Gaussian Processes for Personalized Interpretable Volatility Metrics in the Step-Down Ward

Glen Wright Colopy, Stephen J. Roberts, and David A. Clifton

***Abstract*—Patients in a hospital step-down unit require a level of care that is between that of the intensive care unit (ICU) and that of the general ward. While many patients remain physiologically stabilized, others will suffer clinical emergencies and be readmitted to the ICU, with a subsequent high risk of mortality. Had the associated physiological deterioration been detected early, the emergency may have been less severe or avoided entirely. Current clinical monitoring is largely heuristic, requiring manual calculation of risk scores and the use of heuristic decision criteria. Technical drawbacks include ignoring the time-series dynamics of physiological measurements, and lacking patient-specificity (i.e., personalization of models to the individual patient). In this paper, we demonstrate how Gaussian process regression models can supplement current monitoring practice by providing interpretable and intuitive illustrations of erratic vital-sign volatility. These personalized volatility metrics may provide significantly advanced warning of deterioration, while minimizing the false alarms that induce so-called alarm fatigue. While many AI-based approaches to healthcare are criticized for being uninterpretable "black-box" methods, the cause of alarms generated from the proposed methods are explicitly interpretable and intuitive. We conclude that intelligent computational inference using methods such as those proposed can enhance current clinical decision making and potentially save lives.**

***Index Terms*—Precision medicine, forecasting, Gaussian processes, patient monitoring, statistical learning, time series analysis.**

I. Introduction

THE challenge to (i) identify a deteriorating hospital patient and (ii) bring this information to a clinician's attention is beset with clinical and technical challenges. In the hospital step-down unit (SDU), where a patient's intensity of care transitions between the intensive care unit (ICU) and the general ward, such alarms systems must be timely, interpretable, suitable for application to a general heterogeneous patient population.

Current proposals to address these (or similar) clinical challenges may be either heuristic (e.g., Early Warning Scores - EWS, such as the National EWS - NEWS or Modified EWS - MEWS) or empirical (i.e., a score learned and validated from data). Empirical alarm systems tend to focus on bringing a particular quantifiable facet of patient physiology to light (e.g., abnormally high or low physiological measurements, anomalous trajectories, or deranged waveform morphology).

We propose to quantify erratic vital-sign volatility as one such useful metric that may address the timeliness, transparency, and generalizability criteria described earlier. The remainder of the paper describes several methods by which to derive these volatility metrics. Empirical evaluations are given to demonstrate why these volatility metrics may usefully supplement current approaches to monitoring.

II. Clinical Need

A. The Role of Step-Down Units in Healthcare

An SDU provides care intermediate between that of an ICU and an in-patient ward. The SDU manages the recovery of stabilized acutely-ill patients after discharge from the ICU; however, the intensity of monitoring in the SDU is less than that of the ICU in accordance with patient condition. The SDU may also receive patients from the general ward who require an escalation in care [1]. Patients admitted to the SDU may enter for a variety of clinical conditions; monitoring techniques that can generalize to a heterogeneous patient population are therefore desirable for condition monitoring in the SDU.

B. Severity of Patients in the Step-Down Unit

Although SDU patients are physiologically stable in general, a significant portion of SDU patients experience a clinical emergency event, or require emergency re-admission to the ICU. Bose *et al.* [2] and Yousef *et al.* [3] determined respectively that 31% and 34% of SDU patients on a single ward experienced cardiorespiratory instability during their stay. Mortality rates were 2% and 3% respectively. Various studies across different hos-

Manuscript received May 14, 2018; revised October 8, 2018 and November 19, 2018; accepted November 21, 2018. Date of publication January 22, 2019; date of current version May 6, 2019. The work of S. J. Roberts was supported by the Royal Academy of Engineering/Man-AHL Research Chair. The work of D. A. Clifton was supported in part by a Royal Academy of Engineering Research Fellowship, in part by a Research Fellowship at Balliol College, Oxford, and in part by an EP-SRC "Grand Challenge" Fellowship. *(Corresponding author: Glen Wright Colopy.)*

G. W. Colopy was with the University of Oxford, Oxford OX1 2JD, U.K. He is now with Current Health, Edinburgh EH1 3LY, U.K. (e-mail: gwcolo@gmail.com).

S. J. Roberts and D. A. Clifton are with the , Department of Engineering ScienceUniversity of Oxford, Oxford OX1 2JD, U.K. (e-mail: sjrob@robots.ox.ac.uk; david.clifton@eng.ox.ac.uk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2019.2890823

pitals have estimated ICU readmission rates (within the same hospital stay) to be 3.9%-9% [4], 8.8% [5], and 0%-18.3% [6].

C. Severity of Readmitted ICU Patients

Readmission to the ICU has significant implications for patient outcomes: Campbell *et al.* [5] estimated mortality rates of 40.2% for ICU patients readmitted within the same hospital stay. Cooper *et al.* [4] estimated readmission mortality to be 24.7%, in contrast to 4.0% mortality of patients who were not readmitted.

These high levels of mortality motivate the use of principled methods to identify (and, ideally, predict) physiological deterioration. We propose to implement such methods via intelligent computerised inference to supplement the knowledge of clinical staff.

D. The Clinical Value of Early Warning

Not all deaths after discharge from the ICU are preventable [5]–[7] but there is evidence that earlier response can reduce mortality rates [5], [8]. Hogan *et al.* [9] estimated that 5.2% (52 of 1000) of acute hospital deaths were preventable. Donaldson *et al.* [10] identified over 2,000 preventable patient deaths over 29 months across the UK. Both [9] and [10] identified the primary cause of most of the preventable deaths to be undetected early warning signs, and a failure to act on the evidence of deterioration. Yousef *et al.* [3] reported "a mean of 6.3 hours elapsed between the onset of a clinically apparent cardiorespiratory instability and the activation of our rapid response system", with early warning ranging from 0 to 15 hours.

These statistics suggest the presence of an observable period of deterioration, prior to emergency readmission. That many such periods of deterioration were observable but not acted upon suggests that clinicians may be more apt to intervene from alarms that are interpretable than for conventional systems of alarms. We therefore wish to generate alarms whose cause may be visualized and readily interpreted, as will be the approach described in this paper.

III. Review of Early Warning Systems

A multitude of patient monitoring systems are used in the ICU, SDU, and other hospital wards. A useful distinction can be drawn between (i) heuristic methods and (ii) empirical methods which, broadly, describe the means by which clinical decision criteria are derived. The methods employed, and the physiology that is of interest, when detecting physiological deterioration varies according to clinical context. We here examine some of the various methods used for EWS systems.

A. Heuristic Early Warning Systems

Heuristic methods include the use of rule-based thresholds, for example NEWS [11], MEWS [12], and the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) I-II [13], [14] which raise an alarm if a vital sign exceeds a prespecified threshold. These thresholds are usually set according to expert clinical experience concerning a general population of stable patients. A review of these single-parameter systems can be found in [15], and of multi-parameter systems in its companion article [16]. Typical critiques of such methods include their lack of patient-specificity, arbitrary threshold values, disregarding the presence of any trends that may exist over time, and disregarding the history of vital-sign values preceding the current set of observed values.

B. Empirical Early Warning Systems

Empirical approaches to patient monitoring learn explicit relationships between physiological data (e.g., vital signs, lab-test results) and clinical outcomes of interest (e.g., mortality, emergency ICU readmission). Examples include regression-based methods such as the Simplified Acute Physiology Score (SAPS) II-III [17] and APACHE III-IV [18]. Novelty detection methods [19], such as the Visensia EWS algorithm [20], are a popular means of quantifying divergence from a class of interest.

Machine learning methods have also been proposed to assist in clinical prognosis and diagnosis. Examples include the use of support vector machines (SVMs) for sepsis classification [21] and vital-sign time-series interpolation [22], neural networks and autoencoders for gout and leukemia classification [23], and vector autoregression for causal inference on time-series. Applications of Gaussian process regression (GPR) to patient monitoring is described in the next section

IV. Gaussian Processes for Patient Monitoring

GPR is a particularly popular model for patient monitoring due to its probabilistic framework (to handle noisy measurements) [24], flexibility to represent a wide range of functional forms, and its ability to model time-series with irregularly-spaced time-stamps [23]. Particularly relevant examples include Durichen *et al.* [25] who used multi-task GPR to model the correlation between nurse-recorded observations in heart rate, respiratory rate, and blood pressure to impute missing data values. Wong *et al.* [26] used GPR to impute missing vital-sign values, as well, to calculate EWSs. Pimentel *et al.* [27] used multi-task GPs to estimate and cluster vital-sign trajectories to distinguish between deteriorating and non-deteriorating patients. Clifton *et al.* [28] used GPs with extreme value theory (within what the authors term to be extreme function theory) to perform novelty detection in physiological time-series.

We propose to use GPR to derive personalized volatility metrics. As described above, the non-parametric flexibility of GPR allows it to model a wide range of personalized time-series dynamics. The probabilistic framework of our approach allows it to explicitly model measurement noise in physiological time-series data. The Bayesian non-parametric framework allows the GP to regularize its real-time inference over a small number of interpretable hyperparameters. Finally, as shown in Figure 1, the posterior estimates of the hyperparameters of a GP can concisely and visually represent the model, thereby allowing interpretability when presented to clinicians.

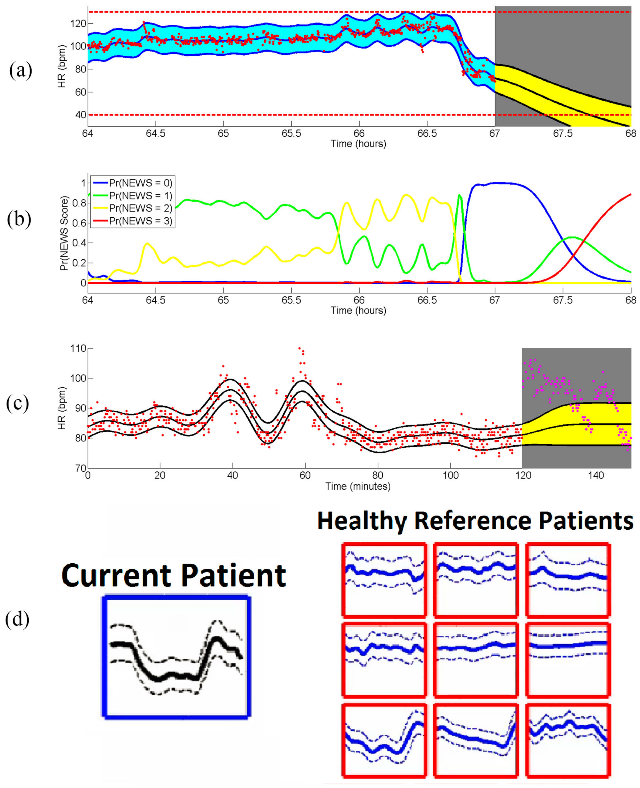

Fig. 1. Gaussian process inference for deterioration detection. In (a) a GP fits a patient's time-series and identifies the sharp downward trend in HR, and forecasts the probability that HR will fall below the 40 bpm threshold within the next hour, which would trigger a clinical alarm. On the same patient, this same GP fit and forecast can be used (b) to estimate the probability of the HR achieving a particular NEWS early warning score. On a different patient, in (c) the GP forecast may be used as a step-change detector, to quantify the deviation of a patient's vital-sign trajectory from its expected trajectory. As an alternative to comparing vital-signs at only a single point in time, in (d) a segment of the patient's time-series is compared to a dictionary of healthy patients' time-series.

A. GP-Based Approaches to Detecting Deterioration

As illustrated in Figure 1, there are many possible warnings that an intelligent computer system may bring to a clinician's attention via GP inference using vital signs.

For example, a GP can forecast the probability density of a vital-sign time-series' future values. This allows for the application of probabilistic reasoning to questions such as whether the vital-sign will 1(a) exceed the thresholds of a trigger system, or 1(b) achieve a particular EWS, for example those defined by the thresholds of the NEWS scoring system. Compared to sporadic monitoring of vital signs, the GP-based approach allows for a transparent and principled method to handle the trends and noisiness of vital-sign measurements. Both 1(a) and 1(b) use GP modelling to infer a patient's deviation from an aggregate healthy population.

Alternatively, we may wish to 1(d) compare a segment of a patient's time-series to segments from a dictionary of healthy patients. This method would incorporate both (i) magnitude of values, like current systems, as well as (ii) time-series dynamics, which are currently ignored.

Noting the homeostatis involved in physiology, in which the body seeks to return to physiological normality, we may also focus on 1(c) unusual dynamics. In particular, if we interpret abnormally-rapid increases or decreases of a particular vital sign as evidence of homeostatic reaction, then we may circumvent the need explicitly to learn examples of vital-sign abnormality. That is, we can use "step-change" detection to identify departures from physiological normality, rather than explicitly modelling physiological abnormality. (This is advantageous, because we tend to have many more examples of normality than abnormality for patients; also, physiological abnormality will differ substantially from patient to patient, and thus it can be very difficult to model in explicitly in the patient-specific setting.)

B. Selection of GP-Based Step-Change Detection

The vital signs of deteriorating patients do not always degrade gradually into abnormality. This is, perhaps, unsurprising in light of the large number of patient deteriorations that go undetected in clinical practice. When these predictable degradations do occur, they are typically too close in time to the emergency event to provide actionable early warning. This suggests that GP applications such as in Figure 1(a) and 1(b) are unlikely to garner significant gains in terms of early warning of deterioration, since there is rarely a prolonged period of evidence for extreme values before they occur.

The comparison of a current time-series to a dictionary of reference patients, as in Figure 1(d) can provide early warning gains over currently available methods. Such a method may be thought of as an expansion of the kernel density estimate (KDE) method (discussed in Section VII) into the personalized probabilistic time-series domain. Furthermore, the method is extensible (to many or few time-series or non-time-series features) and transparent in its decision criterion. Although time-series matching can easily be run in real-time, it does require significant memory to hold the reference dictionary.

This paper will focus on step-change detection methods, as illustrated in Figure 1(c). Such methods are extensible (to many or few time-series features) with minimal computational effort. This makes such a method a realistic contender for clinical implementation across a variety of clinical environments (e.g., both those with and without significant computational resources). From a clinical standpoint, such methods are transparent and interpretable for real-time inspection by the clinician. The technical details of GPR-based step-change detection (and an SVM-based comparator) will be covered in a later section.

V. Clinical Data

A data set comprising 333 adult patients was collected in the surgical-trauma SDU at the University of Pittsburgh Medical Center (UPMC) Presbyterian Hospital. The patients were recorded as phase 1 of a 3-phase trial to optimise and validate the efficacy of a KDE-based monitoring system described in [20]. Phase 1 was designed to optimise the KDE-based system's clinical alarm threshold. (As described below, we will use this KDE-based system as one of our baseline comparators against GPR-based monitoring.)

TABLE I
UNIVARIATE MET ALARM THRESHOLDS

	Lower Threshold	Upper Threshold
HR (bpm)	40	140
RR (bpm)	8	36
SpO_2(%)	85	-
SBP (mmHg)	80	200
DBP (mmHg)	-	110

Each patient's data record contains unique time-series for each of five vital-signs: heart rate (HR), respiratory rate (RR), SpO_2, systolic blood pressure (SBP), and diastolic blood pressure (DBP), acquired by Phillips bedside monitors. The time-series of each vital-sign comprised (i) vital-sign measurements and (ii) their associated time-stamps. Patients' vital-signs were recorded continuously, with individual patient records lasting from less-than an hour to several weeks on ward. HR, RR, and SpO_2 were acquired at approximately $\frac{1}{5}$ Hz. SBP and DBP were recorded approximately once every 30 minutes.

112 clinically-validated emergency events, called C”-events, occurred in 59 patient's vital-sign time-series. These 59 patients are called C”-patients. Each patient record includes the time-stamp, duration, and primary cause of any C”-event. C”-events are defined as a single vital-sign's prolonged non-artefactual exceedance of the thresholds defined in Table I. Non-artefactual exceedances of these thresholds warrant emergency medical intervention. Identical or nearly-identical criteria have been used in numerous other studies at the UPMC SDU to define “cardiorespiratory instability”, for example in [1]–[3], [29]. The presence of 112 emergency C”-events when, in practice, only 7 MET calls were made for abnormal vital-signs supports the understanding that continuous monitoring can add value to the intermittent observation of nursing staff.

VI. Experimental Design

A. Data Selection

For the purpose of detecting deterioration, we will focus on a patient's first C”-event since the vital-signs subsequent to that first C”-event may be affected by clinical intervention. (Similar reasoning is found, e.g., when developing the APACHE IV system [18], which use patient exclusion-criteria to avoid the confounding affects of previous emergency interventions.)

Acknowledging the small number of C”-patients, we took all 59 C”-patients and 89 non-C”-patients (selected at random) to test the efficacy of early deterioration detection. The remaining $333 - 59 - 89 = 185$ patients were used as a training set to (i) learn how to parametrize GPR and SVM models fit to HR, RR, and SpO_2 time-series and (ii) train the KDE baseline comparator.

B. Early Warning Performance Metric

Each of the comparator methods is evaluated according to its trade-off between two performance metrics: (i) the false positive alarm rate (FPR), and (ii) the time of early warning (TEW).

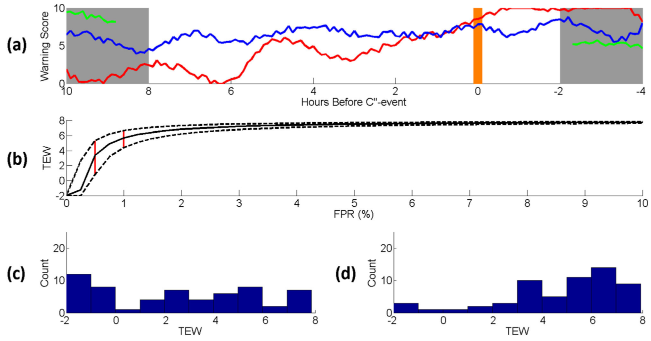

Fig. 2. The construction of a TEW vs. FPR plot. In (a) the early warning scores of three patients (red, blue, and green) are shown near their respective C”-event, at time 0 (orange). Plausible alarms are considered if they occur between 8 before until 2 hours after the event (white), and not considered if they occur outside of this time period (grey). In (b), for any alarm sensitivity we have a corresponding false-alarm rate in the non-C”-patient set (which we aggregate into a single proportion across all 89 non-C”-patients) and the TEW across the 59 C”-patient. To visualise the dispersion of TEW values, we plot the 33, 50, and 67 percentiles of the 59 C”-patients. The TEW distribution at two distinct FPR values is shown in (b) in red, and the constituent TEW values are plotted in (c) and (d), illustrating how, for each patient, TEW increases monotonically with FPR.

The trade-off between TEW and FPR closely resembles the familiar ROC curves. However, the TEW vs. FPR performance metric incorporates both (i) the clinical ambiguity of a patient's time-series prior to deterioration, as well as (ii) the time-value of early alarms. The calculation of a TEW vs. FPR plot is described in Figure 2: Alarms on non-C”-patients are false positive alarms. For a generic EWS calculated over time, in 2(a), true positive alarms are those alarms on C”-patients falling in the time period of 8 hours before until 2 hours after the C”-event (in orange). A patient's TEW is the time between a C”-patient's first true alarm (within this window) and his first C”-event (in orange). Alarms prior to 8 hours before, or following 2 hours after the C”-event (in the greyed-out region) are not considered due to their ambiguous status. That is, it is less certain that an alarm in this region is specific to the abnormal physiology related to *this* C”-event.

A false negative would be the failure of the EWS to escalate sufficiently to surpass an alarm threshold within the alarmable-window. The TEW of such cases is censored at -2 hours, which is 2 hours after the C”-event at 0 hours. This is the worst possible TEW result. The desired TEW vs. FPR plot of 2(b) is achieved by modulating the threshold required to trigger, which, in turn, modulates our alarm sensitivity in the window of 2(a), but at the cost of more frequent false positives among non-C” patients.

At a given alarm threshold, each patient will differ in the TEW due to differing personal physiology in the period surrounding their C”-event. We are therefore interested in the the distribution of TEWs for all 59 C”-patients at any particular FPR. To visualize this, we plot the 33, 50, and 67-percentiles of the TEW distribution at each FPR. For example, at two different FPR values, marked in red in 2(b), we can see that the span of TEW quantiles differ since they are drawn from the 59 individual-patient's TEWs, shown in 2(c) and 2(d). We are interested in the distribution of TEW because it informs impor-

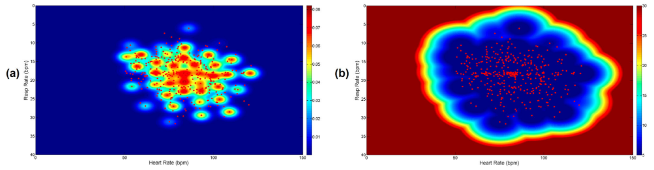

Fig. 3. KDE-based model of patient vital-sign abnormality. The displayed figures show only HR and RR, however, the KDE model was fit across all 5 recorded vital-signs: HR, RR, SpO$_2$, SBP, and DBP. The 185-patient training set comprises millions of 5D vital-sign measurements. For practical implementation, these data are reduced to 400 k-means centroids (●). In (a) the KDE fits a joint probability density to the 400 centroids (●). This produces high-likelihood regions were the training data is present, and low-likelihood regions elsewhere. In (b) the negative log-likelihood of the KDE creates a novelty score which is high when data are far from those data seen in the training set. Since the training set is comprised of non-C”-patients, we assume that deviation from these points may indicate deterioration.

tant clinical considerations, such as worst-case performance on the hardest patient cases. Such patients are of special interest to machine-monitoring applications, since they have the greatest potential to benefit, compared to easy-to-identify deteriorating patients.

VII. Deterioration Detection Methods

A. Baseline Comparator: KDE Method

There are many methods from which to select a baseline comparator. To represent the technical state-of-the-art in empirical patient monitoring, we select a KDE-based novelty detection algorithm. A further 6 methods, based on current heuristic practice were also tested. These were single-parameter trigger systems for each of the 5 vital-signs, along with a NEWS-based algorithm. However, the TEW vs. FPR results for each of these 6 methods were inferior to that of the KDE and are not shown.

A KDE-based model of patient normality reasonably represents the current technical state-of-the-art in empirical methods in use today, particularly for the UPMC data set under consideration:

The UPMC data set was first collected in order to train and evaluate such a model in 2008. As a testament to its success, KDE-based monitoring was retained in the UPMC SDU after the conclusion of the study, and continues to generate publications using UPMC SDU data. The KDE-based model, commercially known as Viscensia, was FDA-approved in 2012. Using the Visensia system in a single-site prospective study (the same UPMC SDU as this study), Hravnak *et al.* [29] found that the time period of using the Visensia system had a statistically-significant decrease in the number and duration of cardiorespiratory instabilities per admission, compared to the time period prior to using the system. Mortality decreased from 2% (before Visensia) to 1% (after using Visensia), but a statistical comparison was not made.

In Figure 3(a), the KDE models the joint distribution of the vital-signs from a “healthy” patient group (the training set’s 185 non-C”-patients described earlier). In Figure 3(b) the novelty of a new measurement is quantified by the negative log-likelihood of the new measurement with respect to the KDE. The KDE’s warning score is used any time in which at least 3 of the 5 vital-signs are available.

Notably, the KDE is an IID model of vital-signs: Vital-sign abnormality is only a function of their current values and not of their time-series dynamics.

B. GP Model

Ebden [30] provides a concise introduction to GPs for regression and classification. An overview of Gaussian process covariance functions can be found in “The Kernel Cookbook” by David Duvenaud [31], with a more in-depth coverage throughout multiple chapters of his thesis [32].

The GP extends the multivariate Gaussian (of pre-defined dimensionality n) to an infinite-dimensional stochastic process. We define the GP to be a stochastic process for which any finite subset of points, along a domain t, follows an MVN distribution. This MVN has both a mean vector, $\mathbf{m}$, and covariance matrix, $\mathbf{K}$ to describe any observed data points.

To populate the elements $\mathbf{m}$ and $\mathbf{K}$ for *any* finite subset (conditional on specifying all t), we replace the mean *vector*, $\mathbf{m}$, with a mean *function* $\mu(t)$. That is, mean of the Gaussian at point t is $\mu(t)$.

Similarly, the covariance of any two points y_i and y_j, located at points t_i and t_j, is defined by covariance *function* $k(t_i, t_j) = \mathrm{Cov}[y_i, y_j]$. This function $k(t_i, t_j)$, which will be described in detail below, is positive semi-definite and typically decreases as t_i and t_j separate in distance. Note that the covariance function k takes the *location* (time) of the random variables y_i and y_j as arguments, not y_i and y_j themselves.

We return to our initial definition of a GP as “a stochastic process for which any finite subset of points, along a domain t, follows a multivariate Gaussian”. Given this specification of $Y(t) \sim \mathrm{GP}(\mu(t), k(t, t'))$, then for any finite vector $\mathbf{t} = [t_1, ..., t_n]$ we have a random vector $\mathbf{Y}(\mathbf{t}) = [y_{t_i}, ..., y_{t_n}]$, with the now-familiar mean vector $\mathbf{m} = \mathrm{E}[\mathbf{Y}(\mathbf{t})] = [\mu(t_i), ..., \mu(t_n)]$.

The covariance matrix is

$$\mathbf{K} = \begin{bmatrix} k(t_1, t_1) & \cdots & k(t_1, t_n) \\ \vdots & \ddots & \vdots \\ k(t_n, t_1) & \cdots & k(t_n, t_n) \end{bmatrix}.$$

While GPs do not necessarily specify a functional form over $y(t)$, *a priori*, functional characteristics may be made implicit via the *a priori*-specified mean and covariance functions, which, typically, are parametric. Selection of the covariance function $k(t, t')$, typically receives the lion’s share of attention for GP model selection. This may be sensible, given that, while the prior mean may be “washed out” where data is observed, the implicit effect of the covariance function will always influence the posterior estimate of $y(t)$.

Collating the hyperparameters of the mean and covariance functions into a single vector, $\boldsymbol{\theta}$, we can infer appropriate values of $\boldsymbol{\theta}$ though the posterior log marginal likelihood (LML)

$$\log p(\mathbf{y}|\theta) = -\frac{1}{2}(\mathbf{y}-\mathbf{m})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{m}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi), \quad (1)$$

where $\boldsymbol{\theta}$ influences $\log p(\mathbf{y})$ via the the mean function and covariance function which determine the values of the mean vector, $\boldsymbol{m}$, and the covariance matrix, $\boldsymbol{\Sigma}$.

This crucial inferential step typically either (i) optimises the LML (e.g. via gradient ascent), or (ii) integrates across the LML (e.g. via Markov Chain Monte Carlo as in [33]). Access to parallel computation can assist MCMC via (i) running multiple parallel MCMC chains, or, alternatively, (ii) parallel processing of the proposals for the likelihood-ratio step between points [34].

Due to the real-time application of patient monitoring algorithms, it is desirable to reduce computational burden, where possible. One way to achieve this is to frame the GP as an equivalent state-space model, which requires less-burdensome inference. [35] demonstrates how these requirements can be reduced to $O(m^3 n)$ for covariance inversion and $O(m^2 n)$ for memory, by reformulating $k(t, t')$ as an m^{th}-order, scalar, linear time-invariant stochastic differential equation. This computation-saving manipulation does not affect interpretation of the GP models subsequently described. For this reason we used a state-space approximation of the GP, with hyperparameters fit via optimisation of Equation 1.

C. GP-Based Step-Change Detection

Figure 4 illustrates how a GP-based step-change algorithm sequentially fits and forecasts the future distribution of 4(a) BR, 4(b) SpO2, and 4(c,d) HR values. When the future values are consistent with the prediction, as in 4(d), the corresponding LML, as shown in 4(h) will be high. However if the future values are not consistent with the forecast distribution, as in 4(a-c), then the corresponding LML will be lower, as in 4(g). It is reasonable to present the various LML values within a forecast window via a summarising statistic, such as the mean. This also mitigates the affect of occasional outlying measurements. Since conventional alarm scores are high in the presence of abnormal physiology, we will measure step-change warning scores in terms of negative log marginal likelihood (NLML).

Since the step-change detector forecasts over a time-window that contains one NLML value per vital-sign measurement with the window (as seen in Figure 4(g) and 4(h)), the step-change detector's tunable parameters include (i) the metric to summarize NLML values within the forecast window, and (ii) the time-length of the window. To reduce research degrees of freedom, we will relegate our choices to the simplest and most obvious choices for these tunable parameters. We will examine the mean NLML of 1-minute forecast windows.

Using the 185 training set non-C"-patients, we can assess forecasting robustness of various combinations of covariance functions with regularising priors for our GP model. Since our goal is step-change detection of large forecast errors, we select kernel-prior combinations that maximise the lowest 1%-10% of forecast LML for each patient. A patient-by-patient evaluation of these criteria is helpful to avoid Simpson's paradox, by which a model with inferior performance across all patients individually may appear to have superior performance on aggregate. Using these criteria, we selected an additive two kernel Matérn 5/2 covariance function, plus white noise, to model HR time-series. We selected a single-kernel Matérn 3/2 covariance function, plus white noise, to model RR and SpO$_2$ time-series. The Matérn 3/2 and 5/2 covariance functions encode that the time-series are once- and twice-differentiable, respectively, which is more realistic for erratic vital-sign time-series than the smooth (infinitely-differentiable) radial basis function (RBF) kernel.

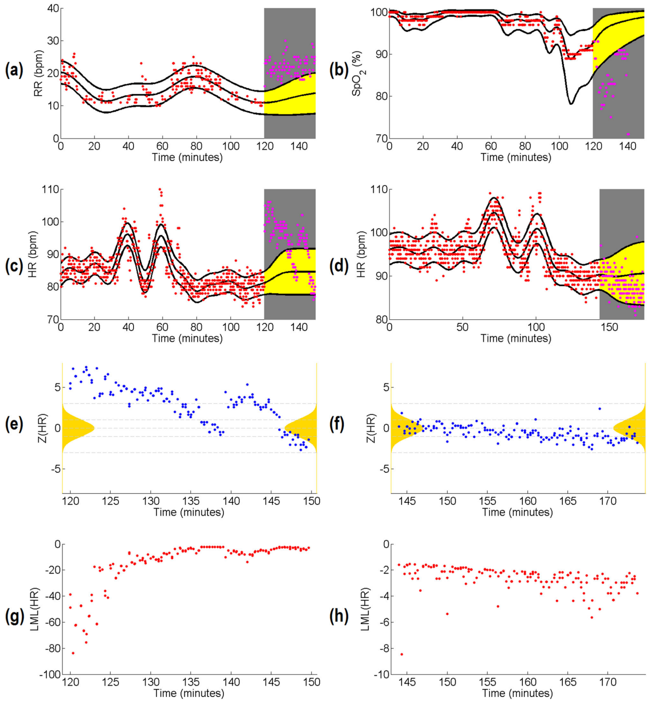

Fig. 4. Illustration of a GP-identified step-change in (a) RR, (b) SpO$_2$, and (c) HR. In (d) the HR time-series shows no step change. GPs are fit to observed data (●) and forecast the distribution of unseen data in the future (●). In (c) the asymmetric GPR confidence bounds on SpO$_2$ are due to a $\log(101 - \mathrm{SpO}_2)$ transformation, to minimize the proportion of the posterior distribution greater than 100%, which is physically impossible. Since the marginal Gaussian distribution changes through the forecast window, the z-scores of the forecast-window HR from (c) and (d) are shown in (e) and (f), respectively. In (e) and (f) a $N(0, 1)$ reference distribution (gold) is provided with (- -) denoting mean ± 0, 1, and 3 standard deviations. The forecast LML of each measurement in the forecast windows of (c) and (d) are shown in (g) and (h), respectively. The LML measurements within a specific time window may be summarized, e.g., by the mean or another statistic. Step-change warning scores are the negative of these LML values, NLML.

In Figure 5, we show how step-change metrics may be used as a continuously-monitored warning score in the same manner as the NEWS or the KDE methods described above. It is noteworthy that unlike the NEWS or KDE method, which tend to be persistent, the step-change detector (by its nature) produces transient warning metrics. This can be seen in 5(b) where step-change NLML is escalated for only a short period of time, e.g. for HR step-change NLML (red) near hours 71, 73, and

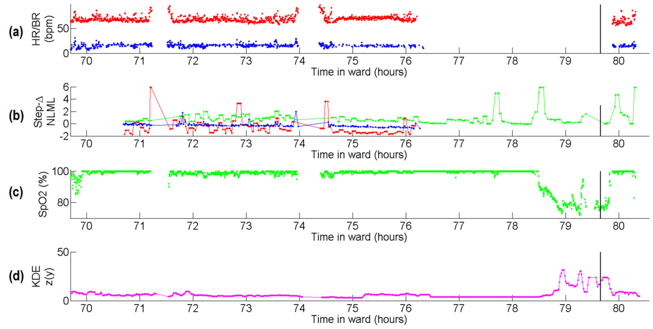

Fig. 5. Time-series of a patient's vital-signs and early warning scores leading up to an emergency event near hour 80 (black vertical line). The patient's vitals in (a) HR (●) and RR (●), and (c) SpO_2 (●) each display various step-change dynamics. SBP and DBP are not shown. In (d), the 5-vital KDE score of the patient vitals has been calculated. It is seen here to escalate at the approach of the emergency event. In (b) the step-change detection novelty score for each individual vital sign is shown for HR (●), RR (●), and SpO_2 (●). No step-change score is available in the absence of measurements.

74.5 and for SpO_2 step-change NLML (green) near hours 77.5, 78.5, and 80. This is to be expected since the flexibility of the GP allows it to quickly adjust to the new (volatile) data and resume precise forecasting. In contrast, the KDE-based warning score in 5(d) is persistent when elevated, e.g., between hours 79 and 80.

Since the warning scores produced by step-change detection are transient instead of persistent, the warnings scores of a step-change detector are apt to be missed if monitored only sporadically by clinical staff. In this, step-change detection would only be appropriate in a continuous computer-assisted monitoring setting since the score indicative of deterioration would need to be recorded and brought to the attention of clinical staff.

D. Baseline Comparator: SVM-Based Step-Change Detection

SVMs provide an alternative method to the GP for non-linear time-series regression. In contrast to the GP, which evaluates the suitability of parameters via the posterior likelihood function, the SVM is typically parametrized via cross validation techniques. Previous work in SVMs for time-series applications has noted their difficulty in forecasting (compared to alternative time-series tasks such as interpolation, as in [22]). Like the GPR method, the SVM can learn time-series trends and forecast. Since SVM is non-probabilistic, a forecast's performance will be evaluated by root mean squared-error (RMSE) and, in turn, will serve as the metric for SVM-based step-change.

A common critique of machine learning literature is that baseline methods are less rigorously tuned to the desired application, compared to the proposed method. To address this, multiple cross-validation methods (CVMs) were compared to find the best method of parametrizing the SVM's RBF kernel for forecasting tasks. The three contending cross-validation methods (abbreviated to CVM-1, CVM-2, and CVM-3, respectively) were

- *CVM-1:* traditional k-fold cross-validation, with training data randomly allocated to each fold,
- *CVM-2:* "windowed" k-fold cross validation, with the training data partitioned into k-contiguous time windows, and
- *CVM-3:* forecast-only validation, where the k-folds for cross validation are k sequential training and forecast windows at the end of the data set.

An illustration of each CVM is provided in the appendix of supplemental material. We will note that with continuously acquired vital-sign data, CVM-1 selects parameters according to strong interpolation performance, CVM-3 for strong forecast performance, and CVM-2 for a combination of forecast, backcast, and interpolation performance. To validate our choice in CVM, the 185 training set patients were divided into 45 patients to learn vital-sign specific (i) ranges for the SVM parameters and (ii) the preferred CVM. The remaining 140 patients were used to compare SVM to GPR forecast (the GP's covariance function having also been selected by performance on the same 45 patients). For each vital-sign, CVM-2 was found to have marginally better forecast RMSE than CVM-3, and both outperformed CVM-1.

VIII. Results

A. GPR vs. SVM for Step-Change Detection

140 patients were used to compare SVM to GPR forecast. In Figure 6, patient-specific forecasting of GPR and SVM were compared. Results indicated that GPR provided more accurate forecast performances more frequently for most patients. Accordingly, for our step-change detection method, it is sensible to select GPR to derive the metric since it (i) is more successful learning and forecasting patient dynamics 6(a-c), (ii) provides probabilistic estimates (of which SVMs are incapable), and (iii) seems to run faster 6(d-f). For thoroughness, the early warning performance of the SVM was also evaluated, and this is included in the appendix.

B. TEW vs. FPR Results

We wish to emphasize how personalized volatility metrics may helpfully supplement more common extreme-value oriented EWS. Accordingly, the results section will focus on the comparison of the GPR-derived personalized volatility metrics against the population-wide, extreme-value oriented KDE method. A performance comparison for the GPR versus SVM-based methods can be found in the supplemental materials.

Figure 7 shows the TEW vs. FPR performance of the 5-vital-sign KDE against the performance of GPR-based step-change detection models over 1, 2, and 3 vital-signs. We focus on early warning performance in the 0% to 2.5% FPR range due to the desirability to mitigate alarm fatigue. The strong performance of step-change detection methods suggest that step-change detection, or similar methods may provide a useful supplement to the current state-of-the-art in patient monitoring.

Among the univariate step-change detectors, 7(a) HR and 7(b) RR both outperform the KDE, despite each using only a single vital-sign compared to the KDE's five vital-signs. The 7(c) step-change on SpO_2 is nearly the same as the KDE, except

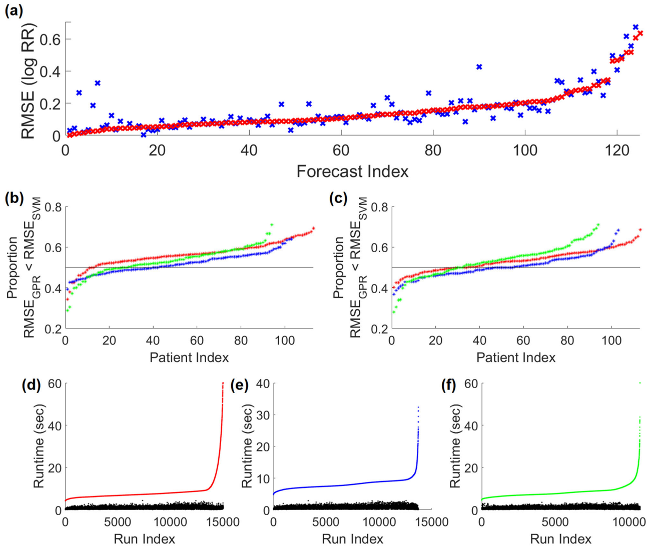

Fig. 6. GPR vs SVM forecast performance. SVMs cannot produce probabilistic forecasts therefore forecast performance was compared by RMSE. In (a) a single patient's forecast RMSE on RR is shown for GPR (×) and SVM (×). Each method provided a superior forecast 50% of the time, however the SVM had more instances where the forecast was substantially worse. To collate the results in (a) across all 140 patients (while accounting for per-patient and per-forecast confounding) forecast performance at (b) 1-minute and (c) 5-minute forecasts are shown for HR (●), RR (●), and SpO2 (●). The y-axis shows the proportion of forecasts for which GPR was more accurate, and results are stratified by patient. A value greater than 0.5 indicates that GPR was better for most forecasts for that patient. Patients with less than 2 hours of a vital-sign were not included in the analysis. Subplots (d), (e), and (f) compare the runtimes to parametrize for each forecast, with SVM (●)(●)(●) and GPR (●). These runtime differences should be seen as rough comparisons, given that neither method was optimally coded, e.g., had they been programmed in C.

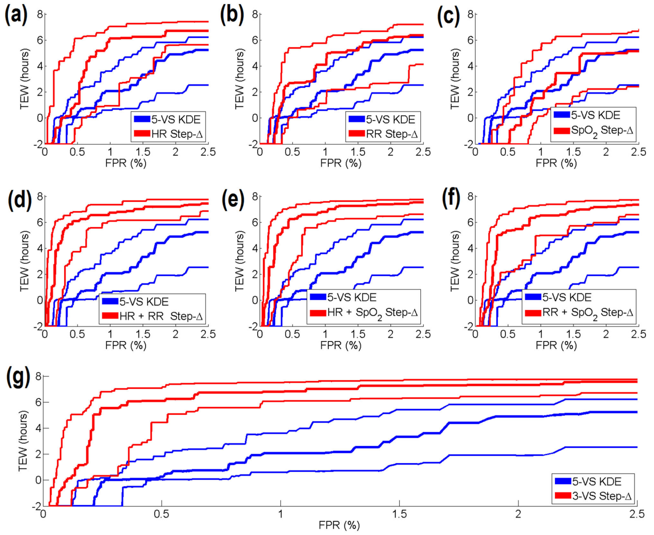

Fig. 7. TEW vs. FPR plots of KDE baseline comparator (_) next to univariate step-change detectors (_) on (a) HR, (b) RR and (c) Spo2; Bivariate step-change detectors (_) on (d) HR+RR, (e) HR+SpO2, and (f) RR+SpO2; Trivariate step-change detector (_) on (g) HR+RR+SpO2 (red). Lines represent 33, 50 and 67 percentiles of TEW at the respective FPR. Each FPR value is calculated from 600 random time points for each of the 89 non-C"-patients in the test set, i.e. $600 \times 89 = 53,400$ total time points. Results were nearly identical when FPR was calculated using minutely predictions from the first 72 hours of monitoring.

for the KDE's superior performance in the 0% to 0.5% FPR range. An explanation for the superior performance of univariate step-change detection over the KDE will be discussed shortly, however this outcome is positive in several respects:

First, the step-change method is not dependent on the magnitudes of vital-signs. The information contained within the step-change detection metrics is significantly different from KDE or NEWS-based scores. This suggests that such a metric may be a useful supplement to current monitoring. Second, the strong monitoring results on only a single vital-sign suggests that a vast number of clinical variables may not be necessary to achieve optimal or near-optimal monitoring performance. This may be useful, i.e. for intelligent monitoring in resource-constrained settings where fewer monitoring modalities are available. However, both of these benefits are dependent on the interpretablity of the step-change detector when it brings warning scores to the attention to clinical staff.

A final note on the univariate step-change methods is that, within the FPR range of 0% to 0.5%, the KDE frequently outperforms the step-change detector in median and 33-percentile performance. This region roughly corresponds to warnings about 1 hour or less prior to the emergency event. Contributing factors to this are several-fold. One factor is the missingness in individual vital-channels near the emergency event, as illustrated for HR and RR in Figure 5(a). This puts univariate methods at a disadvantage when the particular vital-sign under consideration is missing. However, this would also put the KDE method at a disadvantage when 3 or more of the five-total vital signs (that is, any 3 between HR, RR, SpO_2, SBP, and DBP) are missing since the KDE would no longer produce a score whereas univariate monitoring system on either of the other two vital-signs would continue.

More important than the issue of missingness, is the definition of C"-events: C"-events, by definition, have highly-abnormal vital-sign values. This means that the KDE-based warning score is nearly-guaranteed to be high in proximity to emergency events because at least one vital-sign will be sufficiently abnormal to contribute to a high warning score. In contrast, the emergency events are not defined according to vital-sign volatility, on which step-change methods are based. Step-changes are, therefore, not guaranteed to occur at the time of event.

When KDE performance is compared to 5(d,e,f) bivariate and 5(g) trivariate step-change detection, there are fewer caveats to the results, since results are almost uniformly superior. While the improved performance itself may be unsurprising (having already seen that univariate methods themselves produced superior results) the magnitude of difference in performance motivates further inspection to understand why this may be.

IX. Discussion

There are several possible factors that contribute to the difference in performance between step-change based monitoring and KDE-based monitoring. The following factors were determined to be particularly important:

1) The physiology at the time of emergency differs from the physiology preceding the emergency.

2) Personalisation of monitoring improves FPR, or, conversely, population-based risk assessment necessitates high FPR.

We will illustrate each of these two items below to add transparency to why GP-based volatility metrics (via step-change detection) may to a useful supplement to current monitoring methods, which may only alarm when a patient's vital-signs are univariately or jointly extreme.

A. The Physiology at the Time of Emergency Differs From the Physiology Preceding the Emergency

Vital-sign measurements tend to differ between (i) the time of the emergency event, and (ii) the time preceding an emergency events. In other words, the majority of patients' abnormal values are not preceded by a long period of gradual decent towards abnormal values. Instead vital-signs are more frequently characterised by “shocks” to one or more vital-signs, which are quickly corrected by homeostatic mechanisms. This can be verified by examining plots of patient time-series in the the time period surrounding emergency events. For example, in Figure 5, the emergency event triggered by low-SpO_2 near hour 80, was only preceded by about 1 hour of decreasing SpO_2. Otherwise, there was nothing abnormal in the absolute values of HR, RR, or SpO_2. This means that NEWS or KDE-based warning scores would only begin to increase (compared to the general population) in the final hour before deterioration. Therefore, for the KDE to achieve a TEW greater than 1 hour, FPR would need to increase substantially. In contrast, step-change detection identifies at least 3 prominent HR step-changes, 2 prominent RR step-changes, and two prominent SpO_2 step-changes, in addition to several smaller step-change episodes. This results in approximately 1 step-change episode per hour. It is unclear whether the KDE warning score was artificially depressed between hours 76 to 80 due to the missingness of HR and RR, however, the KDE does not appear to identify any episodes for which alarms should be raised between hours 70 to 76, in which data for all vital signs are available.

In summary, magnitude-based EWS methods (e.g., KDE, NEWS) have both advantages and disadvantages:

An advantage is that alarms occur for vital-sign measurements that are univariately or jointly abnormal (high or low). Since the emergency events of this data set were also annotated as such for their abnormally high or low values, the KDE can alarm reliably *at* the time of the annotated emergency events. In contrast, the step-change detector has no such guarantee that the requisite physiology (a step-change) will occur at the time of the event.

However the TEW metric places a premium on *advanced* warning, since earlier warning facilitates preventative clinical intervention. If vital-signs are neither abnormally high or low far in advance of the emergency event, then magnitude-based monitoring methods are disadvantaged.

B. Personalization Improves FPR

Like the heuristic NEWS approach, the KDE-based EWS suffers from attempting to use a single model to describe all patients. This means that it only matters (i) whether a patient exhibits physiology abnormal to the entire population, not (ii) whether the patient exhibits physiology abnormal to himself. Age- and sex-based early warning scores attempt to adjust for obvious confounding demographic information, however, the intra-group variability is likely to be substantial, given that inter-patient and intra-patient variability is high. Inter-patient variability is high, even when stratified by C”-status.

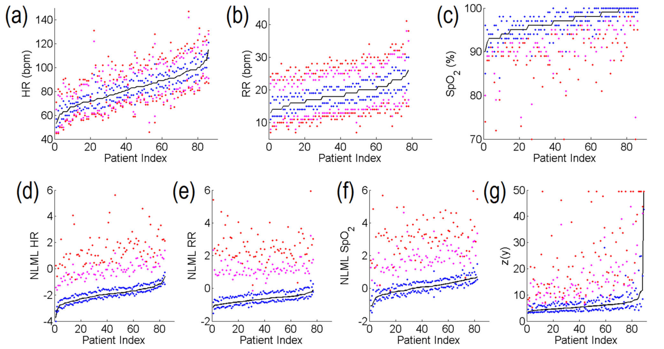

Fig. 8. Non-C”-patient intra- and inter-patient variability in (a) HR, (b) RR, (c) SpO_2, contrasted to (g) KDE warning scores, and step-change NLML in (d) HR, (e) RR, and (f) SpO_2. For each patient the following percentiles are marked: median (-), 25 and 50 (●), 5 and 95 (●), and 2.5 and 97.5 (●). Patient indices are ordered in ascending median value for each metric.

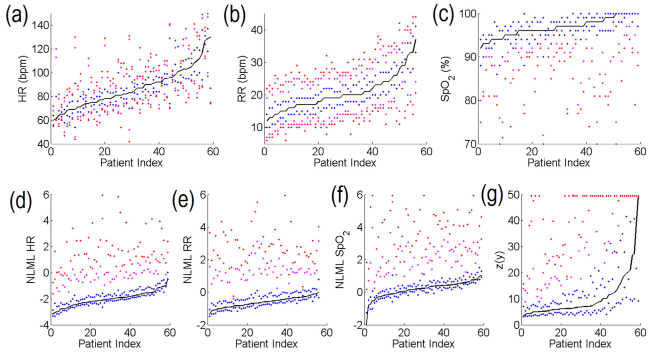

Fig. 9. C”-patient intra- and inter-patient variability in (a) HR, (b) RR, (c) SpO_2, contrasted to (g) KDE warning scores, and step-change NLML in (d) HR, (e) RR, and (f) SpO_2. For each patient the following percentiles are marked: median (-), 25 and 50 (●), 5 and 95 (●), and 2.5 and 97.5 (●). Patient indices are ordered in ascending median value for each metric.

To illustrate this, the inter- and intra-patient variability of vital-signs, KDE warning scores, and step-change warning scores are plotted in Figure 8 for 89 non-C”-patients, and Figure 9 for 59 C”-patients.

Inspection of the intra-patient ranges in 8(a-c) and 9(a-c) show patient vital-signs occur in completely different dynamic ranges. In extreme cases the upper 2.5 percent of one patient's vital-signs may be less-than the lower 2.5 percent of another patients vital-signs. This, effectively, removes the possibility of alarming on low values for patients with high-valued vital-signs or alarming on high values for patients with low-valued vital-signs. More importantly with respect to the FPR metric, this means for an “average” patient in the middle of this range could not achieve any type of alarm without the monitoring system inducing a nearly-constant state of alarm in other patients.

In contrast, the dynamic range of step-change NLML for HR, RR, and (to a lesser-extent) SpO_2 is highly consistent and compact across all patients. Fewer patients have dynamic ranges in step-change NLML that overlap with the extreme (alarm-generative) values of other patients. For example, the 2.5 quantile values of HR NLML in C”-patients in 9(d) are more extreme than those of non-C”-patients in 8(d). This indicates that C”-patient have more pronounced HR step-changes than non-C” patients, as we would expect. More important for maintaining a low FPR, though, is that no patient has average values that would fall within the high NLML range. This leads to a two-fold conclusion: (i) a single threshold can delineate between low-NLML and high-NLML step-changes across all patients, and (ii) highest NLML step-changes occur more frequently in C”-patients than they do in non-C”-patients. It is not surprising then, that step-change detection demonstrates a successful trade-off between TEW and FPR.

The KDE method falls in between thresholding on raw vital-signs and the step-change method. Comparing KDE warning scores between the 89 non-C” patients in 8(g) and the 59 C”-patients in 9(g), it is immediately apparent that C”-patients experience a much higher rate of high warning scores than non-C”-patients. The difference is much greater, even, than the difference between NLML step change values of C” and non-C”-patients. This is expected, since the KDE-based novelty score is persistent in the presence of abnormality. However it can be seen that the 95 (●) and 97.5 (●) percentiles of KDE novelty are still highly-intermingled with more central values on an inter-patient basis. To avoid the high FPR that make an early warning system infeasible in clinical practice, the high and low percentiles must be clearly delineated. Otherwise the system will generate a near-constant rate of alarms in a subset of the non-C”-patients, which escalates average FPR.

X. Conclusion

Current practice in vital-sign EWSs focuses on identifying patients with extreme vital-sign measurement values. This is sensible, given that emergency events themselves are typified by vital-signs with extreme values. However, clinical reasoning over time-series offers many ways to identify early physiological indications which current practice tends to ignore. Such reasoning requires an intelligent system, both to compute warning metrics and inform the clinician.

The probabilistic representation of GP modelling allows us to incorporate a richer range of physiological features beyond magnitude, such as volatility and uncertainty in the patient’s current and future vital-sign values. The described methods can provide useful clinical insight beyond or paired with current practice, even when using only a single vital-sign. This is helpful in implementation, since the benefits of a step-change detector may be realised with only a single vital-sign, whereas other empirical monitoring systems may require a diffuse range of vital-signs in order to provide an EWS. Both the baseline KDE method and step-change detection (via GPR or SVM) have their advantages, however the KDE and related methods are limited due to (i) their population-based approach to modelling patient abnormality, which increases FPR, and (ii) the fact that many patients do not exhibit extreme-valued vital-signs until shortly before the emergency event, which decreases the TEW provided by such methods.

The described step-change methods may be applied to a variety of settings including those with constrained computational resources or with a single vital-sign under consideration. Importantly, the described step-change detection models can be run in real-time, even with minimal computational resources, and present salient interpretable physiology to clinical staff to explain the cause of the alarm. By displaying the vital-sign step-change that precipitated the alarm, the GP-based step-change detector is far from a black-box algorithm, it is an intelligent monitoring system to supplement and enhance a clinician’s understanding of deterioration.

XI. Future Work

Reiterating our conclusion that personalized volatility metrics are promising candidates with which to supplement (not replace) current clinical metrics, we would suggest several steps moving forward. First, a clinical review of the volatility metrics, and their associated interpretable figures, as in Figure 4(a-c), would be helpful to identify which of the step-changes clinicians would like to be brought to their attention. Currently, we have not identified precisely which step-change alarms would have elicited clinical intervention, and therefore our method’s effect on clinical outcome requires prospective validation. Clinical feedback would also help establish strategies to combine these personalized volatility metrics with the more traditional extreme-value-based EWSs.

In this paper, we constrained our search to a generally-applicable warning score, due to the heterogeneous patient population in the UPMC set. However there are many critical care wards in which patients are suffering from a more homogeneous set of clinical ailments. In this case, it would be helpful to bring in a larger number of condition-specific physiological parameters, to better tailor treatment to the patient population. In these settings waveform data (which was unavailable in the UPMC set) would also be desireable to (i) better evaluate measurement noise and (ii) identify informative arrhythmias with disease-specific connotations.

References

[1] M. Hravnak, L. Edwards, A. Clontz, C. Valenta, M. DeVita, and M. Pinsky, “Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system,” *Arch. Internal Med.*, vol. 168, no. 12, pp. 1300–1308, 2008.

[2] E. L. Bose *et al.*, “Cardiorespiratory instability in monitored step-down unit patients: Using cluster analysis to identify patterns of change,” *J. Clin. Monitoring Comput.*, vol. 32, no. 1, pp. 117–126, Feb. 2018.

[3] K. Yousef, M. R. Pinsky, M. A. DeVita, S. Sereika, and M. Hravnak, “Characteristics of patients with cardiorespiratory instability in a step-down unit,” *Amer. J. Crit. Care*, vol. 21, no. 5, pp. 344–350, 2012.

[4] G. S. Cooper, C. A. Sirio, A. J. Rotondi, L. B. Shepardson, and G. E. Rosenthal, “Are readmissions to the intensive care unit a useful measure of hospital performance?” *Med. Care*, vol. 37, no. 4, pp. 399–408, 1999.

[5] A. J. Campbell, J. A. Cook, G. Adey, and B. H. Cuthbertson, “Predicting death and readmission after intensive care discharge,” *Brit. J. Anaesthesia*, vol. 100, no. 5, pp. 656–662, 2008.

[6] A. Vlayen, S. Verelst, G. E. Bekkering, W. Schrooten, J. Hellings, and N. Claes, “Incidence and preventability of adverse events requiring intensive care admission: A systematic review,” *J. Eval. Clin. Practice*, vol. 18, no. 2, pp. 485–497, 2011.

[7] E. A. Martinez *et al.*, “Identifying meaningful outcome measures for the intensive care unit,” *Amer. J. Med. Quality*, vol. 29, no. 2, pp. 144–152, 2013.

[8] M. D. Buist, “Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: Preliminary study,” *BMJ*, vol. 324, no. 7334, pp. 387–390, 2002.

[9] H. Hogan, F. Healey, G. Neale, R. Thomson, C. Vincent, and N. Black, “Preventable deaths due to problems in care in English acute hospitals: A retrospective case record review study,” *BMJ Qual. Saf.*, vol. 21, pp. 737–745, 2012.

[10] L. J. Donaldson, S. S. Panesar, and A. Darzi, “Patient-safety-related hospital deaths in England: Thematic analysis of incidents reported to a national database, 2010-2012,” *PLOS Med.*, vol. 11, no. 6, pp. 1–8, Jun. 2014.

[11] Royal College Physicians, *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*, Updated reprot of a working party. London: RCP, 2017.

[12] C. P. Subbe, R. G. Davies, E. Williams, P. Rutherford, and L. Gemmell, “Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions,” *Anaesthesia*, vol. 58, no. 8, pp. 797–802, 2003.

[13] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence, “APACHE—Acute physiology and chronic health evaluation: A physiologically based classification system,” *Crit. Care Med.*, vol. 9, no. 8, pp. 591–597, 1981.

[14] W. Knaus, E. Draper, D. Wagner, and J. Zimmerman, “APACHE II: A severity of disease classification system,” *Crit. Care Med.*, vol. 13, no. 10, pp. 818–829, 1985.

[15] G. B. Smith, D. R. Prytherch, P. E. Schmidt, P. I. Featherstone, and B. Higgins, “A review, and performance evaluation, of single-parameter ‘track and trigger’ systems,” *Resuscitation*, vol. 79, no. 1, pp. 11–21, 2008.

[16] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, “Review and performance evaluation of aggregate weighted ‘track and trigger’ systems,” *Resuscitation*, vol. 77, no. 2, pp. 170–179, 2008.

[17] R. P. Moreno *et al.*, “SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission,” *Intensive Care Med.*, vol. 31, no. 10, pp. 1345–1355, 2005.

[18] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, “Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients*,” *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.

[19] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, “A review of novelty detection,” *Signal Process.*, vol. 99, pp. 215–249, 2014.

[20] A. Hann, “Multi-parameter monitoring for early warning of patient deterioration,” Ph.D. dissertation, Dept. Eng. Sci., Univ. Oxford, Oxford, U.K., 2008.

[21] E. Gultepe, J. P. Green, H. Nguyen, J. Adams, T. Albertson, and I. Tagkopoulos, “From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system,” *J. Amer. Med. Inform. Assoc.*, vol. 21, pp. 315–325, 2013.

[22] L. Clifton, D. A. Clifton, M. A. F. Pimentel, P. J. Watkinson, and L. Tarassenko, “Gaussian process regression in vital-sign early warning systems,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 6161–6164.

[23] T. A. Lasko, J. C. Denny, and M. A. Levy, “Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data,” *PLOS ONE*, vol. 8, no. 6, pp. 1–13, Jun. 2013.

[24] O. Stegle, S. V. Fallert, D. J. C. MacKay, and S. Brage, “Gaussian process robust regression for noisy heart rate data,” *IEEE Trans. Biomed. Eng.*, vol. 55, no. 9, pp. 2143–2151, Sep. 2008.

[25] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, “Multitask Gaussian processes for multivariate physiological time-series analysis,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, Jan. 2015.

[26] D. Wong, D. A. Clifton, and L. Tarassenko, “Probabilistic detection of vital sign abnormality with Gaussian process regression,” in *Proc. IEEE 12th Int. Conf. Bioinform. Bioeng.*, Nov. 2012, pp. 187–192.

[27] M. A. F. Pimentel, D. A. Clifton, and L. Tarassenko, “Gaussian process clustering for the functional characterisation of vital-sign trajectories,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2013, pp. 1–6.

[28] D. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, “An extreme function theory for novelty detection,” *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 28–37, Feb. 2013.

[29] M. Hravnak, M. A. DeVita, A. Clontz, L. Edwards, C. Valenta, and M. R. Pinsky, “Cardiorespiratory instability before and after implementing an integrated monitoring system*,” *Crit. Care Med.*, vol. 39, no. 1, pp. 65–72, 2011.

[30] M. Ebden, “Gaussian processes: A quick introduction,” 2015, arXiv:1505.02965.

[31] D. Duvenaud, “The Kernel cookbook: Advice on covariance functions,” 2014. [Online]. Available: http://www.cs.toronto.edu/duvenaud/cookbook/

[32] D. K. Duvenaud, “Automatic model construction with Gaussian processes,” Ph.D. dissertation, Dept. Eng. Univ. Cambridge, Cambridge, U.K., 2014.

[33] I. Murray, R. Adams, and D. MacKay, “Elliptical slice sampling,” in *Proc. 13th Int. Conf. AISTATS*, 2010, no. 9, pp. 541–548.

[34] T. Cui, C. Fox, G. Nicholls, and M. O’Sullivan, “Using parallel MCMC sampling to calibrate a computer model of a geothermal reservoir,” Faculty Eng., Univ. Auckland, Auckland, New Zealand, Tech. Rep. no. 686, pp. 1–18, Jan. 2011.

[35] J. Hartikainen and S. Särkkä, “Kalman filtering and smoothing solutions to temporal Gaussian process regression models,” in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, pp. 379–384, Aug. 2010.