t \).

The performance of the EWS systems as a diagnostic tool to detect patient deterioration in "event" patients is analysed for 5 cases: 1) OBS T&T - when only using clinical observations, 2) OBS. T&T + CONT. T&T - when using clinical observations and continuous data, 3) OBS. T&T + CONT. T&T WITH PERSISTENCE - when using clinical observations and continuous data, considering a persistence criterion to generate an alert, that is, an alert for continuous data will only exist if 4 min of abnormality exist in a 5 min window, 4) CONT. T&T WITH PERSISTENCE - when using only continuous data with the persistence criterion, and 5) CONT. T&T - when using only continuous data with no persistence criterion.

This analysis is repeated for the subgroup of "event" patients that was escalated only after arrival (A2 and B2 patients). The number of normal patients remains the same, while the number of escalated patients is now 16, giving a total of 205 patients.

**V. RESULTS**

**A. FG-QEWS**

Fig. 2 and Fig. 3 show the results of QEWs refinement. QEWs quantisation can be immediately observed. The analysis of the effect of quantisation is shown in section V-E.
Fig. 3. T&T scores of Fine-grained vs coarse-grained QEWS.

Fig. 4. Histogram of patient’s time of stay in the ED by starting clinical area.

**B. Periods of data collection**

Fig. 4 presents the histogram of the length of stay and time of data collection of 347 patients from the ED dataset. The length of stay and time of data collection are shown per patients’ starting ED clinical area. For example, more abnormal patients, requiring a higher level of clinical care, may initially first start in the Majors and then move to the "Resus" area. The total length of stay was 1788 hours and the total time of data collection was 1585 hours. The periods of data collection are determined as described in section IV-B. TABLE II shows the total time of data collected for each vital-sign and the mean percentage of time the patients were monitored for each vital sign. The vital signs were collected in average for 60% of the patients length of stay in the ED if the temperature data is not considered.

**C. Periods of physiological abnormality**

TABLE III shows the percentage of collected data that was classified as being abnormal (score 3) for each vital sign and for each EWS system. The RR contributes for the significant differences between the ViEWS and the JR-QEWS criteria, both evidence based approaches. While about 66% of patients show some RR abnormality time in ViEWS, only 12% show abnormal RR time in the JR-QEWS. This corresponds to a total of 115 hours of RR abnormality for ViEWS against a total of 7 hours JR-QEWS. The length of time of abnormality for RR in MET was 5 hours, while the OR-EWS scored 26 hours of physiological abnormality. The respiratory rate is one of the vital signs for which these systems differ most. The FG-QEWS shows the same amount of RR abnormality than the MET, scoring less 2 hours of abnormality than JR-QEWS. There are no differences between the time of abnormality scored by QEWS and FG-QEWS, which is expected since the abnormal thresholds 3, have the same values. This means that quantisation only affects the aggregate score and not the abnormal scores of the individual vital-signs.

**D. Frequency of physiological abnormalities**

TABLE IV shows the total number of abnormality periods for the subset of ED patients from section IV-B when scored by the EWS systems. The OR-EWS and the ViEWS present 2 and 3 times more abnormal periods than the JR-QEWS respectively. In the case of ViEWS, this is related to the very sensitive threshold for the high RR (Table 10). This threshold influences the aggregated score as can be seen by the high number of periods in the "High & Low" column, given that the thresholds for abnormal low values are similar to the OR-EWS ones for most of the vital-signs (SpO\textsubscript{2}, RR and SysBP). The greater number of abnormalities in the FG-QEWS suggests that there are some combinations of vital sign abnormalities that are not being captured by the quantized JR-QEWS. The greatest difference between the JR-QEWS and its fine-grain version is in the number of abnormalities due to low values of HR.

TABLE V shows that OR-EWS, ViEWS and FG-QEWS show more number of abnormalities for Majors than "Resus" patients. MET has a similar number of abnormalities and JR-QEWS presents more abnormalities for "Resus". Since "Resus" patients are more abnormal, confirmed by the duration of abnormality, the frequency of abnormalities, alone, is not a good indicator to rank the patients by abnormality. A more abnormal patient will have longer periods than transient ones and so it is inferred that "Resus" patients will have longer periods of abnormality than Majors patients showing fewer abnormalities scored by the EWS systems.

**E. Effect of quantisation in EWS systems**

Fig. 2 and Fig. 3 show the fine-grain QEWS. For the continuous data, FG-QEWS only removes 2.5 min of abnormality, caused by the quantisation, from the total 1585 hours of data collection, from 347 patients. On the other hand, the results
showed that if an alerting threshold of 4 min of abnormality was used, FG-QEWS would alert more 33 more times than QEWS, affecting 26 patients.

For the clinical observations, FG-QEWS would only lead to one clinical observation from not scoring as abnormal, due to QEWS quantisation, being scored 3.99 instead. This happened for patient ED00572 that was not escalated. However, FG-QEWS would have scored more than 45 abnormal observations than QEWS, affecting 41 patients. 12 of these patients were not escalated, and seven of these were discharged home.

It seems that more combinations of vital signs, that would score abnormal in a fine-grain EWS system, are being missed due to quantisation.

**F. Duration and Frequency of physiological abnormality per patient escalation group**

1) **Continuous Data:** Fig. 5, shows a boxplot of the duration of physiological abnormalities scored by six EWS systems on A1, A2 and NED patients. The total numbers of patients used were 51, 14 and 88 respectively. This boxplot requires two different statistical analyses to study the differences in the duration of physiological abnormalities scored by the different EWS systems: 1) inter-patient group analysis where each EWS system is tested on different groups of patients and 2) intra-patient group analysis where the all EWS systems are compared on the same patient group. For both cases the Kruskawalli test is used to study if the differences between the medians are statistically significant between 3 or more samples, and the Wilcoxon Rank Sum test is used to confirm statistical significance. The analysis is shown for all EWS systems except QEWS, giving more focus to the JR-QEWS (currently used in clinical practise) and FG-QEWS, the fine grain version.

In the inter patient group analysis, it is observed that the median time of abnormality scored by each EWS system was greater for A1 patients (varies from 0.57 to 1.67 hours), followed by A2 (varies from 0.15 to 1.02 hours) and NED patients (varies from 0 to 0.05 hours), MET and ViEWS scoring the lowest and the highest respectively. Overall it is shown that there is a statistically significant difference between the median duration of physiological abnormality, between escalated and NED patients, p-value is <0.01. However the difference in the median duration of abnormality between A1 and A2 patients is not statistically significant for all systems. The p-value for each case was: 0.66 for MET, 0.07 for OR-EWS, 0.16 for JR-QEWS, 0.13 for ViEWS and 0.15 for FG-QEWS.

In the intra patient group analysis evaluates wheter if the EWS systems give different times of abnormality within the same patient group. In the case of A1 patients, the difference in the median duration of physiological abnormality, between escalated and NED patients, p-value is <0.01. However the difference in the median duration of abnormality between A1 and A2 patients is not statistically significant for all systems.
patients only the medians between 1) MET and JR-QEWS, 2) OR-EWS and ViEWS and 3) OR-EWS and FG-QEWS were considered as belonging to the same distribution (p-value ≥0.01). This is an interesting result since it indicates that JR-QEWS is as conservative as the MET criteria when scoring periods of abnormality in NED patients.

The same analysis is now applied to the number of abnormal periods scored by the different EWS systems on continuous data from the same patient groups (Fig. 6).

For the inter patient group analysis, the results follow the same trend as the ones for the duration of abnormality. The median number of periods of abnormality scored by all the EWS is different with significance (p-value <0.01) between NED and escalated (A1 and A2) patients, overall being higher for escalated patients (3 to 9 abnormalities) than NED patients (0 to 3 abnormalities), MET and ViEWS presenting the lowest and the highest, respectively.

The intra patient group analysis shows that for A1 patients only the MET shows a significant different median (p-value <0.03) when compared with all other EWS systems. For A2 patients there is a significant difference between 1) JR-QEWS (3.5) and ViEWS (9) and 2) MET (2.5) and ViEWS (9), with p-values 0.05 and 0.02 respectively. Finally for NED patients the difference in the medians was significant between MET and FG-QEWS, the later showing a higher median, which differs from the continuous data results. This could be due to the fact that the GCS and the Temp are being used to score abnormality in clinical observations and not in the continuous data case. Finally for NED patients all the EWS systems have a median number of abnormalities of 0, but the statistical tests indicate that the medians do not come from the same distribution in the case of 1) JR-QEWS and ViEWS and 2) JR-QEWS and FG-QEWS and 3) ViEWS and FG-QEWS.

G. Performance of EWS systems in the ED dataset

Fig. 8 and Fig. 9 show ROC charts for the performance of the EWS systems when the window to detect an escalation in an event patient is changed from 10 min to 1 hour. The first figure represents the case where all event patients are used, while in the second chart only patients escalated after arrival are used. These two cases are studied next.

All event patients

For this analysis, case 4) CONT. T&T WITH PERSISTENCE and case 5) CONT. T&T, present very low performance due to the fact that for some escalations at arrival (A1), no continuous data was available. The ROC chart of case 4 is shown for comparison. The first 3 cases: 1) OBS T&T; 2) OBS. T&T + CONT. T&T and 3) OBS. T&T + CONT. T&T WITH PERSISTENCE are discussed here. As it can be seen the specificity does not change as a consequence of the approach used to classify FP and TN, as these do not depend on the time window t. The system that performed worst was the MET applied to case 2, and the system that performs best is the OR-EWS applied to the clinical observations (case 1) with a sensitivity and specificity of 82% and 76% respectively. JR-QEWS was the system that performed better for case 2 and
Fig. 9. Performance of EWS system when considering patients escalated after arrival (A2 and B2 patients) for 5 cases: 1) OBS T&T; 2) OBS T&T + CONT T&T; 3) OBS T&T + CONT T&T WITH PERSISTENCE; 4) CONT T&T WITH PERSISTENCE and 5) CONT T&T.

case 3 with a sensitivity of 86% and 83% for case 2 and case 3 respectively. OR-EWS shows the highest sensitivity for case 2 (89%), but it is the MET who shows the highest specificity for case 1 (88%). ViEWS is the system with lowest performance in all cases. The combined T&T approach with and without the persistence criterion detects more TP patients then only the T&T based on clinical observations. When the persistence criterion was changed to 2 min of abnormality in a 3 min window, there was no change in the sensitivity but the specificity decreased, with the increase in FP. Figure 25 and Figure 26 show examples of two patients that were escalated only when the continuous T&T system was present. Both patients were escalated because one of the vital-signs became very abnormal before the escalation (catastrophic deterioration). In the case of patient ED00479 there was a clear hypotension period at 15:30 hours where the SysBP was 65 mmHg, 30 min before the escalation. All systems indicate periods of patient deterioration starting at 12:00 hours, 30 min after the last observation before the escalation. The clinical observations missed these periods of patient deterioration that would trigger a clinical review by an expert doctor at least 2 hours before the escalation, as defined by the T&T rules. This patient was labelled as being escalated due to neurological abnormality (B2 patient). In the case of patient ED00177 all continuous T&T systems alerted 20 min before the escalation at 3:54 AM, due to the tachycardia period before the event. Only ViEWS scored intermittent abnormal respiration periods 3 hours before the escalation. In this case the irregular decreasing heart beat is missed by all EWS systems and the irregular RR is only identified by ViEWS.

Event patients escalated after arrival

In this case performance is evaluated on the group of patients that includes A2, B2 and non-escalated patients, to discuss the cases when there should have been continuous data to capture the escalations. As expected the trends are the same: 1) The system that performs better is OR-EWS when only clinical observations are used, with 82% sensitivity and 76% specificity; 2) JR-QEWS performs better in the case of the combined OBS T&T and CONT T&T with the persistence criterion, with 100% sensitivity, but only 63% specificity, 3) ViEWS is the multi-parameter system that performs worst on all cases except on "CONT T&T WITH PERSISTENCE" where OR-EWS performs worst.

VI. DISCUSSION

A. Use of EWS systems in the ED

The total number and duration of physiological abnormalities were higher for patients that were admitted first to the "Resus" area than for the ones admitted first to the "Majors" are for all EWS systems, both for continuous and clinical observation data. It was also shown that the median values of these parameters is higher for escalated patients than for NED patients, for all EWS systems, with statistical significance. However the same was not true between patients escalated at (A1) and after arrival (A2) in the ED. The number of A2 patients was very low (14) which may have contributed to the lack of statistical significance.

ViEWS showed the highest duration and frequency of abnormality for NED patients. This is another indication that ViEWS will generate more false alerts for patients that are stable, especially due to the RR upper threshold (25 bpm) that causes it to score 5 times more abnormality than any of the other considered EWS systems. On the other hand ViEWS can miss periods of deterioration due to bradycardia and hypertension, which was the case for one A2 patient in this dataset. MET showed the lowest duration of abnormality in A1 and A2 patients, which is an indication that it can miss periods of deterioration caused by combinations of abnormal vital-signs.

B. The need for continuous monitoring in the ED

It was shown in this paper that the combination of the traditional and continuous T&T system improved the sensitivity in detecting "event" patients from 82% to 89% in the
case of OR-EWS, the system with best performance when only using the T&T on clinical observations. However its specificity in this case was very low, 25%. By applying the persistence criterion to the continuous T&T system the specificity was improved to 57%. JR-QEWS was the system with best overall performance for the combined T&T system. When the persistence criterion is considered its sensitivity and specificity are 84% and 63% respectively. It was shown that for two patients the continuous data would capture periods of abnormality due to the combination of abnormal vital-signs, scored by most or some of the EWS systems, between clinical observations and before the patient was escalated, due to clear physiological abnormality before the escalation.

C. The need for improvement in data fusion systems

For Patient ED00479 a data fusion system based on any of the multi-parameter EWS systems would have alerted 3 hours before the escalation and 2.5 hours before the clear period of hypotension. According to the escalation pathway implemented in the JR ED, the patient would have been reviewed by an expert doctor at least 2 hours before the escalation. For patient ED00177 only ViEWS would have alerted 3 hours before the escalation, based on intermittent RR abnormality. In the time-series of this patient there is a clear decreasing heart rate trend, and an increasing heart variability trend before the escalation, both starting at 21:30 PM, 6.5 hours before the escalation. Also the RR data (that started at 0:00 AM) is irregular varying between 15 and 30 breaths/minute. The patient was moved at 23:30 to the CDU area, and RR monitoring was added. These two conditions preceded the tachycardia event that triggered the escalation at 4:00 AM. Atrial fibrillation (AF) with rapid response rate was the diagnosis for this patient and it is the most common cardiac arrhythmia. It may cause no symptoms, but it is often associated with palpitations, fainting, chest pain, or congestive heart failure. Atrial fibrillation may be treated with medications to either slow the heart rate to a normal range (“rate control”) or revert the heart rhythm back to normal (“rhythm control”). A data fusion system, connected to a bedside monitor, should be able to capture these three described trends and evaluate them as abnormal. A calibrated Gaussian Process would be able to capture 1) the decreasing mean heart rate trend, with increasing variance around the mean (the increasing heart rate variability trend), and 2) the high RR variance around the mean RR. This patient was reviewed by a doctor at 2:40 AM, and a medication plan was draw up but not in time to prevent the escalation at 3:20 PM, that was reported in the clinical notes at 3:45 PM, and therefore labelled at that time.

D. Quantisation of EWS systems

In this paper a fine-grain EWS system is extended from the evidence based QEWS system. The FG-QEWS scored an extra 22 hours of physiological abnormality than QEWS. The number of periods of abnormality or clinical observations that would be avoided with the fine-grained version of the system is insignificant. Much more significant are the periods of abnormality that were missed by QEWS because of its quantisation affecting the clinical observation of 41 patients, and 33 periods longer than 4 min affecting at least 26 patients. In the performance analysis, FG-QEWS ranked 3rd, performing better than MET and ViEWS to detect “event” patients. This system scored more 37 hours of abnormality than the quantised JR-QEWS, 10 hours being from abnormal HR. This happens because the system is not corrected by clinical judgment as JR-QEWS was, in order to be applied to clinical practice. According to our performance analysis
Fig. 11. Patient ED00177 (A2 patient) was correctly classified as an "event patient" due to the periods of abnormality scored by the continuous T&T before the escalation at 03:45.

This correction was well applied as it increased JR-QEWS specificity. This study shows that without correction from clinical judgement, to tune the system for populations that differ in their vital-sign distribution, such as young and elderly patients, the fine-grained version of QEWS would decrease its specificity while maintaining a similar sensitivity.

E. Limitations of this study

One of the objectives of the study was to analyse the duration and frequency of abnormalities in clinical observations and continuous data between the identified patient groups. The group of patients that escalated after arrival (A2) was small to consider statistically significant difference in the way that the EWS systems score abnormality within this patient group. In the study dataset there are 29 A2 patients, but it was decided that the EWS systems should be compared for patients whose length of stay was similar, so the length of stay would not correlate strongly with abnormality [20].

As pointed by Wong [12], the outcome measure in the ED study, escalations, was usually triggered by the T&T alerting thresholds being met, and was therefore biased towards T&T scores. It could be argued that OR-EWS performed better because it was the EWS system being used at the time the data was collected. It is also true that there is no gold standard methodology to evaluate the performance of EWS systems, for the purposes of detecting patient deterioration.

VII. Conclusion

NHS standards suggest patients from the ED should be monitored, diagnosed and receive initial treatment within four hours of their arrival in the department. Then they should be discharged to a location appropriate to the diagnoses made. The 347 analysed patients’ average length of stay was $4.5 \pm 2$ hours, and in average they were continuously monitored 60% of their time. The patient-based performance analysis enabled us to test the performance of EWS against the escalation outcome framework where the purpose is to diagnose patients who need to be escalated to avoid further deterioration. It was shown that in this way it is possible to understand what periods of abnormality that result from the combination of abnormal vital-signs, are being captured by the different EWS systems.

As catastrophic deterioration, such as cardiac arrest, is usually preceded by abnormal vital-signs [21], an effective EWS system should be able to detect them. It was shown that the OR-EWS was the system with best performance to detect “event” patients when scoring clinical observations. It was also shown that its sensitivity would be improved by using a combined T&T system also based on continuous data acquired by bedside monitors connected to patients from "Resus", Majors and CDU areas. The specificity on the other hand was low, which leads to a research opportunity to develop datafusion algorithms that are able to discriminate between relevant abnormal periods, which precede deterioration, and transient ones.

It was shown escalated patients had a higher duration and frequency of physiological abnormality than the NED ones when scored by each of the scoring systems. It was also shown that the current EWS systems have a great limitation in detecting deterioration that depends on the combination of vital-signs time series dynamics rather than the combination of their amplitudes (as explicited by patient ED00177). Finally it was shown that although the FG-QEWS performed better than ViEWS the system needs to be tuned for the population that needs to be scored to avoid lowering specificity.
REFERENCEs

Vital-Sign data fusion for detection of patient deterioration in the Emergency Department

Mauro Santos
St Hilda’s College
University of Oxford

Supervised by
Professor Lionel Tarassenko
Submitted: Michaelmas Term,
September 1, 2012

This Transfer Report is submitted to the Department of Engineering Science, University of Oxford
Contents

1 Literature Review ............................................. 1
   1.1 Introduction ............................................. 1
   1.2 Continuous Monitoring Systems for Detecting Patient Deterioration .... 2
   1.3 Novelty Detection ......................................... 7
   1.4 Gaussian Processes for Time-series Modelling ......................... 11
   1.5 Future Work .............................................. 13
      1.5.1 2nd Phase ED trial data collection and validation ............... 13
      1.5.2 Machine Learning ..................................... 14

A DPhil Project Plan ........................................ 16
   Bibliography .................................................. 17
Chapter 1: Literature Review

1.1 Introduction

The Emergency Department (ED) is often the first contact with a hospital for a patient that experiences an acute health problem. The UK Government’s Department of Health has mandated that at least 98% of patients presenting to an ED be seen, treated, and either admitted or discharged within four hours [33]. Since the patient’s condition is usually unknown prior to arrival, the ED must provide initial treatment for a broad spectrum of illnesses and injuries. The ED can be extremely busy when other wards like the Intensive Care Unit (ICU) are close to full capacity and patients may be held in the ED until ICU beds become available [25].

UK National Health Service (NHS) hospitals are required to use “track and trigger” (T&T) systems in which vital-sign observations are collected periodically from patients and scored [38]. If the scores exceed a pre-defined threshold, care of the patient is escalated. In the ED, observations are taken more frequently than on general wards [17]. This additional workload combined with frequent overcrowding of the department results in low levels of correctly completed T&T scores and the possibility of missing patient deterioration between observations.

It is known that patient morbidity is improved if diagnosis leads to early treatment [5] [8], and it has been shown that early treatment in the ED can decrease length-of-stay and level of care required in the ICU [25]. Furthermore, decreased length of stay with a consequentially increased patient throughput may have a substantial financial benefit.
The use of continuous monitoring systems that alert nursing staff by integrating information from multiple vital signs is standard in the ICU. Such an approach could be applied to acute wards, such as the ED, to help continuously screen for conditions that could otherwise be missed between clinical observations.

However, care must be taken when applying methods from the ICU to patients in the ED. The ICU and ED have different conditions and patient types. While ICU beds have fixed beds-side monitors, the ED does not have such monitors outside the Resuscitation Room ("Resus") and Majors area and nurses have to make regular observations of the patients’ vital signs. The most modern T&T systems allow nurses to input vital-sign observations into an electronic hand-held PDA, with a validation step that checks the input. The PDA then automatically calculates the patient score using a calibrated T&T chart [41]. It would be ideal if a patient that is initially observed with such PDA-based systems and who is then transferred to a bed with a bed-side monitor could have their vital-sign history integrated, complementing the infrequent clinical observations with data derived via continuous monitoring. Furthermore, some clinical observations, such as the level of consciousness\(^1\) [55], are not currently captured automatically by continuous monitoring systems, and are important to identify patient deterioration [38, 55].

1.2 Continuous Monitoring Systems for Detecting Patient Deterioration

Data fusion is the process that allows the integration of vital signs, which may be continuously or manually acquired providing a status index that indicates the patient’s health.

Most of the data fusion methods described in the literature have been designed for use on high-dependency units (HDUs), where continuous monitoring is standard. Previous work on automatic early warning of patient deterioration can be separated into two categories: those that generate alarms based on a single vital sign, and those that aggregate

\(^1\)which is manually scored on a semi-objective scale from “alert” down to “unresponsive”
multiple vital signs before generating alerts.

Single-channel systems are those in which alert thresholds are set for each vital sign. The bedside monitor will alert when any one of the vital signs is outside the range defined by these thresholds. The thresholds may be chosen by the manufacturers of the bedside monitor, by nurses, or by an automatic selection scheme based upon an offset from the previous values being recorded. Such systems typically have a very high sensitivity and very low specificity [59]. This results in nursing staff ignoring the alarms or increasing the alarm thresholds to increase specificity while lowering sensitivity to a point where the alerts have little or no clinical value. Additionally, if alarms are temporarily disabled by a clinician, other staff may not be informed, making the system potentially more hazardous than having no alarms at all [2].

Tsien et al. [60] tried to improve the accuracy of single-channel alarms using supervised learning techniques, in which a clinician identified intervals of data that were deemed artefacts in vital signs acquired from neonatal ICU patients, preventing the latter from being used to generate false alarms. These labels were used together with features extracted from a window of data to train decision trees and logistic regression models. These models were used to classify sections of single-channel data as either “artefactual” or not. It was shown that the best-performing models were those derived from more than one signal type. Clifford et al. showed that false alarms in ECG arrhythmia detectors could be minimized using the arterial blood pressure waveform [15].

Although such approaches integrate data from multiple sources, the extra information has only been used to validate a single-channel alarm. More recent systems use a data fusion method in which information from multiple channels is fused together prior to alarm generation. There have been two types of approaches to detecting patient deterioration using data fusion: expert systems, where human expert knowledge is used to construct a rule-based classifier, and machine learning, where detecting abnormality is “learnt” automatically from training data.

Oberli et al. propose an expert system approach to the data fusion problem [12]. The
vital signs are first converted into a set of quantitative classes which describe a physiological condition, such as bradycardia (abnormally low heart rate) or “normal” heart rate, based on training information given by a set of clinicians. The experts then define rules by proposing a diagnosis for every class thus derived. If any of the rules associated with abnormal physiology are satisfied, an alarm is produced, which also provides information describing the cause. This system was tested with 20 adult patients undergoing cardiac surgery for heart-lung bypass, together with standard monitors. The sensitivity of the expert system was 92% and the positive predictive value was 97%, compared with a sensitivity of 79% and a positive predictive value of 31% for standard monitors. The alarms generated by the expert system were deemed to be clinically relevant 60% of the time, compared with 45% of the alarms from standard monitors. The success of this system was likely to be due to the fact that the set of rules was optimized for these patient conditions, taking advantage of the homogeneity of the patient group.

In Schoenberg et al.’s expert system, a customizable “logic engine” was produced that was able to interpret information from multiple single-channel monitors. The proposed system analyzes a set of user-defined features that can be extracted from the raw vital signs. Thresholds were then defined for each feature, based on expert advice, and a feature was assigned a score if it exceeded the threshold value. The sum of the scores was then compared to a further threshold, which triggered an alarm if exceeded. The device was compared to single-channel alarms that staff classified as being clinically relevant during a period of over 120 hours in an ICU. The positive predictive value of the proposed system was 32% compared with 3% from the conventional single-channel system.

More recently Charbonnier and Gentil have presented a system based on trend analysis in which, each vital-sign is converted into a semi-quantitative temporal feature represented by symbols such as {Increasing, Decreasing, Steady}. The quantitative data such as the start- and end-times of the event, and the start- and end-values of the vital-

\[\text{Positive predictive value}\] is the proportion of patients with positive test results who are correctly diagnosed.
sign are also determined. The resulting collection of features is then used in a rule-based system to trigger an alarm when the trend is deemed to be persistent and severe. The proposed system was tested on-line using 36h of data recorded on adult patients in an ICU. When compared with a traditional single-channel alarm system, the proposed system showed that it could reject 33% of false single-channel alarms without missing any clinically relevant alarms.

The major limitations of knowledge-based expert systems are: 1) rule-based systems are simple when compared with the number of possible outcomes for alarm generation; 2) medical knowledge is difficult to capture and translate into rules; 3) rule formation is labour-intensive for clinicians, and the systems have to be repetitively evaluated in the clinical context to obtain acceptable levels of performance and 4) re-evaluation is required when the rules change [2].

Most of the recent approaches based on machine learning techniques have adopted a novelty detection scheme. The latter is the identification of new or unknown data when compared with a set of “normal” data [32]. Novelty detection has been used in applications where a much larger number of examples of the “normal” condition is available than those for the “abnormal” condition. Even when the latter exist they typically do not represent all of the possible modes of abnormality that a complex system can enter. This concept has already being applied in some studies in neonatal ICU [14], HDU [57] and one observational study in the ED [17].

Williams et al. have shown that a Factorial Switching Kalman Filter (FSKF) model can be applied successfully to complex monitoring data from neonatal ICU patients [14] using a novelty detection approach. Given a set of observations, the FSKF is used to calculate the state of the patient.

Tarassenko et al. [55] reviewed the use of a data fusion system, “Biosign” [55], to trigger the timely intervention of a Medical Emergency Team (MET) for at-risk patients in a step down ward. Biosign (now named as Visensia) adopts a probabilistic model of normality that uses 5 vital signs. The fusion of vital signs acquired from the patient and their
comparison against a model of normality, allows the determination of a single-parameter representation of patient status, the Patient Status Index (PSI). Visensia alerts occurred either when a single parameter deviates by approximately $\pm 3$ standard deviations (SDs) from its mean “normal” value or when two or more parameters depart from normality, but by some smaller quantity. In a trial with high-risk patients, the system alerts were generated every 8h on average, and 95% of these were classified as “true” by clinical experts. The authors also concluded that the proposed data fusion algorithm was capable of detecting critical events in advance of single-channel alerts from conventional patient monitors.

Clifton et al. [17] presented preliminary results of an observational study where novelty detection techniques were used to detect physiological deterioration in patients from the ED. Two machine learning techniques were used: 1) a kernel density estimator, and 2) a one-class support vector machine (SVM). Both approaches were used to score bedside monitor vital-sign data from 472 adult patients during their stay in the ED. The results showed that when given the task of identifying the clinical escalations in patients corresponding to abnormality both machine learning methods provided an increase in sensitivity and specificity compared with existing manual Early Warning Scoring (EWS) systems.

Clifton et al. concluded by stating that moving from a population-based to a patient-based modelling approach could improve sensitivity to abnormalities. Similarly, Zhang proposed a personalized model that attempts to increase alarm specificity by automatically tuning alarm thresholds on a per-patient basis [7]. In his study, both neural networks and classification trees were used to generate personalized alert thresholds. It was shown that neural networks performed consistently better than classification trees and simple single-channel system. However, in the case of the ED, small quantities of training data would be available for patient-specific models due to the short average length-of-stay (less than four hours), introducing significant model uncertainty. Clifton et al. [17] suggest that in this case a Bayesian approach is likely to be more appropriate.
1.3 Novelty Detection

Novelty detection is the identification of new unknown data or signal that a machine learning system is not aware of during training [32]. Novelty (outlier [9], anomaly [10], exception [34]) is a pattern in the data that does not conform to the expected behaviour. An inevitable consequence of the degree of system complexity is the large number of possible failure modes, the effects of which on observable (sensor) data are often poorly defined. Furthermore, examples of abnormal behaviour in high-integrity systems, such as, gas-turbine engines from airplanes or even humans, are few and far between, being usually insufficient to construct accurate fault-detection systems. The novelty detection paradigm machine learning approach allows the identification of abnormal system behaviour which is not consistent with the normal state of the system [56].

The framework of novelty detection usually includes the following steps: 1) input data preprocessing, that removes artefacts from the data; 2) feature extraction that represents input signals using a smaller set of quantities, named features, to reduce the dimensionality of the input; 3) construction of feature vectors and normalization (using for instance zero-mean, unit-variance transform); 4) data visualization, to study how the D-dimensional feature vectors are distributed over the space of normal data, specially near the boundaries of normality. Neuroscale [29] and Sammon Mapping [50] are methods that convert D-dimensional data into 2-D or 3-D space; 5) the obtained feature vectors are used to create a model (region) of the normal operation of a system. The information about novelty of test data is given as a final result. Abnormal operation of a system is detected by novel points outside normal region.

Overview of novelty detection methods

An taxonomy for novelty detection methods is now described, with a focus on methods that have already been applied to medical data. Chandola et al. [9] define a novelty detection taxonomy, where the different method were divided into, classification methods, Nearest Neighbor methods, clustering methods, Statistical methods and others.
In classification-based novelty detection a classification model is built for normal patterns based on labelled training data and few anomalous examples are added to the data. The model is used to classify each new unseen pattern. Common classification approaches include "Rule based", Neural Networks and SVM-based and the latter two have been used to classify medical data. Neural networks have been used extensively in novelty detection. They have the advantage that a very small number of parameters need to be optimized for training networks and no a priory assumptions about the properties of data are made \cite{6,31}. The goal of this approach is to train the neural network on the normal training data and then detect novelties by analyzing the response of the trained neural network to a test input. If the network accepts \(^3\) a test input, it is normal and if the network rejects a test input, it is novelty. Many neural network architectures exist. As an example, Penny and Frost \cite{40} used Multilayer Perceptron (MLP) neural networks architecture to detect outliers in medical diagnostics data.

SVMs are based on the concept of determining optimal hyperplanes for separating data from different classes \cite{7}. The one-class SVM proposed by Schölkopf et al. in \textit{51} is considered here, since it has been used in the medical area to identify deterioration (novel) periods of vital-signs time-series from patients recovering from cancer surgery \textit{18} and to identify brain activity associated with a specific motor activity in fMRI analysis \textit{19}. In this method a hyper-plane is constructed which is maximally distant from the origin and such that the margin \(^4\) is positive. The algorithm computes a binary function: the function returns +1 in "small" regions that contain data and -1 elsewhere. The data is mapped into the feature space corresponding to the kernel and is separated from the origin with maximum margin. A variable \(v\) which takes values between 0 and 1 is used to control how hard or soft the boundary is around the data.

\(^3\)Accepting a test input means that the reconstruction error, given in the testing phase, between an input and a reconstructed output is below an optimal threshold.

\(^4\)There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. The hyperplane is chosen so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane. The latter is separated from the normal data points, with output \(y_i=1\), by the positive margin.
The Nearest Neighbor (NN) approach for novelty detection is based on the assumption that normal points have close neighbors while novelty points are located far from other points [24]. This approach can be 1) distance based, such as the case of the k-NN algorithm, where the object class is determined by the class most common to its $k$ nearest neighbors; or 2) density based, for which the computation of a local outlier factor is included to indicate how strongly an instance can be considered to be novel [9].

The clustering-based method assumes that normal data belong to large and dense clusters, while novel data do not belong to any cluster. Each point is assigned a degree of membership to each cluster and a defined threshold then defines the final membership of the point. Novel patterns are examples that are not assigned a final membership [21].

Statistical approaches are based on modelling data based on its statistical properties and using this information to estimate whether a test sample comes from the same distribution or not. The estimation of the model of normality can be done using parametric or non-parametric methods [56]. Parametric techniques, assume an underlying data distribution, and the parameters of the distribution are estimated from the observed data. The most commonly used form of distribution for continuous variables is the gaussian distribution, which is defined by its mean and variance parameters. These are estimated from the data using the maximum likelihood method. Gaussian Mixture Models (GMMs), or other mixtures of different types of distributions, such as the gamma distribution, are used to model more complex forms of data. A numerical technique such as the expectation-maximization [20] algorithm is used in this case to estimate the parameters of the GMMs. These models might need a large training dataset to estimate the parameters accurately and also the chosen functional form may not be a good model of the distribution that generates the data.

The selection of the novelty threshold can be made in a principled fashion by introducing another parametric technique called Extreme Value Theory (EVT). Given a set of normal training $X = \{x_1 \ldots x_m\}$, EVT estimates the probability distribution of the maximum of that set, $\max(X)$. The threshold is set according to where the maximum
of the normal data will occur, and thus provides a principled method of setting novelty thresholds. The use of "classical" EVT in novelty detection for biomedical applications was first developed by Roberts et al. [46][48] and Numerical solutions to set multivariate novelty thresholds using EVT can be found in the work of Hugueny et al. [26] and Clifton et al. [16]. Other parametric approaches include Hidden Markov Models (HMM), which are ideal to detect novel sequences in time-series data [22], and Hypothesis testing in which a t-test can be used to compare a sample with a Gaussian distribution. If the test shows significant difference between two sets of measurements (normal profile and test profile), then the second set is considered to contain novel patterns [32].

Nonparametric approaches which make as few assumptions as possible about the form of the distribution, include kernel density estimators, such as histograms, box-plots, Parzen windows. The latter can be used to estimate the density of the training (normal) data. \( p(x) \) is estimated in 2 steps [7]: 1) locating a hyperspherical Gaussian window, or kernel, with width \( \sigma \), on each of the D-dimensional feature vectors in the training dataset, \( x_i \), where \( i=1, \cdots, N \); (a one-step "training phase") and 2) evaluating the sum of the Gaussian distributions using the squared Euclidean distances between the test feature vector \( x \) and the training vectors \( x_i \), normalized by a factor that ensures \( p(x) \) integrates to 1. \( p(x) \) will have a higher value where the concentration of training data is greatest. Two parameters must be carefully chosen, the number of training (normal) patterns and \( \sigma \), the width of the Gaussian Kernel. \( N \) must not be to large to avoid undue computational complexity. Methods for reducing the training data to a number of prototype points, with outlier removal, have been developed here [56]. If \( \sigma \) is too large, the local variation in the D-Dimensional space is not captured. If the value is too small, the estimate of the pdf is too noisy and follows the data too closely (overfitting). A simple method to set \( \sigma \) is to use the average distance of the 10 nearest neighbors from each vector in the (normal) training dataset, averaged across all the selected prototype feature vectors. As with the case of GMMs, novelty increases as \( p(x) \) decreases and the novelty threshold can be set using the EVT approach.
Novelty detection can be used with Supervised, Semi-supervised and Unsupervised learning paradigms. The methods may be used off-line, when the data is evaluated retrospectively, or on-line, where the normal behaviour of a system is changing over the time. In this case the need arises to dynamically update the normal profile of a system. For this reason Gaussian Processes are introduced next, as a way to model time-series dynamics and improve the novelty score with such information [61].

1.4 Gaussian Processes for Time-series Modelling

One limitation of the data fusion model currently being tested to detect patient deterioration in the Emergency Department, is its inability to evaluate the vital-signs time-series trends for a patient [61]. Only the most recent vital sign vector is used to access the patient’s physiological status. This differs from standard clinical practice, in which trend analysis is considered to be an important part of the detection of deterioration.

Time-series analysis is a broad field, with applications in economics [58], meteorology [43], among others. A number of frameworks, such as Kalman filters [28] and particle filters [4], exist to effectively evaluate time-series data but in this review only Gaussian Processes (GPs) are considered. A GP is defined as a stochastic process for which any finite combination of samples have a joint Gaussian distribution [44]. It has been shown that GP models have a close relationship to a Bayesian neural network [13, 36]. GPs can be applied in the context of classification, regression or optimization [39]. GP regression has been used in a number of diverse applications including, time-series visualization [37], regression of noisy Heart Rate data [53], predicting patient condition in intensive care [23] and multi-sensor time-series prediction [39, 45].

GPs can be defined as being a probability distribution over a infinite number of variables, such that the distribution over finite subset of them is a multi-variate Gaussian. The Gaussian distribution provides useful mathematical properties for a regression problem. In particular, the marginals of a multivariate Gaussian distribution are themselves
Gaussian and the distribution of a subset of variables conditioned upon any other subset is also Gaussian. These properties allow the evaluation of subsets of potentially infinite lists of function outputs and inputs [39]. The first property allows us to marginalize over any unobserved values of a function, even in the case of an infinite number of such values. Since the marginals are Gaussian and considering the second property, inference is possible, constrained to the space of Gaussians. A GP distribution is defined by a mean and covariance function and can be represented by the distribution [45]:
\[ p(y(x)) = N(\mu(x), K(x, x)) \]
where \( y = y_1, \ldots, y_n \) refers to a possible vector of function outputs, evaluated in locations \( x_1, \ldots, x_n \), \( \mu \) the mean function evaluated at the locations of the \( x \) variables, \( K \) the covariance kernel function [45] which provides the covariance element between any two (arbitrary) sample locations, \( x_i \) and \( x_j \).

Covariance functions can be created by adding or multiplying other covariance functions. This allows for multi-scale behaviour to be modelled. An example is given by Stege et al. [53], who attempted to infer missing heart rate data using GPs. 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.

Both mean and covariance functions are specified by a set of hyperparameters (that is, parameters of the GP prior), and marginalizing these hyperparameters, is the challenge of GP prediction. The selection of the covariance and mean functions is important to model the system dynamics. A common choice for the mean function is to make it zero or non-zero constant for all inputs. Popular non-constant mean functions include polynomials of various orders [39]. In the case of regression the mean function will dominate forecasts in regions far from the data, so its specification should be more thoughtful [45]. The parameterization of the covariance matrix is free but it must lead to a positive semi-definite matrix. Fortunately, a wide variety of well-studied functions exist [1, 54]. These can be combined to model periodicity, delay, noise, long-term drifts and other phenomena. Examples of commonly used covariance functions are the squared exponential, the Matérn and the rational quadratic function.
One of the main areas of interest has been in developing methods to reduce the computational cost of GP regression, both in the training and prediction phases. Both have non-trivial dependencies upon the hyperparameters used to generate the mean vectors and covariance matrices (hyperparameter marginalization). A method such as maximum likelihood can be used as a numerical integration solution. More recently more sophisticated methods such as Laplace approximation \[30\], Variational Bayes \[7\] or Expectation Propagation \[35\] have been explored.

### 1.5 Future Work

#### 1.5.1 2nd Phase ED trial data collection and validation

A new ED study is currently being executed 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 started on 23rd of April 2012 and 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 \[41\], 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. In the third phase of the study, nurse observations will continue to be recorded electronically. In addition, the Visensia algorithm, modified for greater
sensitivity to hypotension, 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 randomized. However, practically, it was deemed impractical to return to paper-based observations after the introduction of an electronic system.

The first phase of the trial finished in the first week of July. It is estimated that the second phase will start in October. If everything goes according to the protocol the trial itself will not take more time than 4.5 months ending in the middle of February 2013. The data from the first stage (approximately 3000 patients) is still being collected from the T&T paper charts and patient clinical notes. Data from the second and third phases should take less time to collect since there is no need to transcribe T&T paper charts. The transcription error for the first phase data has already been done for 200 patient notes and it has been found to be 2%. In the new study, the outcome measure is 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.

1.5.2 Machine Learning

Novelty Detection - Static Modelling

A selection of the methods presented in this literature review will be used to build a population based (static modelling) model to score abnormality in vital-signs from the ED dataset. This will also include a selection of optimal vital signs variable to include in the model and also make the population based models more tuned to specific populations. Continuous variables such as age, and categorical variables such as patient presenting complaint and discriminator, which are captured at triage time, will be used to tune these population-based models. Age has been confirmed as an important factor to take into consideration to score abnormal vital-signs. This work will run in parallel with the
data collection phase.

**Time series modelling with Gaussian Processes**

Gaussian processes will be used as a tool for modelling vital sign data and for regression. Work conducted in the second year of research will evaluate what combination of covariance matrices best model the vital sign time series. Also there has been very recent work on clustering time-series modelled by Gaussian Processes using curve analysis [52]. It will also be explored if vital-sign trajectories of the ED database can be clustered in groups of patients with similar clinical conditions.

**Patient event detector**

The second ED trial will have vital-signs data from 9000 patients but this data will not be labelled in the short and medium term. Therefore a methodology is needed to automatically find clinical relevant events that would distinguish the most abnormal and normal patients. Semi-supervised learning techniques [11] that learn the clinical event periods from the validated labelled data from the first ED trial and identify them in the new dataset, will be used. Both the static and dynamic models studied before will be used here. A good example is the work of Khalid et al. [27], who developed an approach whereby the static novelty detection models of normality and the clinically labeled abnormal vital-signs were used in a two-class classification approach to detect patient deterioration. Clustering of dynamical models could complement this work in detecting patient deterioration.

**Patient-specific Models for rest of hospital stay**

One of the objectives of the ED trial is to investigate whether the continuous vital-signs data captured during a patient’s stay in the ED could be used to suggest a risk score for patients that are admitted to other hospital wards. This risk information should be communicated to the monitoring systems used in the wards to which the patients are transferred. The long-term goal is to develop patient-specific models from the ED data and use these to better monitor the patient in the next ward.
Appendix A: DPhil Project Plan

Figure 1: DPhil project plan, completion deadline: October 2014.
Bibliography


