A Large-Scale Clinical Validation of an Integrated Monitoring System in the Emergency Department

David A. Clifton, David Wong, Lei Clifton, Sarah Wilson, Rob Way, Richard Pullinger, and Lionel Tarassenko

***Abstract*—We consider an integrated patient monitoring system, combining electronic patient records with high-rate acquisition of patient physiological data. There remain many challenges in increasing the robustness of "e-health" applications to a level at which they are clinically useful, particularly in the use of automated algorithms used to detect and cope with artifact in data contained within the electronic patient record, and in analyzing and communicating the resultant data for reporting to clinicians. There is a consequential "plague of pilots," in which engineering prototype systems do not enter into clinical use. This paper describes an approach in which, for the first time, the Emergency Department (ED) of a major research hospital has adopted such systems for use during a large clinical trial. We describe the disadvantages of existing evaluation metrics when applied to such large trials, and propose a solution suitable for large-scale validation. We demonstrate that machine learning technologies embedded within healthcare information systems can provide clinical benefit, with the potential to improve patient outcomes in the busy environment of a major ED and other high-dependence areas of patient care.**

***Index Terms*—Biomedical informatics, biomedical signal processing, machine learning.**

I. Introduction

THE rapid pace of development in "e-health" technologies within integrated healthcare systems (such as electronic patient records) has far outpaced their uptake in clinical practice. There is a perceived "plague of pilots" [1], in which prototype systems do not penetrate into clinical use, and there is a consequence lack of evidence required for adoption at scale [2].

We address this problem by describing a large clinical trial in which the care of 10 000 patients in the Emergency Department of a major research hospital[1] switches over to the use of integrated healthcare systems that we have designed around best-practice principles in machine learning.

Adopting new healthcare systems at scale in a clinical environment is a time-consuming and resource-intensive process, particularly in building the large bodies of evidence required to support adoption. This paper describes the trial needed to provide this evidence, in which algorithms for detecting physiological deterioration are embedded within an integrated healthcare system and are compared with the existing standard of hospital care (where the latter is introduced in Section II). The infrastructure of the system is described in Section III; Section IV investigates the shortcomings of existing methods of using that infrastructure for patient care, and proposes techniques to overcome these shortcomings. We describe how methodologies for evaluating the success of electronic systems in clinical studies are inadequate: Section V describes the dangers involved with using traditional evaluation methods, and proposes a methodology that avoids these problems. Results, obtained from comparing existing and proposed methods, are presented and discussed in Section VI. Conclusions are drawn, and future work is considered, in Section VII.

II. Existing Standard of Care

Before the introduction of novel "e-health" systems at scale becomes possible, considerable evidence must be acquired concerning the existing standards of care, which may be obtained, for example, from clinical trials. For systems that monitor patient physiology, we must determine the efficacy of existing methods of patient observation.

Adverse events occur when the physiological condition of patients is not recognized or acted upon [3]. This has resulted in clinical guidance being provided for the U.K. in which the monitoring of certain vital signs[2] was recommended, followed by the suggestion that manual "early warning score" (EWS) systems are used [4]. The latter involve the clinician making a manual observation of a patient's vital signs, applying univariate scoring criteria to each vital sign in turn (e.g., "score 3 if heart rate exceeds 140 beats per minute"), and then escalating care to a higher level if any of the scores assigned to individual vital signs, or the sum of all such scores, exceeds some threshold. There are several disadvantages of this existing standard of care, against which novel methods must be evaluated:

1) The scores assigned to each vital sign, and the thresholds against which the scores are compared, are mostly determined heuristically, according to clinical opinion. This leads to significant differences between EWS

Manuscript received August 16, 2012; revised November 20, 2012; accepted December 10, 2012. Date of publication December 20, 2012; date of current version June 27, 2013. The work of D. A. Clifton was supported by the Centre of Excellence in Personalized Healthcare funded by the Wellcome Trust and EPSRC under Grant WT 088877/Z/09/Z. The work of D. Wong and L. Clifton was supported by the NIHR (National Institute for Health Research) Biomedical Research Centre Programme, Oxford.

D. A. Clifton, D. Wong, L. Clifton, and L. Tarassenko are with the Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, OX3 7DQ, U.K. (e-mail: david.clifton@eng.ox.ac.uk; wong@robots.ox.ac.uk; lei.clifton@eng.ox.ac.uk; lionel@robots.ox.ac.uk).

S. Wilson, R. Way, and R. Pullinger are with the Oxford University Hospitals NHS Trust, Oxford, OX3 9DU, U.K. (e-mail: sarah.wilson@ouh.nhs.uk; rob.way@ouh.nhs.uk; rick.pullinger@ouh.nhs.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2012.2234130

[1]John Radcliffe Hospital, Oxford University Hospitals NHS Trust, Oxford, U.K.

[2]Heart rate (HR) measured in beats per minute, respiration rate (RR) measured in breaths per minute, peripheral blood-oxygen saturation (SpO_2) measured as a percentage, systolic blood pressure (BP) measured in mmHg, and body temperature.

TABLE I
EWS SYSTEM USED IN THE ED DURING INITIAL DATA ACQUISITION

Score:	3	2	1	0	1	2	3
HR:	≤ 40	41 - 50	51 - 100	101 - 110	111 - 129	≥ 130	
RR:	≤ 8			9 - 18	19 - 24	25 - 29	≥ 30
SpO_2:	≤ 91			≥ 92			
BP:	≤ 90	91 - 99		100 - 179			≥ 180
Temp:		≤ 35.0		35.1 - 37.9			≥ 38.0

systems used in different hospitals, or even between wards in the same hospital [5]. An attempt has recently been made to define these quantities using a large evidence base of vital-sign data [6]. We argue that standardization of the approach, using an evidence-based procedure, can overcome the disadvantages of heuristic methods.

2) EWS systems are used only after routine observation of patient vital signs, which may be performed as infrequently as once every 4 h in some wards. Patients may deteriorate significantly between observations. It may be that automated systems can operate continuously; however, existing bed-side monitoring systems suffer from a false-alarm rate of up to 86% [7]. This motivates an integrated approach, based on machine learning methods, which can drive down the false-positive alarm rate to clinically useful levels, as we will describe later in this paper.
3) Many hospitals use EWS systems manually, where the clinician assigns scores using a lookup table, adds the scores, and then compares the resulting totals to predefined thresholds. There is a significant error-rate associated with this manual arithmetic [8]–[10], suggesting that automated methods can avoid such errors.
4) Scores in EWS systems are univariate (so that they may be used easily by the ward staff), and, therefore, do not take into account the covariance between vital signs. When comparing the sum of scores from individual vital signs to a threshold, an assumption is implicitly made that the vital signs are independent, which is unlikely to be true in practice. Automated methods, such as those proposed in Section III, can learn the dependence between vital signs, and thereby better assess patient physiology than basic EWS systems.
5) The evidence base for evaluation of EWS systems is poor, with most hospitals using scores that have been derived heuristically, rather than implementation being based on clinical data [5]. The clinical trial described in this paper aims to address this need for evidence.
6) EWS systems are population dependent, by necessity. Some systems exist for pediatric patients or other specific populations. Automated systems offer the possibility of patient-specific alerting, which is discussed later.

Two EWS systems are considered by the work described in this paper: an older system in use in the ED of the John Radcliffe Hospital at the time of initial data collection, and the evidence-based method described in [6] that has since been implemented in the ED. These two EWS systems are shown in Tables I and II, respectively.

III. INFRASTRUCTURE

This paper considers an integrated system that interfaces to a peer-to-peer network of bed-side monitors, hand-held PDAs, and wall-mounted touch screens, being introduced into the ED of the John Radcliffe Hospital as part of a 10 000-patient clinical trial, approved by the local medical ethics committee. Bed-side monitors acquire vital-sign data from patients in real time using ECG electrodes, pulse oximeters, and sphygmomanometers, at sampling intervals of approximately $T_s = 20$ s (and where BP is measured approximately every 30–60 min). A secure wireless network communicates vital-sign data to a central server, where algorithms process the data, identify periods of communication and sensor failure, and then analyze the data with respect to models of patient physiology (as described in Section IV).

As with the manual EWS systems described in Section II, the goal is to identify periods of patient deterioration, and so each patient is assigned a patient status index (PSI), which takes low values when patient physiology is stable, and which takes higher values when patient physiology is deemed to be indicative of patient deterioration. The PSI and other results of analysis are then displayed at the patient bedside, and can be further summarized on the wall-mounted touch screens (for providing an overview of the health status of all patients in a ward), and to clinicians' hand-held PDAs.

It is important to note that fully automated patient monitoring systems cannot entirely replace the process of manual patient observation, because the ward staff are required to perform patient reviews; in many wards, nurses are encouraged to make physical contact with the patient to estimate pulse rate and to determine respiration rate. However, rather than relying on the traditional paper-based methods of recording and scoring patient physiology, our trial includes the facility for clinicians to input vital-sign data from manual patient reviews into their hand-held PDAs. These data are then automatically scored using an EWS (as shown in Table II), and transmitted to the central server for more detailed multivariate analysis combined with the continuous data acquired from bed-side monitors, as described in Section IV. This effectively provides benefit through, as described in Section II, performing continuous, automated assessment of patient condition, and taking a multivariate approach that does not treat the vital signs as if they were independent random variables.

During the initial phase of the trial, over 105 GB of patient data were acquired, totaling approximately 2170 h of patient data. The data completeness for each vital sign is shown in Table III. Such data loss is typically caused by disconnection of sensors from the patient, either by accident (such as a pulse oximeter falling from its place on the earlobe or finger) or by intention of the clinician.

We note that the inclusion of continuous temperature measurement into the electronic patient record, and its use in both manual and automated analyses, is problematic. Continuous temperature measurement using sensors is difficult; our trial initially attempted to collect data using a skin-mounted thermistor, to collect skin temperature as a proxy for the core temperature typically used by clinicians in the diagnosis of patients.

TABLE II
EVIDENCE-BASED EWS SYSTEM SUBSEQUENTLY ADOPTED IN THE ED

Score:	3	2	1	0	1	2	3
HR:	≤ 42	43 - 49	50 - 53	54 - 104	105 - 112	113 - 127	≥ 128
RR:	≤ 7	8 - 10	11 - 13	14 - 25	26 - 28	29 - 33	≥ 34
SpO_2:	≤ 84	85 - 90	91 - 93	≥ 94			
BP:	≤ 85	86 - 96	97 - 101	102 - 154	155 - 164	165 - 184	≥ 185
Temp:	≤ 35.4		35.5 - 35.9	36.0 - 37.3	37.4 - 38.3		≥ 38.4

TABLE III
DATA QUANTITY AND DATA LOSS DURING INITIAL DATA ACQUISITION

	HR	RR	SpO_2	SBP
Time (hours)	1,645	1,629	1,664	1,776
Data loss (%)	24.2	24.9	23.3	18.2

However, the drop-out rate for this channel exceeded 75%, the standard deviation of the resulting temperature signal exceeded 2 °C, and the mean varied significantly between patients (sometimes exceeding a mean shift of 5 °C between patients). These poor signal statistics made identification of patient deterioration using continuous temperature data impossible, where changes of 1–2 °C can be significant indicators of physiological distress.

We therefore used manual measurements of temperature. While manual methods of tympanic measurement (measured using an in-ear probe) exhibit low variance, are close to core temperature, and are the "gold standard" for the existing standard of care, the resultant data are often inappropriate for use due to device configurations that can cause significant mean shifts that could make "healthy" patients seem "abnormal" in their temperature, and "unhealthy" patients seem "normal." For example, a time series of tympanic measurements from devices at the Oxford University Hospitals NHS Trust is shown in Fig. 1(a). The measurement devices were known to have had settings switched, causing a small offset to be applied to the N temperature data $\mathbf{X} = \{x_m\}, m = 1 \ldots N$ at some point during the period shown in the figure. A Bayesian change-point detector [11] can be used to find a posterior distribution $p(m|\mathbf{X})$ over the indices m of the data

$$p(m|\mathbf{X}) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(m, \mu_1, \mu_2, \sigma|\mathbf{X}) \mathrm{d}\mu_1 \mathrm{d}\mu_2 \mathrm{d}\sigma \quad (1)$$

$$\propto \frac{1}{\sqrt{B}} \left[\sum_{i=1}^N x_i^2 - \frac{1}{m} S_l^2 - \frac{1}{N-m} S_r^2 \right]^{(1-\frac{N}{2})} \quad (2)$$

where $S_l = \sum_{i=1}^m x_i$ and $S_r = \sum_{i=m+1}^N x_i$ are the sums of the data up to and after change-point m, respectively, and where constant $B = m(N-m)$. Here, we have assumed that the data up to and after the change-point have distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, and we have used (1) to integrate out the nuisance parameters μ_1, μ_2, σ, leaving us with the posterior change-point distribution over m. The result shown in Fig. 1(b) shows that a step-change occurring after 15 months has been detected. We note that the formulation has constrained $p(m|\mathbf{X})$ to a single change-point, and hence short-term deviations, such as a temporary decrease at around 12–13 months, do not affect detection of the change-point. Accurate detection

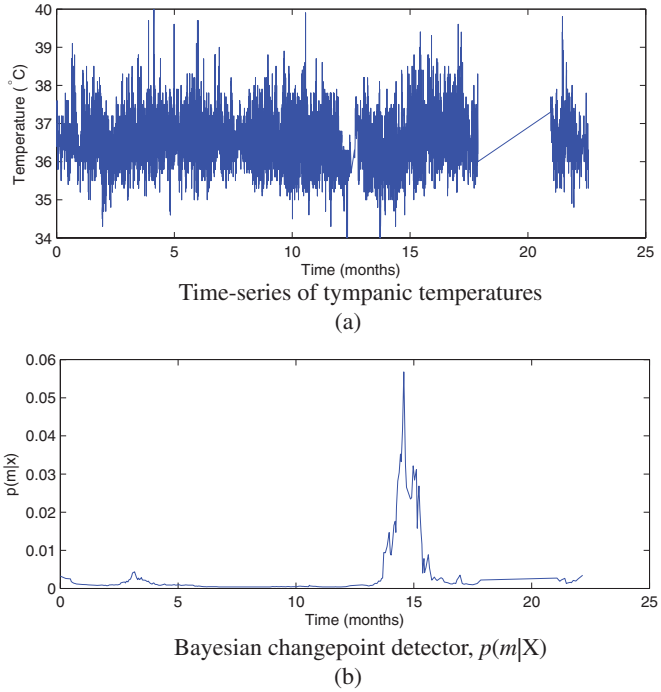

Fig. 1. (a) Time series of tympanic temperature measurements, $\mathbf{X} = \{x_m\}$, $m = 1 \ldots N$. (b) Corresponding change-point distribution $p(m|\mathbf{X})$.

of regime changes, such as the simple reconfiguration of temperature probes, allows us to cope with such step-changes in input in a principled manner. We note here that the time series of temperatures was taken over a long period, and over many patients, such that the variability between individual patients can be assumed to be a constant effect over the whole time series, and thus not affect change-point detection of a mean shift.

IV. NOVELTY DETECTION

The aim of clinical practice, in its use of integrated EWS systems, can be framed in machine learning terms as a *novelty detection* task, in which deviations away from normality are identified. The EWS performs this task by assigning scores to vital signs as they depart from normality. However, novelty detection may also be performed using principled machine learning techniques, appropriate for use in the automatic analysis of large quantities of physiological data.

This paper takes a novelty detection approach, in which a model of "normal" patient physiology is constructed. Test data are then compared to this model, and deemed "abnormal" if they differ significantly from it, where abnormality is defined differently according to the various novelty detection approaches

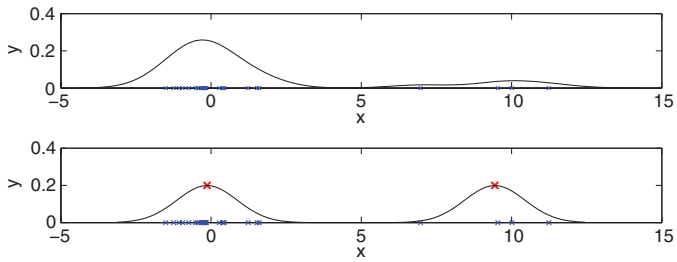

Fig. 2. Dataset $\{x\}$ and its distribution $y = p(x)$ are shown in the upper plot. $K = 2$ kernel means $\{y\}$ and the resulting kernel estimate $p(y)$ are shown in the lower plot.

described later. Novelty detection is typically performed in preference to a multiclass approach to classification when there is an insufficient quantity of data to model abnormal states accurately. This may occur when the abnormal states are so numerous that each cannot be fully specified (as in the case of high-dimensional datasets acquired from complex systems, such as human patients), or when examples of failure are rare (as in the case of some physiological conditions).

This section describes several approaches to novelty detection, which will be compared for use with the data obtained at the central server described in Section III.

A. Inference With Kernel Estimates

Previous work [12] has modeled the joint distribution $p(\mathbf{x})$ of vital signs $\mathbf{x} \in \mathbb{R}^5$, for the vital signs[3] shown previously. Each vital sign was standardized with respect to its own mean and variance, $x' = (x - \mu)/\sigma$. The joint distribution of the (normalized) training data was estimated using a mixture of Gaussian distributions: a kernel estimate [13] with 500 components. The procedure used to estimate this distribution involved first summarizing (using the k-means clustering algorithm [13]) a set of approximately 2.3×10^6 data, corresponding to over 3000 h of vital-sign data acquired from acute patients. This summarization process is necessary due to the large size of the dataset.

The likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ of previously unseen test data $\mathbf{x}$ is then evaluated with respect to the kernel estimate (parameterized by $\boldsymbol{\theta}$) and used to generate a corresponding novelty score, $z(\mathbf{x}) = -\log p(\mathbf{x}|\boldsymbol{\theta})$. This novelty score takes high values when the test data are "abnormal" with respect to the model of normality. Thus, the novelty score may be seen as a probabilistic version of the manual EWS, with the advantages that the vital signs are not treated independently (because the joint density of all vital signs is estimated), and that it is not heuristic.

A threshold κ is defined on z such that test data $\mathbf{x}$ are deemed "abnormal" with respect to the joint pdf if $z(\mathbf{x}) > \kappa$. Communication failures, network and sensor noise, and other transients can cause temporary, artifactual spikes in continuous novelty scores. To avoid false-positive alerts caused by these transients, our method only communicates a novelty alert to the clinician (thus calling them back to the patient's bedside) when $z(\mathbf{x}) > \kappa$ for 4 minutes out of any 5-minute window of data.[4] The value of κ was similarly selected using cross-validation with an independent validation set, selected from over 18 000 h of vital-sign data acquired from acute patients [12].

B. Modified Kernels

The summarization process described previously, in which an integrated dataset of physiological data is reduced to a smaller set of distribution means $\{\mathbf{y}_i\}, i = 1 \cdots K$, can adversely affect the quality of the resulting estimate of the joint distribution, $p(\mathbf{x}|\boldsymbol{\theta})$. Fig. 2 illustrates this disadvantage with an exemplar set of univariate data (shown as crosses on the x-axes of the plots in the figure). The required distribution $p(x)$ is shown in the upper plot; however, summarizing the dataset using a smaller number of distribution centers $\{y_i\}$, and then forming a kernel estimate of the distribution of those centers $p(y)$, is a poor approximation of $p(x)$. This is shown in the lower plot in the figure, where the k-means algorithm has placed $k = 2$ distribution centers that describe each of the two clusters of data, even though one has a significantly smaller mode than the other in the desired distribution $p(x)$. This problem is caused by the k-means algorithm, which minimizes squared distances between the dataset $\{x_i\}, i = 1 \cdots N$ and the centers $\{y_j\}, j = 1 \cdots K$, and by the kernel estimate from Section IV-A which assigns equal prior probabilities $\pi_j = K^{-1}$ to each of the kernels.

The appropriate weighting of the kernels in the density estimate may be obtained using a modified kernel estimate [14]

$$p(\mathbf{x}|\mathbf{y}_i) = \sum_{i=1}^{K} \frac{\pi_i}{\sigma} \Phi_i \left(\frac{\|\mathbf{x} - \mathbf{y}_i\|}{\sigma} \right) \tag{3}$$

which is the kernel estimate from Section IV-A to which has been added priors π_i, and where $\sum_{i=1}^{K} \pi_i = 1$. In the above, the ith kernel Φ_i is a Gaussian distribution, characterized by mean $\mathbf{y}_i$ and isotropic covariance Σ with diagonal elements σ. The priors are determined from the proportion of data $\mathbf{x}_i$ that fall within the ith cluster, $\pi_i = N^{-1} \sum_{j \in \mathbf{y}_i} \mathbb{1}$, where $\mathbb{1}$ is the indicator function.

Fig. 3 shows 2-D visualizations of the results of applying the modified kernel estimate to the 4-D training set of 3000 h of patient data described previously, where the visualization has been performed using the SASS [15] projection, which attempts to preserve distances between data before and after projection. Noting that the training set comprises both "normal" and "abnormal" patients, we follow [12] in discarding those 20% of the resulting kernels (i.e., 100 of the $K = 500$ kernels) deemed to be "outlying." We will consider two definitions of "outlying": one where the 100 kernels with the lowest associated priors π_i have been discarded (and so which have the lowest population), and one where the 100 kernels furthest from the centroid of the data have been discarded. These are shown in the left and right plots, respectively, in Fig. 3. We will refer to these modified

[3]These are vital signs where the systolic blood pressure was replaced by the mean of systolic and diastolic blood pressure (the *systolic–diastolic average*, or SDA) in order to take into account changes in both blood pressure measurements that are typically acquired, rather than just the systolic.

[4]We note in passing that this is statistically equivalent to generating a novelty alert if $r_{20} > \kappa$, where r_n is the nth order statistic, taken from the order statistics of a window of scores $z(\mathbf{x})$ of test data $\mathbf{x}$.

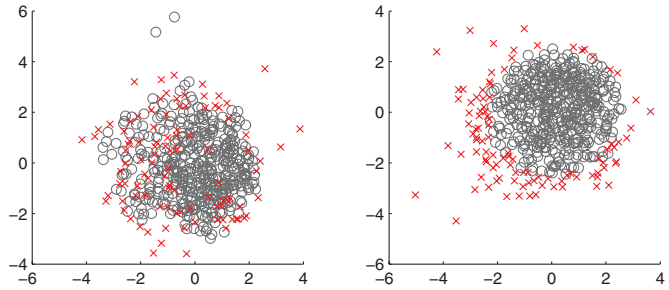

Fig. 3. Two-dimensional visualizations of 4-D kernel means $\{\mathbf{y}_i\}$. Those kernel means with the 100 smallest populations and 100 greatest distances to the centroid of the data are shown by $\times$ in the left and right plots, respectively.

kernel methods as K_{w1} and K_{w2}, respectively, and to the original (unmodified) kernel estimate from Section IV-A as K_0.

C. One-Class Support Vector Machines

We also consider the use of a one-class support vector machine (SVM), trained using the same data as that from which the kernel estimate described previously was obtained. We used the method proposed by Schölkopf *et al*. [16], in which the objective function is defined by separating the training data from the origin in the feature space defined by the SVM kernel, where we have used the squared-exponential kernel.[5]

The degree to which the SVM objective function is penalized by misclassifications (and thus the curvature of the decision boundary) is controlled by the "C" parameter in the conventional nomenclature, the value of which, along with the width parameter σ shared by all of the isotropic kernels in the model, was selected using tenfold cross validation.

A quantity N of d-dimensional data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^d$ are mapped into a (potentially infinite-dimensional) feature space $\mathbb{F}$ by some nonlinear transformation ϕ: $\mathbb{R}^d \rightarrow \mathbb{F}$. A kernel function Φ provides the dot product between pairs of transformed data in $\mathbb{F}$, such that $\Phi(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. A Gaussian kernel allows a point to be separated from the origin in $\mathbb{F}$ [16], hence is chosen for us in the work described by this paper: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2\right)$.

The decision boundary between "normal" and "abnormal" subspaces in $\mathbb{F}$ is $z(\mathbf{x}) = w_o \cdot \phi(\mathbf{x}) - \rho_0$, with parameters

$$w_o = \sum_{i=1}^{N_s} \alpha_i \phi(\mathbf{s}_i) \tag{4}$$

$$\rho_o = \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_{i=1}^{N_s} \alpha_i \Phi(\mathbf{s}_i, \mathbf{s}_j) \tag{5}$$

where $\mathbf{s}_i$ are the support vectors, of which there are N_s, and where k is the Gaussian kernel. Here, $w_o \in \mathbb{F}$, $\rho_o \in \mathbb{R}$, and the α_i are Lagrangian multipliers used to solve the dual formulation, more details of which may be found in [16] and which are not reproduced here. Test data $\mathbf{x}$ arriving in the electronic patient record are classified as being "abnormal" if $z(\mathbf{x}) > 0$, and "normal" otherwise. In order to allow a fair comparison with the probabilistic methods K_0, K_{w1}, and K_{w2} described in Sections IV-A and IV-B, an alert was generated if test data were classified "abnormal" for 4 minutes in any 5-minute window of test data, in order to avoid false-positive alerts due to transients arising from sensor artifact.

[5]This method typically performs similarly to the other popular one-class SVM formulation, the *support vector data description*, as proposed by Tax and Duin [17].

V. Evaluating Monitoring Systems

The evaluation of automated patient monitoring systems is not straightforward and remains an area of contention. This section describes the challenges involved, and proposes an evaluation strategy appropriate for the complex evaluation of an automated system running in real time.

A. Clinical Labels

In order to evaluate system performance, we would ideally have accurate labels of "normal" and "abnormal" episodes of data. The "gold standard" in classification problems is often a set of labels provided by domain experts—in the case described by this paper, such experts are clinical specialists. However, exhaustive labeling is typically not possible in practice due to the size of the datasets and the difficulty in determining patient abnormality from inspection of the vital signs alone. Furthermore, intra- and interexpert variability is such that the subjective nature of the labeling process becomes significant. This is a fundamental obstacle to the evaluation of automated systems, which are based on the communication and analysis of very large quantities of data.

Therefore, a one-sided approach is sometimes taken to evaluating large-scale systems, in which clinical experts are asked to review data in the electronic patient record corresponding to periods of suspected patient abnormality. These could be taken from, for example, "hard outcomes" such as death, unforeseen admission to intensive care, etc. However, the number of patients with these outcomes is typically small unless data are acquired from large numbers of patients, which may take many years. Therefore, this study takes the approach in which *clinical escalations* were taken as being possible indications of patient abnormality. These escalations are events that took place during the patient's stay, where the patient's clinical notes indicate that care was escalated to a high level due to some perceived abnormality.

There are many reasons for which a patient's care may be escalated in practice, only some of which are likely to correspond to derangement of the vital signs, and which could, therefore, be expected to be identified by an automatic method. Two experts in emergency medicine independently reviewed patient records (but not the continuous vital-sign data acquired from sensors and bed-side monitors) and identified those periods that corresponded to escalations that should be expected to have a corresponding change in the vital signs. Any differences in opinion between the two experts were resolved by the independent assessment of a third clinical expert. This labeling of data in the electronic patient record by experts is extremely time consuming, and is a primary reason why such large clinical trials

are not undertaken for the principled evaluation of automated monitoring systems.

B. Evaluation Methodology

Communications theory originated the received-operating characteristic (ROC) curve, which has since become the primary means of evaluating the decision outputs of medical devices. However, the evaluation of performance with respect to events occurring within time-series data is not straightforward.

1) Independence: The classical method for evaluating classifier performance is to construct a confusion matrix, which quantifies the number of true and false, positive and negative classifications (TP, FP, TN, and FN) made by the classifier, with respect to some "ideal" classification. The sensitivity and specificity of the classifier may then be plotted as a function of some variable of its operation (typically a parameter that controls its decision threshold), to give the ROC curve. Indeed, some EWS systems [18] were evaluated by maximizing the area under the ROC curve (AUROC), which corresponds to maximizing the accuracy[6] of the classifier.

Such approaches can be appropriate when evaluating the classification of independent entities, such as mammograms or blood samples taken from different patients, but are problematic when used to analyze time-series data, such as that considered in this study within the electronic patient record, which are not independent. The results could be biased, for example, by a small number of "abnormal" patients with long hospitals stays (which often occurs, because length-of-stay is correlated with physiological abnormality); these patients contribute a large proportion of data to the set of "abnormal events,", but where those events are largely *dependent*. Thus, the performance of the classifier would be skewed toward how well it performed for this small subset of patients.

We suggest that there is no "right" answer to the problem of how to evaluate a time-series classifier, and that, ultimately, it is probably inappropriate to reduce the performance of a system down to a single metric (e.g., accuracy/AUROC).

2) Patient-Based Analysis: If we wish to use ROC-based performance metrics in an evaluation (without reducing them to a single statistic), we must select a basic unit of analysis other than individual samples of vital-sign data within the electronic patient record. The assumption of independence between basic units can avoid being broken by performing the analysis on a *per-patient* basis. In this study, we adopt the following convention:

"Event" patients: This group comprises all patients containing one or more "events," defined to be those with escalations in their vital signs that occurred during their stay in the ED. Section VI will provide results from phase I of our study, which comprised 476 patients, 52% of which were male. The mean age was 61 years (range 18–108, IQR 43–79). There were 34 escalations in this population, indicating the scarcity of labeled event data.

[6]Accuracy is defined to be (TP + TN)/(TP + TN + FP + FN).

"Normal" patients: This group comprises all patients who had no clinical escalation of any kind, and corresponds to 217 (46%) of the 476 patients in phase I of our study.

We note that there is a set of patients that belongs to neither of the two sets described previously, being those who had no escalations due to their vital signs, but who were also not "definitely normal"; for example, they may have had escalations for reasons other than those related to vital signs, and which an automated monitoring system could, therefore, not be expected to detect. Similarly, some patients had no vital-sign data transmitted by our system, and could, therefore, not be used in either of the sets described previously.

We define a *TP* classification to be an "event patient" for which the first event was successfully detected; conversely, we define an *FN* classification to be an "event patient" for which the first event was not successfully detected.

A confounding factor in the evaluation of monitoring systems is determining which machine alerts count as "early warning" of a forthcoming event, and which alerts are merely false positive. This definition must depend on the dynamics of the system being monitored; in the assessment of patient vital signs, for example, if an alert was generated by a machine within 1 h of a clinical escalation, we might deem that the alert was predictive of the escalation; however, if the alert occurred 10 h before the escalation, we might deem it to be a false alert. Clinicians have no standard definition of "early warning," and so we define an event to have been successfully detected if an alert within some time τ prior to that event. We will consider the performance of each system as τ is varied in the interval $\tau = [0\ 60]$ min, representing the range of times that could constitute "early warning" of an event in the context of patient vital-signs monitoring. We argue that this approach allows the comparison of various integrated systems without dependence on this window length, τ.

We define a *TN* classification to be a "normal patient" for which there were no alerts generated; conversely, we define an *FP* classification to be a "normal patient" for which one or more alerts were generated; i.e., the "normal" set of patients should have had no alerts generated for them, because they were deemed to be physiologically stable throughout their connection to the monitoring system.

VI. Results

Fig. 4 shows the TP and FP results for the various analysis methods described previously, when evaluated using the large-scale methodology described in Section V. We compare the kernel estimate K_0 from Section IV-A, the two modified kernel estimates K_{w1} and K_{w2} from Section IV-B, the SVM from Section IV-C, the heuristic EWS system that was used in the hospital at the time of the study (termed EWS_a), and the "evidenced-based" EWS system proposed in [6] since adopted throughout the Oxford University Hospitals NHS Trust (termed EWS_b).

EWS_a and EWS_b, which would be applied manually, in practice, were evaluated when applied to continuous data at frequencies of 30 min and 2 h. We refer to the 2-h EWS systems as

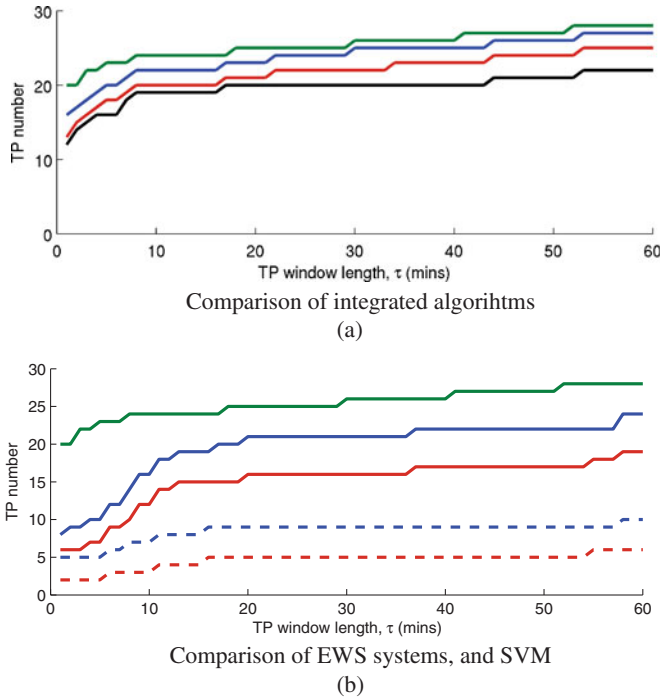

Fig. 4. TP numbers for manual and automatic patient systems, with a maximum TP = 29 event patients: (a) automatic methods SVM, K_{w2}, K_{w1}, and K_0 are shown from uppermost to lowermost, respectively (i.e., in green, blue, red, and black, respectively); and (b) SVM, and manual methods EWS_{a2}, EWS_{b2}, EWS_{a1}, and EWS_{b1} are shown from uppermost to lowermost, respectively (i.e., by green, blue solid, red solid, blue dashed, and red dashed lines, respectively).

EWS_{a1} and EWS_{b1} for the old and new EWS systems, respectively. Similarly, we refer to the 30-min EWS systems as EWS_{a2} and EWS_{b2} for the old and new EWS systems, respectively.

The SVM and kernel-based methods were both trained using data obtained from a previous clinical study, as described in Section IV. These training data were acquired from patients deemed representative of those encountered in the ED, and where studies indicate that "stable" patient physiology across high-dependence departments in the U.S. and U.K. is generally similar [6].

Examination of the results shown in Fig. 4 demonstrates that the SVM is the superior classifier in terms of identifying patient deterioration. In order of decreasing sensitivity to patient deterioration, the automatic methods K_{w2}, K_{w1}, and K_0 suggest that the discriminative power of the SVM is closely matched by the modified kernel estimate K_{w2} when those kernels furthest from the centroid of the training data are discarded. Of the automated methods, the basic kernel estimate is the least sensitive to patient deterioration in the results from the phase I data obtained from the clinical trial [19], [20].

The figure also shows a comparison of the manual methods, where it may be seen that the EWS used during the period of data acquisition applied at 30-min intervals (EWS_{a2}) is more sensitive to patient deterioration than the new EWS proposed in [6] applied at 30-min intervals (EWS_{b2}). Unsurprisingly, when these two methods are applied less frequently, at 2-h intervals (EWS_{a1} and EWS_{b1}), the sensitivity for both EWS systems decreases significantly. While these EWS systems can reach the sensitivity of automated methods, they must be performed at unrealistically frequent intervals, which could not be supported by the typical resources available in emergency care.

TABLE IV
SUMMARY OF FALSE POSITIVES FOR ALL 217 "NORMAL" PATIENTS

K_0	K_{w1}	K_{w2}	SVM
34	36	46	58
EWS_{a1}	EWS_{a2}	EWS_{b1}	EWS_{b2}
49	86	29	54

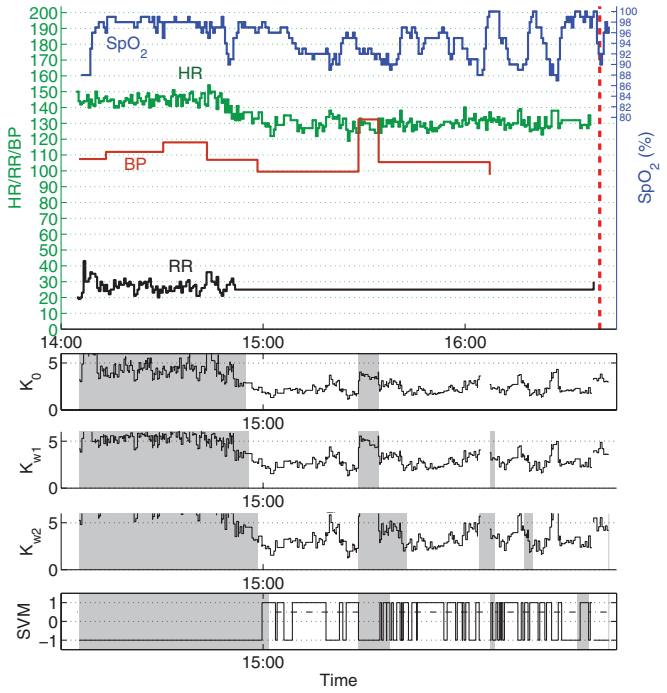

Fig. 5. Vital signs (SpO_2, HR, BP, RR from upper to lower, first plot, in blue, green, red, and black, respectively) acquired via e-health system for an example ED patient, who deteriorated and for whom care was escalated at the time shown by the vertical red dashed line. Underneath, the outputs of the automatic methods, where shaded regions indicate intervals during which an alert was raised.

Table IV shows the number of FPs for each method, obtained from examining their performance with the set of entirely "normal" patients, as described in Section V, for whom we would not expect physiological deterioration to be detected. The table shows that, while manual EWS methods can approach the sensitivity of integrated automatic methods (if performed sufficiently frequently), they generate large numbers of FP alerts. The results also demonstrate that the evidence-based EWS system EWS_b of [6] has a lower FP rate than all methods considered; this is a particular advantage for an EWS, which probably has to err on the side of being insensitive, because every alert generated by the system greatly increases the workload of the ward staff. We note that these values of FP are significantly lower than the false-positive rates from conventional bed-side monitors [7], making them attractive for use in a busy clinical environment, such as an ED.

A case study of a patient that deteriorated during phase I of the clinical trial is shown in Fig. 5, along with corresponding

outputs from the various automatic methods under consideration. We note that this patient had no temperature data available. An escalation of care occurred at the end of the period shown, following periods of erratic SpO_2 (including many desaturations to low oxygen levels), a shift in baseline HR by approximately 20 bpm, and transient hypertension (where the systolic-diastolic average BP exceeds 130 mmHg in this example). Note that the RR signal drops out in the first half of the interval shown, due to poor signal quality obtained from the measurement probes. The outputs of the automated methods demonstrate that this deterioration was detected sufficiently early to be useful to clinicians.

VII. Conclusion

This paper has described a large clinical trial, undertaken in order to 1) evaluate automatic methods for assessment of patients, based on an electronic patient record augmented with machine learning algorithms and 2) address the perceived lack of evidence in e-health research, that has been suggested as one of the primary reasons that such methods have not yet been adopted at scale. The trial that we have described is the first of its kind, and brings with it many challenges that we have had to address, and which have been described in this paper.

Integrated, automatic systems require some sort of supervised training in order to perform useful analyses, and so that resulting systems can be evaluated. While semisupervised techniques have been demonstrated to be effective in some areas [21], this is still an active research area within e-health, and remains future work. Therefore, there is a dependence on the labeling of patient data such that periods of instability can be learned. This requires teams of expert clinicians to review data from the electronic patient record, which is extremely time consuming and costly. We have described methods of reducing this workload, by requiring clinical labeling of those events that are deemed to have been "escalations" in practice; while this is not an exhaustive labeling of the data (which is impractical, for a 105-GB dataset, exceeding 2170 h of data), it has allowed us to train and evaluate machine learning techniques within the integrated healthcare system deployed in the ED.

We have framed the existing standard of care (EWS systems) as a novelty detection task, in which periods of "abnormal" physiology are scored; this allows a direct comparison with our integrated, automatic methods. Such approaches are, by nature, population-based; patient-specific algorithms that learn from individual test patients in real-time remain an area of on-going research. However, the population-based approach is already accepted as being clinically acceptable, because the EWS systems in widespread use in hospitals adopt such methods. Hence, our proposed system may be seen as a continuous, automatic version of the existing clinical standard of care, with the advantage that vital signs can be treated in a multivariate sense, using models that take into account dependence between vital signs; this is a considerable advantage over the existing standard of care, in which vital signs are treated as being independent.

Future work will involve improving the algorithms used to perform data communication and analysis, and in further building the body of evidence required before integrated patient monitoring systems can be adopted at scale. Related clinical trials in the physiological monitoring of other cohorts of patients are underway between the Institute of Biomedical Engineering, Oxford and the Oxford University Hospitals NHS Trust; these trials will adopt the methodologies developed within this paper for assessment and evaluation of evidence.

References

[1] G. Clifford and D. Clifton, "Annual review: Wireless technology in disease state management and medicine," *Annu. Rev. Med.*, vol. 63, pp. 479–492, 2012.

[2] S. Martin, G. Kelly, W. Kernohan, B. McCreight, and C. Nugent, "Smart home technologies for health and social care support," *Cochrane Database Syste. Rev.*, vol. 4, pp. 1–11, 2008.

[3] National Patient Safety Association, "Safer care for acutely ill patients: Learning from serious accidents," Tech. Rep., 2007.

[4] National Institute for Clinical Excellence, "Recognition of and response to acute illness in adults in hospital," Tech. Rep., 2007.

[5] H. Gao, A. McDonnell, D. Harrison, S. Adam, K. Daly, L. Esmonde, D. Goldhill, G. Parry, A. Rashidian, C. Subbe, and S. Harvey, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Med.*, vol. 33, no. 4, pp. 667–679, 2007.

[6] L. Tarassenko, D. Clifton, M. Pinsky, M. Hravnak, J. Woods, and P. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.

[7] C. Tsien and J. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Crit. Care Med.*, vol. 25, no. 4, pp. 614–619, 1997.

[8] A. Smith and R. Oakey, "Incidence and significance of errors in a patient "track and trigger" system during an epidemic of legionnaires' disease: Retrospective casenote analysis," *Anaesthesia*, vol. 61, pp. 222–228, 2006.

[9] C. Subbe, H. Gao, and D. Harrison, "Reproducibility of physiological track-and-trigger warning systems for identifying at-risk patients on the ward," *Intensive Care Med.*, vol. 33, pp. 619–624, 2007.

[10] D. Prytherch, G. Smith, P. Schmidt, P. 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*, vol. 70, pp. 173–178, 2006.

[11] J. Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer-Verlag, 1996.

[12] A. Hann, "Multi-parameter monitoring for early warning of patient deterioration," Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2008.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.

[14] G. Babich and O. Camps, "Weighted parzen windows for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 5, pp. 567–570, May 1996.

[15] D. Wong, I. Strachan, and L. Tarassenko, "Visualisation of high-dimensional data for very large data sets," presented at the Workshop Mach. Learn. Healthcare Appl., Helsinki, Finland, 2008.

[16] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[17] D. Tax and R. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, pp. 1191–1199, 1999.

[18] G. Smith, D. Prytherch, P. Schmidt, and P. Featherstone, "Review and performance evaluation of aggregate "track and trigger" systems," *Resuscitation*, vol. 77, pp. 170–179, 2008.

[19] D. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *J. Signal Process. Syst.*, vol. 65, pp. 371–389, 2011.

[20] S. Hugueny, D. Clifton, and L. Tarassenko, "Probabilistic patient monitoring with multivariate, multimodal extreme value theory," *Commun. Comput. Sci.*, vol. 127, pp. 199–211, 2011.

[21] O. Chapelle, B. Schölkopf, and A. Zien, Eds., Semi-Supervised Learning. Cambridge, MA: MIT Press, 2006.

Authors' photographs and biographies not available at the time of publication.