Bayesian Gaussian Processes for Identifying the Deteriorating Patient

PRS Transfer Report
Glen Wright Colopy
St. Cross College
University of Oxford

Supervised by:
Dr. David A. Clifton
Prof. Steven J. Roberts

September 6, 2015
Abstract

Patients discharged from the ICU will commonly be placed in intermediary care, such as the step-down ward. Although most of these patients will continue to recover and stabilise, a significant portion will suffer cardiac arrest, other clinical emergencies, and readmission into intensive care where the risk of mortality is significantly higher. Gaussian Process Regression (GPR) models are proposed as a flexible, principled, probabilistic method to address the clinical need to continuously monitor patient vital sign time-series in the step-down ward, where the nurse-to-patient ratio is reduced. To assist in the continuous monitoring of patients, the proposed GPR models focus on the robust forecasting of patient heart rate time-series and early detection of patient deterioration. The proposed methods are tested in a step-down ward data set from the University of Pittsburgh Medical Center, comprising 333 patients, 59 of whom had at least one verified clinical emergency event.

To forecast robustly patient heart rate time-series, a series of covariance kernels are proposed as candidates that may capture the generative process of heart rate variability. More complex kernels prove to be beneficial in ensuring robust forecasting. Priors over certain kernel hyperparameters also demonstrate improved forecasting robustness.

For early deterioration detection, two GPR approaches inspired by novelty detection techniques are tested and compared to the deterioration detection methods in current clinical practice. The univariate GPR approach known as “step-change detection” is shown to outperform univariate thresholding methods, as measured by the trade-off between early warning of deteriorating patients and false alarm rates on non-deteriorating patients. The univariate GPR approach known as “time-series matching” is shown to exceed the performance of the univariate threshold and, in some cases, a 5 vital signs kernel density estimate of vital signs from a training set of healthy patients.

Future work will build on the univariate GPR methods in this report to incorporate the correlation structure between vital signals and to use more principled techniques to account for uncertainty in the fitted GPR hyperparameters.
1 Introduction of Clinical Need

1.1 Overview

The goal of this work is to perform time-series analysis using data from patients in hospitals. A step-down unit\(^1\), or SDU, is “a hospital nursing unit providing care intermediate between that of an intensive care unit and a normally-staffed in-patient division” [1]. The goal of an SDU is to manage the recovery of patients after discharge from the Intensive Care Unit (ICU) while reducing the staff-intensive burden of monitoring patients who are more stable than acutely-ill ICU entrants. The ICU accounts for 1.2% [2, 3] of hospital beds in the UK. ICUs in the US account for 9% - 20% [2, 4] of hospital beds. Exact definitions of an SDU and ICU vary between countries [2, 5]. SDU patients are of a more stable condition than those in the ICU and therefore the SDU has a reduced nurse-to-patient ratio (1 nurse to 4-6 patients) compared to the ICU (1 nurse to 1-2 patients) [6]. It is common for SDUs to be staffed by nurses trained in critical care, just as in the ICU [7].

Critical-care wards in the US have been estimated to account for about 13.4% [8], or 17.4%-39% [9] of total hospital care expenditures, and therefore it is important to optimise the management of patients in such settings.

1.2 Clinical Deterioration and Readmission to ICUs

Readmission Rates Although SDU patients are usually in a stable condition, a significant portion of SDU patients experience a clinical emergency, or require emergency re-admission to the ICU. Various studies across different hospitals have estimated ICU readmission-rates 4.2% - 7.6% [10], 8.8% [11], and 0%-18.3% [12]. Readmission rates were 5.8% for postoperative patients and 6.4% for nonoperative patients [10].

Mortality upon Readmission Readmission to the ICU has significant implications for patient outcomes: mortality rates for ICU patients readmitted within the same hospital stay have been estimated at 40.2% [11]. Another study estimated readmission mortality to be 24.7% [10] (in contrast to 4.0% mortality of patients who were not readmitted) and went on to note, “Several studies conducted in single hospitals during the 1980s and early 1990s indicate that

\(^1\)Alternatively referred to as a step-down ward or high-dependency unit.
roughly 5% to 20% of ICU admissions represent readmission during the same hospitalisation, and that patients readmitted to the ICU have disproportionately high in-hospital mortality rates and lengths of stay.” These high levels of mortality motivate the use of principled methods for identifying, and ideally predicting, physiological deterioration.

**Causes for Readmission** There is significant heterogeneity within patients who were readmitted to the ICU after discharge. A study by Cooper et. al. [10] found that congestive heart failure (7.6%) and sepsis (4.5%) were the most common causes of readmission. Over 20% of the total number of readmissions were from cardiovascular-related causes. Fewer than 20% of patients had identical reasons for initial admission and readmission to the ICU. In contrast, another study calculated that 49% were readmitted “for the same or related diagnosis” [11].

A confounding aspect in these studies is that a single “primary” cause for readmission was required, which ignores the correlation across different clinical risks. Despite the absence of information about the correlation between risks, the heterogeneity of causes is clear evidence that it is advantageous to incorporate a variety of factors when deciding whether a patient requires urgent care or readmission. Cardiovascular-related causes were consistently mentioned as risks for both admission and re-admission to the ICU. It is not surprising that patients who were admitted to the hospital on the account of a cardiovascular condition would exhibit subsequent cardiovascular events. Additionally, it is possible that cardiovascular signals are a bellwether for physiological abnormalities that are otherwise difficult to quantify. This suggests that cardiovascular-based indicators are a reasonable starting point for developing a principled model for tracking patient health.

**Time-Frame of Deterioration and ICU Readmission** In the aforementioned study by Cooper et al. [10], ICU readmission was most common 1 day after ICU discharge (and was more common than discharge within the first day). Over half of readmitted patients were readmitted within 3 days and 22% were readmitted after a week or longer. From this, the time-from-admission is clearly a consideration when assessing a patient’s risk of a clinical event or readmission.

It is typically accepted that SDU patients exhibit symptoms of deterioration well in advance of the response of medical emergency teams (METs), where such teams of additional bedside
support are available (this does not include the UK). One study 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” [6]. This suggests that there is a time-window in advance of a verified clinical event in which an alarm would be justified as a “true positive” alarm. Alarms occurring outside such a window, but still preceding a clinical event, would be considered to be less certainly associated with an event. This would also be the case for alarms that did not precede a confirmed clinical event.

1.3 Proposed Clinical Approaches

Not all deaths after discharge from the ICU are preventable [11, 5, 12] but there is evidence that earlier response can reduce mortality rates [13, 11]. A recent retrospective study found that from 1988-2012, ICUs in the US using the APACHE system (a multivariable threshold-based risk scoring system) saw a drop in mortality rates from 17.3% to 12.4% [4]. Notably, “Most of these dramatic relative decreases in hospital mortality rate occurred between 1988 to 1989 and 1993 to 1996.” The study did not try to relate this trend to any confounding factors, such as new technologies or therapies that would also improve treatment outcomes. Mortality improvements varied greatly by diagnosis. These trends were mirrored by a 23% reduction in ICU length-of-stay (LOS) and a 23% reduction in hospital LOS. An increased flow of patients from ICUs into other wards (including SDUs) was suggested to be a contributing factor in these trends. A reduction in overall ICU mortality has also been documented in the UK [14]. Furthermore, there is a notable interplay between earlier ICU readmission and improved survival upon readmission. Earlier-identified of patients who need to return to the ICU from other wards are more likely to survive their hospital stay [11].

Accordingly, this motivates the goals of the work described in this report: first, we aim to forecast accurately the time-series of vital signs from SDU patients. Second, we aim to detect patient deterioration in the ward setting as early as possible. Both goals will be approached via the probabilistic framework of Gaussian process regression (GPR) to model SDU patient time-series. Gaussian processes (GPs) have several attractive characteristics for modelling patient vital signs, which will be discussed (explicitly) in chapters 2 and 4.

The proposed system to forecast patient vital signs and to detect patient deterioration aims
to address the clinical need described above, and promises to be valuable in clinical setting in the following ways:

First, the proposed system should be of little or no additional burden to the clinical staff. It is already common practice for hospitals to monitor patient vital signs via electronic systems. Computer-based methods can also make use of the large amount of continuous data being collected in a way that the intermittent attention of a human observer cannot. The time-series data for thousands of patients (and their associated clinical outcomes) can also be stored in memory, which is invisible to clinicians. A further improvement over existing methods is to learn dynamics of the data for individual patients and adapt accordingly, while also being free from human transcription and calculation error [15].

Second, a computer-based system for continuous monitoring will alleviate some of the risks associated with the reduced nurse-to-patient ratio that exists in the SDU (compared with the ICU). Continuous electronic monitoring will provide patient care between the intermittent observations of the nursing staff. By allowing for clinical warnings in the absence of clinical staff, continuous observation decreases the lag between the first exhibition of abnormal physiology and the action undertaken by medical staff. Earlier action should, in turn, be associated with decreased mortality, and an improvement in other clinical outcomes. Furthermore, nursing staff can focus on patients demonstrating the greatest risk of an adverse clinical event. For example, those patients whose vital signs are forecasted to take values in a range deemed to be indicative of normality. This would further allow the optimisation of hospital personnel.

Third, there are many factors that a computer-based model could incorporate. Given the consistent recognition of cardiovascular events as a cause for readmission, the proposed work begins with a thorough investigation of the information contained in each patient’s time-series of heart-rate (HR) measurements. Future work will include other cardiovascular indicators: blood-oxygen saturation (SpO₂), breathing rate (BR), and systolic and diastolic blood pressure (SBP and DBP). Above, we identified importance of time-based factors, such as LOS for tracking patient deterioration, and proposed methods can incorporate this information where desirable.

The following report describes how to use GPR to model and forecast HR for SDU patients at the University of Pittsburgh Medical Center (UPMC). We will demonstrate that these models
for HR are also useful for early detection of patient deterioration, which permit members of the clinical staff to take steps, with the ultimate aim of avoiding emergency events and improve patient outcomes.

2 Literature Review

2.1 Current Approaches to Patient Monitoring

There is a multitude of patient monitoring systems in clinical use for the ICU, SDU, and other hospital wards. The first section of this literature review will discuss several of these systems and their various common attributes for detecting patient deterioration. A useful distinction can be drawn between heuristic and empirical methods. Heuristic methods have been developed since the beginning of specialisation in intensive care, but were not widely comparable between hospitals until the 1980s [16]. They are typically concerned with the development of rules by which clinicians can quantify the current risk-status of a patient. Empirical methods, on the other hand, generally work towards quantifying the relationship between a set of predictors and various clinical outcomes. The risk of a clinical event can then be examined within the context of the probability distribution that the model describes. Not all methods that describe the clinical outcome as a function of predictive variables have an exclusively probabilistic interpretation: support vectors machines, neural networks, and random forests have all been used in the ICU setting to forecast clinical outcomes such as LOS and in-hospital mortality. These methods will not be covered in the review, due to our focus on detecting deterioration in post-ICU patients.

Heuristic Approaches to Patient Monitoring The current practice of patient vital sign monitoring is to observe a patient’s current measurements and decide whether the patient is currently in an at-risk state. Simplistic methods include rule-based thresholds, such that a warning is signalled if a vital sign exceeds a pre-specified threshold. These thresholds are usually set according to clinical experience about the global population of stable patients. Univariate alarm thresholds may be factory-programmed with the device generating the vital sign measurements [17]. Improvements have been made by considering the simultaneous abnormality of multiple vital signs, for example, the manually-calculated Modified Early Warning Score (MEWS). An important problem with this approach is that risk-assessment is staff-intensive, which prohib-
its continuous monitoring. A patient struggling to maintain homoeostasis might only exhibit abnormalities when the staff is absent, leaving the deterioration undetected. Alternatively, the patient might demonstrate significant variation, but remain entirely within the pre-specified alarm bounds for the global population. Furthermore, the MEWS does not account for vital sign trajectories or signals that are individually healthy but jointly abnormal. MEWS may also be difficult to calculate in the presence of missing information.

**Probabilistic Approaches to Patient Monitoring**  A great improvement of the above is offered by continuous probabilistic monitoring systems that compare a patient's current set of vital signs to those of a global population of healthy patients. Possibly the best known of these methods is the “Parzen-window” kernel density estimate (KDE) [18]. If the observed data stray sufficiently from those of healthy patients then it is a signal of abnormality and potential deterioration. Other probabilistic methods have made use of extreme value theory (EVT) to anticipate whether extreme observations represent a true deterioration, an artefactual measurement, or a reasonably-expected extreme arising from long-periods of continuous monitoring [19, 20, 21]. In essence, these methods replace an absolute threshold with a probabilistic threshold that can account for the correlation between vital signs and lengthy observation periods. Patient risk, though, is still assessed at a single point in time, thereby losing information from previous measurements, and making the unrealistic assumption of i.i.d. observations. To address the correlated time series nature of data, methods have been suggested to estimate patient trajectories, but these frequently rely on manipulating the data to fit into a specific model, for example, into discrete time points [22].

**2.2 Gaussian Processes for Vital Sign Modelling and Prediction**

This first goal of the work described by this report is to forecast robustly the HR time-series of SDU patients. GPR is commonly in time-series modelling. This section will cover common advantages of GPR modelling of time-series, followed by a summary of several important papers on GPR modelling for vital sign time-series (particularly for the purpose of forecasting and interpolation.) The latter papers will be especially useful as a starting point for GPR forecasting in the SDU setting.
Features of GPR Applicable to vital sign Modelling and Forecasting  

GPs have been used since the 1940s but are increasingly popular as a principled, flexible, probabilistic approach to the analysis of complex data. Their flexibility makes them attractive to model a wide-ranges of data [5].

GPR holds several advantages over more traditional regression methods. Most broadly, GPR does not impute a functional form relating the dependent variables to the regressor. Pre-specified functional forms can include those meant to handle non-linearity (e.g., polynomial, fractional-polynomial [23], or Poisson regression) or to specify the extent to which previous observations affect future observations (e.g., models with autoregressive or moving-average components.) Such pre-defined functional forms lead to inherent challenges in the modelling process. Most obviously, the pre-specified model may be misspecified and deviate significantly from the actual form of the generative process, especially where data are absent or sparse.

There are many developments with GPs that are hitherto unexploited in patient monitoring, but which may warrant future investigation: McHutcheon et al. [24] examine uncertainty in a GPR model and the importance of precise time-stamps depending on how quickly the functions are changing with respect to time. In a patient-monitoring framework, poorly-synchronised monitoring devices will likely contain errors in the times of the recorded measurements, creating temporal uncertainty. This is especially true in mobile or home patient-monitoring where information might be relayed between multiple devices. A simple example could be for the Bluetooth connection of a monitoring device becoming disconnected or desynchronised from other devices being used to monitor the same patient. GPs could be designed to accommodate this uncertainty if it were determined to be significant.

GPs are amenable to analysis in change-point detection [25]; for example, to identify when there has been a change in the underlying generative process, so that the previous model can no longer describe the data subsequent to the change-point. Examples of change-points include changes in the model hyperparameters; and drift, or bias, in the observations. Various functional forms can be designed for robust prediction. It is plausible that in the course of deterioration, the underlying dynamics of a patient’s time-series will deviate from the previous dynamics. These applications might be very useful in those instances in which a patient does not have forecastable trajectory of impending deterioration, but which instead can only be
identified after a dynamic change has occurred.

**Previous Applications of GPRs to Vital Sign Monitoring and Prediction**

GPR monitoring have been used for mobile monitoring, which may also include applications in which the patient is ambulatory in the ward, as is the case for the majority of hospital patients [17]. There is a small but growing body of publications describing the use of GPs to model and predict vital signs. The vital signs are mainly recorded using wearable sensors, with the authors exploiting the probabilistic framework of GPs to handle various noise components, such as sensor noise, quantisation, and artefact arising from patient movement.

Stegle et al. [26] examined the use of GPs for free-living HR monitoring with 40 adult subjects. Model-fitting and forecasting were improved by the use of clustering based on “auxiliary” ECG-waveform summary data (for example, the variability of inter-beat intervals, the extrema within these intervals, and the fraction of time these extrema fell outside a credible range). The latter were used to identify noisy periods in the data. The auxiliary variables were used to generate clusters of variables with different levels of noise. The noise model was then a mixture of the different classes with varying noise. After clustering, the kernel accounted for two additive components of HR variability: a Matern$^{3/2}$ kernel for short-term variability, and diurnal periodicity using a locally-periodic [27, 28] or quasi-periodic [29] kernel. Model parameters were found using expectation maximisation to approximate the posterior distribution and then choosing values that maximised log-likelihood.

Clifton et al. [17] provided a proof-of-concept for GPR to improve early warning systems in hospital wards. Two hundred post-operative cancer patients in an SDU-like environment had manually-recorded vital signs. The majority of patients had further recordings using wearable sensors and bed-side monitors. HR, BR, SpO$_2$, SBP, and DBP were recorded. A squared-exponential kernel was used for each vital sign and 10-fold cross validation was used to select the values of the hyperparameters. To simulate missing data, 25% of observations from each patient’s time-series were removed. The performance of GPR for predicting the value of the missing values was compared to that of population-based mean imputation (which is the current clinical practice for missing data), a support vector regressor, and a patient-based mean. GPR values for each time-series were further used as inputs to a novelty score based on extreme value statistics [19]. Using GPR prediction and interpolation, a patient’s novelty score could
be more robustly evaluated, including during periods with large amounts of missing data. In retrospect, the knowledge of the novelty scores with interpolated inputs would have given a median advanced deterioration warning of 13.4 hours amongst the seven “abnormal” patients identified by the study. In another paper by Clifton et al. [30] on the same group of patients, a squared-exponential kernel was used to model each vital sign and hyperparameter values were found by maximising log-negative marginal likelihood.

In a multitask GP (MTGP) application, Durichen et al. [29] demonstrated how the combination of sensors improved prediction accuracy for BR and HR using two data sets. The first data set combined four univariate time-series (three optical markers and one respiration belt) with the goal of predicting the respiratory rate from an optical marker placed on the chest. Quasi-periodic kernels were used to model each of the univariate tasks. Forecast accuracy was shown to improve with each additional task added to the model. The second data set combined manually-collected univariate time-series for BR and HR for prediction in patients recovering from cancer-related surgery. A third data set (using porcine subjects) used MTGPs to capture the correlation between internal and external motion due to respiratory rate, finding quasi-periodic kernels to be more apt for the prediction task than periodic or squared-exponential kernels. For each experiment, hyperparameters were assigned their using gradient descent. A tool-box corresponding to this work is available online [31].

MTGPs have also been used to characterise the trajectories of ward patients to determine whether a patient is likely to experience an adverse clinical event. Pimentel et al. [32] examined HR and BR time-series with the goal of modelling a patient’s convergence towards stability before discharge from the ward. A squared-exponential kernel was additively combined with a quasi-periodic kernel, the former to model within-day trajectories and the latter to capture diurnal variability. Hyperparameter values were determined using a maximum-likelihood estimate grid-search optimiser. The similarity of a test trajectory to a reference GP was calculated using an average of localised likelihoods between the two. The study considered 154 patients, who were labelled as 138 “normal” and 16 “abnormal” depending on whether the patient was discharged home, or into an intensive care unit, respectively. The data set was repeatedly partitioned to create a training set of 90 “normal” patients, on which to develop prototypical clusters of “normal” trajectories. Performance was measured according to sensitivity, specificity,
and AUROC in classifying the held-out “normal” and “abnormal” patient trajectories. The average local-likelihood method outperformed classification based on multivariate dynamic time warping (DTW). This study is also pertinent to the second goal of this report, and the work described in sections 2.3 and 4.3.

2.3 Gaussian Processes for Detecting Deterioration

The second goal of the work described by this report is to provide early detection of patient deterioration. We take two approaches, which we will describe as “step-change detection” and “time-series matching.” Both approaches are motivated by the concept of novelty detection but differ concerning the definition of a model of normality. A brief overview of novelty detection will be provided, followed by a review of the constituent methods particular to the “time-series matching” approach for the SDU patients in the UPMC data set. The aforementioned EVT papers [19, 20, 21] form the basis for the proposed “step-change detection” methods.

Novelty Detection As noted earlier, EVT has a central role in the area of novelty detection for biomedical applications [19, 20, 21]. The principled approach in EVT-based methods is very attractive, but is not the sole criterion by which a decision on “novelty” can be made. Pimentel et al. [33] define novelty detection as “the task of recognising that test data differ in some respect from the data that are available during training.” The value of such methods, it is explained, come when the preponderance of training data is from the “normal” class, thereby hindering the creation of a satisfactory explicit model for the “abnormal” class. The terms “anomaly detection” and “outlier detection” are noted to be popular synonyms, with the caveat that the identification of “outlier” points is usually undertaken with the goal to take remedial measures during preprocessing. The review makes a further useful distinction in describing how novelty detection techniques fall into five general categories: (i) probabilistic, (ii) distance-based, (iii) reconstruction based, (iv) domain-based, and (v) information-theoretic techniques.

This view partitions the approaches of the current clinical standard with our work: heuristic rules-based methods, such as MEWS or APACHE, are “domain-based” methods, in that they partition the vital sign space into areas of increasing risk, but do not specify the density of measurements in a given region. The aforementioned support vector machine approaches,

\[\text{Both approaches will be described further in chapter 4.}\]
instead of using heuristic rules, divide the data space via a hyperplane to optimise a loss function. The current state-of-the-art KDE method [18] falls in the "probabilistic" category, in which a new patient’s observations activate an alarm if they fall sufficiently far the region of high support. Several variants of the “step-change detector” also fall into the probabilistic category. The “time-series matching” approach is similar to the description of the “distance-based category” in that a candidate fitted time-series is compared only to the “$k^{th}$ most similar” training pattern, but an important distinction is that the metric of interest, Kullback-Leibler (KL) divergence is not a true distance, and therefore, might be better categorised within the “information-theoretic” category, noting the relationship between KL divergences and entropy.

**Time-Series Matching** The time-series matching approach will involve the comparison and quantification of test time-series to a dictionary of GPs derived from the time-series of stable patients. As seen in Pimentel et al. [32] there are many ways to measure “distance” between two time-series. Particular examples include “longest common subsequence”, the distance between extracted features, averaged local likelihood, and DTW. A particular objection to DTW is that any “time-warping” would undermine the underlying information pertaining to the “time component” of the patient time-series which distinguishes GPR time-series from i.i.d. methods. A more general objection to any purely distance-based technique is that they may not model the uncertainty if both time-series are inherently stochastic.

KL divergence, on the other hand, accounts for dissimilarity between two distributions. KL divergence between two multivariate Gaussian distributions of equal and finite dimension is well-known. Although KL divergence is not a distance, there are several fast methods for clustering multivariate Gaussians using KL divergence [34]. These algorithms are especially straightforward for multivariate Gaussians with a diagonal covariance matrix. KL divergence can also examine the dissimilarity between GPs.

In bioinformatics, GPs have likewise been used to quantify the “distance” between two uncertain functions. In particular, gene expression curves are useful to understand cellular activity under different conditions. The data are time-series by nature, and are frequently produced in replicates due to the noise induced by laboratory conditions or biological variation. Stegel et al. [35] use a mixture model to model the intervals of curves that are differentially expressed. Kalaitzis et al. [36] use a likelihood-ratio approach, aimed to examine the total differential
expression, instead of identifying specific intervals of deviation. Any of these techniques could form the basis for the comparison of two GPR-fitted time-series from patient vital signs.

3 Data Description

The data set considered in this report comprises 333 adult patients 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 the (KDE-based) monitoring system described in [18]. The latter involves a model of normality that had been constructed using data from two studies at the John Radcliffe Hospital.

Phase 1 of the UPMC trial involved only data acquisition. The KDE was not used to alert clinicians, and the nursing staff was blinded to the model. The goal of phase 1 was to use the resulting data to optimise the value of the novelty threshold for the KDE used to generate alerts. Phase 1 started in November of 2006 and lasted eight weeks. The subsequent two phases were used, respectively, to train nursing staff to respond to alerts generated by the system, and evaluate the use of the system as an intervention. Only data from patients in phase 1 were used in the work described by this report.

3.1 Annotation of Clinical Emergency Events

Over the course of phase 1 (across all 333 Patients), the medical staff made only 7 MET calls based on extremal vital signs. Artefactual measurements were removed from the recorded vital sign time-series, and additional events that would have satisfied MET calling criteria were then identified. The MET calling criteria are shown in Table 1, which was taken from [18]. Notably, the MET thresholds are univariate limits, and any threshold exceedance was required to last for at least 4 of the previous 5 minutes to generate an alarm. A further artefactual issue was identified as the cause of a large number of alarms on BR, and this was remedied. Alarms occurring close in time were merged, leaving 407 events to be validated by a clinician.

The clinicians were provided with the relevant time-series data to determine whether any of the 407 events were non-artefactual. Any events that clinicians confirmed to be non-artefactual exceedances of the MET Thresholds were labelled as so-called C’-events. There were 237 C’-events between 83 patients. Clinicians were then asked whether an MET ought to have been
<table>
<thead>
<tr>
<th>Vital Sign</th>
<th>Lower Threshold</th>
<th>Upper Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (bpm)</td>
<td>40</td>
<td>140</td>
</tr>
<tr>
<td>BR (bpm)</td>
<td>8</td>
<td>36</td>
</tr>
<tr>
<td>(\text{SpO}_2)%</td>
<td>85</td>
<td>-</td>
</tr>
<tr>
<td>SBP (mmHg)</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>DBP (mmHg)</td>
<td>85</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Univariate MET alarm thresholds

called for each C'-event to perform an emergency intervention. C'-events thus confirmed to have warranted an emergency intervention were labelled C''-events. There were 112 C''-events between 59 patients. (This means that there were 24 patients whose vital signs exhibited non-artefactual, sustained MET exceedances, but who were not thought to require emergency intervention.) Patients who experienced a C''-event are labelled C''-patients. Those who did not experience a C''-event are labelled Non-C''-patients.

The presence of 112 emergency C''-events when only 7 were called in practice supports the understanding that continuous monitoring can add value to the intermittent observation of nursing staff. The annotated C'-events each had an associated start-time, stop-time, and primary cause.

### 3.2 Vital Sign Measurements

The (unfiltered) UPMC data comprises a unique time-series for each of the five vital signs: HR, BR, \(\text{SpO}_2\), SBP, and DBP, all of which were recorded by the bedside monitors.\(^3\) The time-series of each vital sign comprised paired time-stamps and measurements.

For several reasons, including differing sampling rates, artefact removal, probe detachment, and different starting times, time-stamps are usually not aligned between vital signs for the same patient. Measurements outside the physiologically-plausible ranges in Table 2 are considered artefactual and removed [18]. 11 patients were entirely missing data, or had less than one hour of data. After removing these 11 patients, many remaining patients had an individual vital sign with less than one hour of data (Table 3). In particular, Non-C''-patients were much more likely to have less than one hour of BR data. After artefact removal, and down-sampling to \(f_s = \frac{1}{60}\) Hz, a prototypical Non-C''-patient and C''-patient are displayed in Fig. 1. In practice,

\(^3\)Temperature was recorded but not used in any of the models discussed, and is therefore not included in our discussion.
<table>
<thead>
<tr>
<th></th>
<th>Lower Threshold</th>
<th>Upper Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (bpm)</td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>BR (bpm)</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>SpO₂ (%)</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Artefactual thresholds applied to phase 1 patients

HR measurements were log-transformed in advance of regression. This ensures that no part of the posterior distribution has support on negative HR values.

Fig. 1 illustrates many of the challenges encountered with continuous vital sign data. First, each patient has several extended periods of missing measurements, which is a challenge in continuous monitoring. For the Non-C”-patient, only a single initial measurement is available, followed by 20 hours of missingness. Between approximately 35-38 hours, the Non-C”-patient has a step-change with elevated HR. Since this period was not marked as a C’-event (where data exceeded the MET limit of 140 bpm), we know that clinical experts considered these measurements to be artefactual. Since sections such as these are neither marked as C’-events or C”-events, and many of the measurements fall within both the artefactual limits (Table 2) and the MET limits (Table 1) it is left to the researcher to identify these questionable measurements and decide how to handle them appropriately. At this stage of the project, measurements of this kind were left in the time-series with the understanding that, if taken at face-value, they would represent highly unusual physiological dynamics. A more principled approach to identifying and handling this data will be proposed in chapter 6.

The C”-patient has three C’-events due to high BR (where the MET limit is 36 bpm), which is reasonable, given that BR is visibly elevated over that of the Non-C”-patient. The C”-patient’s HR is usually about 20 bpm higher than that of the Non-C”-patient. The preponderance of SpO₂ measurements at 100% suggests that the C”-patient received a significant amount of oxygen therapy. (Patients breathing room air typically have SpO₂ values around 95%.) A final observation is SBP and DBP were sampled much more frequently, for the C”-patient compared with the Non-C”-Patient. This suggests that the clinical staff was taking particular interest in this patient following the previous MET-level measurements. This observation supports our choice to focus our predictive methods on only the first C’-event for each C”-patient, which has previously been used to evaluate the efficacy of patient monitoring systems [37].

No radical differences between C”-patients and Non-C”-patients appeared in the univariate
Table 3: Counts of patients with less-than 1 hour of data for a particular vital sign

<table>
<thead>
<tr>
<th></th>
<th>HR</th>
<th>BR</th>
<th>SpO₂</th>
<th>SBP</th>
<th>DBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-C&quot;-Patients</td>
<td>5</td>
<td>21</td>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C&quot;-Patients</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1: A complete electronic time-series for one C"-patient and one Non-C"-patient
distributions of each of the vital signs (Fig. 2). C\textsuperscript{−}-patients appeared to have, on average, heavier right-hand tails in BR than did Non-C\textsuperscript{−}-patients (Fig. 3). It is difficult to draw too many conclusions from these figures given that a C\textsuperscript{−}-patient could have long periods of “normal” physiology with only brief periods of abnormal physiology; otherwise, far more than 7 MET calls would have been made during phase 1. The LOS of C\textsuperscript{−}-patients (Fig. 4) seemed to have a much heavier right-tail than that of Non-C\textsuperscript{−}-patients (Fig. 4). This could be evidence of the clinician’s apprehension over releasing C\textsuperscript{−}-patients or, it could be coincidental due to a smaller sample size.

### 3.3 Description of C\textsuperscript{−}-Events

For the purpose of detecting deterioration, we are particularly interested in each C\textsuperscript{−}-patient’s first C\textsuperscript{−}-event because it is possible that vital signs subsequent to that first C\textsuperscript{−}-event would be affected by clinical intervention (as may be the case with the C\textsuperscript{−}-patient in Fig. 1). There is significant heterogeneity (Table 4), but no obvious relationship between the first C\textsuperscript{−}-event time-stamp and its primary cause (Fig. 5). Within a C\textsuperscript{−}-patient, multiple C\textsuperscript{−}-events tended to share the same primary cause.

<table>
<thead>
<tr>
<th>Cause</th>
<th>High HR</th>
<th>Low HR</th>
<th>High BR</th>
<th>Low BR</th>
<th>Low SpO\textsubscript{2}</th>
<th>High SBP</th>
<th>Low SBP</th>
<th>High DBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>20</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 3: The distribution of vital sign measurements within each patient. The blue boxes show the $25^{th}$ and $75^{th}$ quantile of each patient’s measurements. Individual red marks indicate outliers, as designated by Matlab.

Figure 4: Length of stay by C"-event status
Figure 5: First C"-event by Primary Cause and Time-stamp

4 Methods and Results for Forecasting of Patient HR Data

4.1 Review of Gaussian Process Regression

Ebden [38] provides a helpful introduction to GPR which states, “a Gaussian process generates data located throughout some domain such that any finite subset of the range follows a multivariate Gaussian distribution.” For this work, the GPR models an HR time-series of measurements and timestamps \([y, x]\) as a multivariate Gaussian distribution, \(N(m, K(x, x))\). The mean vector \(m\) is usually assumed to be \(0\) (after subtracting the mean of \(y\)). The covariance matrix of two random vectors \(x\) and \(x'\) is denoted \(K(x, x')\), such that the \((i, j)\)th element is the covariance of \(x_i\) and \(x'_j\). In GPR, covariance between two points is defined by a positive definite kernel function \(k\) (discussed further in chapter 4.3). For two output-input pairs, \((y, x)\) and \((y', x')\), the kernel function defines \(Cov(y, y') = k(x, x')\). Given a set of training observations \([y, x]\) the distribution of test points \([y_*, x_*]\) is

\[
p(y_* | y) \sim N(\bar{y}_*, \Sigma),
\]

such that
\[ \bar{y}_* = K(x_*, x) K(x, x)^{-1} y, \]
\[ \Sigma = K(x_*, x_*) - K(x_*, x) K(x, x)^{-1} K(x_*, x)^T. \]

(2)

The functional form of kernel \( k \) is governed by a vector of hyperparameters \( \theta \). The log marginal likelihood (LML) of the observed points, given a value of \( \theta \) is

\[ \log p(y|x, \theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi. \]

(3)

The term “marginal” indicates the value of has been \( \theta \) integrated out. Maximizing LML (e.g. via gradient descent) is likely to find values of \( \theta \) that aptly describe of the covariance structure of the observed data. Depending on the data \( y \) and the form of \( k \), it is common for multiple values of \( \theta \) to be locally optimal.

4.2 Commonly-Used Covariance Kernels

Consider a generic covariance kernel \( k_a \) which is a positive definite function with inputs \( x \), \( x' \), and an associated set of hyperparameters \( \theta_a \). Here we use “a” to link a specific set of hyperparameters \( \theta_a \) to a specific kernel function, \( k_a \). This convention will be helpful when it is important to distinguish between multiple kernels, and their associated hyperparameters. To evaluate a kernel, the only requirement is \( d(x, x') \), the distance between \( x \) and \( x' \), and a value for each hyperparameter of \( \theta_a \). The evaluation of \( k_a(d|\theta_a) \) defines the covariance between the outputs, \( Cov(y, y') \). The goal is to find one or more values of \( \theta_a \) that describe the covariance structure appropriately.

There are several functional forms that are commonly used to describe how the covariance is determined by \( d(x, x') \). For legibility, \( k_a(x, x'|\theta_a) \), or \( k_a(d|\theta_a) \), will be replaced by \( k_a \) or \( k_a(\theta_a) \) in the text wherever it is unambiguous. A superscript on \( \theta_a \) will describe the method by which \( \theta_a \) was estimated (e.g. \( \theta_a^{MAP} \) is \( \theta_a \) estimated by MAP, defined below). Furthermore, for individual hyperparameters within \( \theta_a \) both subscripts and superscripts will be used to indicate, respectively, the index for multiple hyperparameters fulfilling similar purposes, and \( \theta_a \) to which the hyperparameter belongs. For example, \( \lambda_1^{\theta_a^{MAP}} \) will denote the first length-scale hyperparameter (\( \lambda_1 \)) of kernel \( k_a \) where \( \theta_a \) is fit according to a MAP estimate.

We first consider the squared-exponential, or Gaussian kernel, described by:
\[
k_{SE}(d|\theta_{SE}) = h^2 \exp\left(-\frac{d^2}{2\lambda^2}\right), \text{ s.t.} \theta_{SE} = [h, \lambda].
\]

Holding \(\theta_{SE}\) constant, an increase in \(d\) will decrease \(k_{SE}\), which is a reasonable attribute in many instances when covariance output \(y\) (e.g. HR) is expected to decrease as observation times \((x, x')\) become increasingly far apart. Furthermore, \(k_{SE}\) is infinitely differentiable, making it an appropriate model for extremely smooth functions. The hyperparameter \(h\) serves to modulate the total level of covariance between any two points and, hence, is called the “output scale” or “signal variance.” The latter is apt because \(h\) is the exact level of variance when two points are zero distance apart. The hyperparameter \(\lambda\) modulates the effect of the distance between two points prior to exponentiation (otherwise known as the “kernel distance”). It can be seen that by decreasing \(\lambda\), \(k_{SE}\) diminishes much more rapidly with distance. Conversely, a large \(\lambda\) value slows down the exponential decay, allowing for points to remain co-influential over a longer distance. Common names for \(\lambda\) are the “length scale”, “input scale”, or the “effective range”.

The next two kernels we consider are of the Matérn family of covariance functions. They are, respectively, the Matérn\((\frac{3}{2})\) covariance function and the Matérn\((\frac{5}{2})\) covariance function:

\[
k_{\text{Mat}}(\frac{3}{2})(d|\theta_{\text{Mat}}(\frac{3}{2})) = h^2(1 + \frac{d\sqrt{3}}{\lambda}) \exp\left(-\frac{d\sqrt{3}}{\lambda}\right), \text{ s.t.} \theta_{\text{Mat}}(\frac{3}{2}) = [h, \lambda].
\]

\[
k_{\text{Mat}}(\frac{5}{2})(d|\theta_{\text{Mat}}(\frac{5}{2})) = h^2(1 + \frac{d\sqrt{5}}{\lambda} + \frac{5d^2}{3\lambda^2}) \exp\left(-\frac{d\sqrt{5}}{\lambda}\right), \text{ s.t.} \theta_{\text{Mat}}(\frac{5}{2}) = [h, \lambda].
\]

Being once-differentiable and twice-differentiable, respectively, the Matern kernels are attractive options to model the sharp discontinuities that frequently manifest in noisy short-term dynamics of vital sign measurements, such as HR.

A final, but important, covariance function is the white-noise covariance function, defined as:

\[
k_{WN}(x, x'|\theta_{WN}) = \sigma_n^2 \delta(x, x'), \text{ s.t.} \theta_{WN} = [\sigma_n].
\]

The Kronecker delta function \(\delta\) does not operate on the distance between two points, but
instead operates on their identity, or their indices. So \( k_{WN} \) is only non-zero when comparing a point to itself, and is used to represent additive Gaussian noise along the main diagonal of the covariance matrix of the training points. Common terms for the \( \sigma^2_n \) hyperparameter are the “noise variance”, “residual variance”, or “nugget” (in the field of geostatistics). The white-noise kernel will be appended to each of the candidate kernels to represent the measurement noise inherent in each point of the time-series. It is worth noting that when the white-noise kernel is non-zero, the covariance function is a sum of the signal-variance and the noise-variance. In essence, these terms compete with each other to describe the sources of variance within the model, but whereas the signal-variance is present throughout the model, the noise-variance is only present where measurement error is described.

A useful property of valid covariance functions is that their sums and products are valid covariance functions as well. Accordingly, if the kernel \( k_a = \sum_{i=1}^{m} k_i \), for \( k_i \) each with a respective set of hyperparameters \( \theta_i \), then the respective hyperparameter set of \( k_a \) is \( \theta_a = \bigcup_{i=1}^{m} \theta_i \). Each hyperparameter \( h, \lambda, \sigma_n \) has interpretable units of measurement: in our application, \( h \) and \( \sigma_n \) are in units of log(HR), and \( \lambda \) is in minutes.

### 4.3 Kernel Choice in Forecasting of Patient HR Data

#### Candidate Covariance Kernels for Modelling Heart-rate

To model HR time-series, the following candidate covariance functions are proposed:

\[
k_0 = k_{SE_1} + k_{WN} \\
= h_1^2 \exp\left(-\frac{d^2}{\lambda_1^2}\right) + \sigma^2_n \delta(x, x'), \text{s.t.}
\]

\( \theta_0 = [h_1, \lambda_1, \sigma_n] \).

\[
k_1 = k_{SE_1} + k_{SE_2} + k_{WN} \\
= h_1^2 \exp\left(-\frac{d^2}{\lambda_1^2}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + \sigma^2_n \delta(x, x'), \text{s.t.}
\]

\( \theta_1 = [h_1, \lambda_1, h_2, \lambda_2, \sigma_n] \).
\[\begin{align*}
k_2 &= k_{\text{Mat}(\frac{2}{3})} + k_{SE_2} + k_{WN} \\
&= h_1^2(1 + \frac{d\sqrt{3}}{\lambda_1}) \exp\left(-\frac{d\sqrt{3}}{\lambda_1}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + \sigma_n^2 \delta(x,x'), \text{s.t.} \\
\theta_2 &= [h_1, \lambda_1, h_2, \lambda_2, \sigma_n].
\end{align*}\]

\[\begin{align*}
k_3 &= k_{\text{Mat}(\frac{2}{3})} + k_{SE_2} + k_{WN} \\
&= h_1^2(1 + \frac{d\sqrt{5}}{3\lambda_1^3} + \frac{5d^2}{3\lambda_2^3}) \exp\left(-\frac{d\sqrt{5}}{\lambda_1}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + \sigma_n^2 \delta(x,x'), \text{s.t.} \\
\theta_3 &= [h_1, \lambda_1, h_2, \lambda_2, \sigma_n].
\end{align*}\]

\[\begin{align*}
k_4 &= k_{SE_1} + k_{SE_2} + k_{SE_3} + k_{WN} \\
&= h_1^2 \exp\left(-\frac{d^2}{\lambda_1}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + h_3^2 \exp\left(-\frac{d^2}{\lambda_3^2}\right) + \sigma_n^2 \delta(x,x'), \text{s.t.} \\
\theta_4 &= [h_1, \lambda_1, h_2, \lambda_2, h_3, \lambda_3, \sigma_n].
\end{align*}\]

\[\begin{align*}
k_5 &= k_{\text{Mat}(\frac{2}{3})} + k_{SE_2} + k_{SE_3} + k_{WN} \\
&= h_1^2(1 + \frac{d\sqrt{3}}{\lambda_1}) \exp\left(-\frac{d\sqrt{3}}{\lambda_1}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + h_3^2 \exp\left(-\frac{d^2}{\lambda_3^2}\right) + \sigma_n^2 \delta(x,x'), \text{s.t.} \\
\theta_5 &= [h_1, \lambda_1, h_2, \lambda_2, h_3, \lambda_3, \sigma_n].
\end{align*}\]

\[\begin{align*}
k_6 &= k_{\text{Mat}(\frac{2}{3})} + k_{SE_2} + k_{SE_3} + k_{WN} \\
&= h_1^2(1 + \frac{d\sqrt{5}}{\lambda_1} + \frac{5d^2}{3\lambda_1^3}) \exp\left(-\frac{d\sqrt{5}}{\lambda_1}\right) + h_2^2 \exp\left(-\frac{d^2}{\lambda_2^2}\right) + h_3^2 \exp\left(-\frac{d^2}{\lambda_3^2}\right) + \sigma_n^2 \delta(x,x'), \text{s.t.} \\
\theta_6 &= [h_1, \lambda_1, h_2, \lambda_2, h_3, \lambda_3, \sigma_n].
\end{align*}\]

For each candidate kernel, each of the constituent kernels is intended to describe a source of variance across a physiologically-relevant time-scale. Table 5 describes the anticipated length-scales, determined by the total number of kernels. The interest in each kernel, as a candidate for describing the variance in HR time-series, is as follows:

Kernel \(k_0\) hypothesises that, aside from random measurement error, the HR time-series has a single, infinitely-smooth latent function. Since it is the simplest kernel under consideration we will take it to be the baseline comparator for all other methods. Each other kernel \((k_1, \ldots, k_4)\) will need to demonstrate an added improvement over to \(k_0\) warrant its additional complexity.
Having only a single length-scale over which to describe variance, the corresponding length-scale $\lambda^{\theta_0}$ has a wide-range of possible values (Table 5), depending on the level of detail to be fit. Ostensibly, the performance of $k_0$ will be hampered by forcing a choice of only one length-scale, particularly for a MAP estimate. (A posterior mean estimate could mitigate this short-coming.)

Kernel $k_1$ also assumes an infinitely smooth latent function, but with two sources of variance: $\lambda^{\theta_1}_2$ describes trends across hours, and $\lambda^{\theta_1}_1$ describes trends on the order of minutes. $\lambda^{\theta_1}_1$ is expected to fit short-term detail (such as beat-to-beat HR variability), while $\lambda^{\theta_1}_2$ describes longer-term variability (such as that due to BP).

The kernels $k_2$ and $k_3$ make the same assumptions as $k_1$ regarding the length-scales of the sources of variance, but deviate from the assumption that the shorter-term source of variance is smooth. Instead, $k_2$ encodes once-differentiable short-term signal variance, and $k_3$ encodes twice-differentiable short-term variance added to a smooth longer-term trend ($\lambda^{\theta_2}_2$ and $\lambda^{\theta_2}_3$).

The kernel $k_4$ describes an infinitely smooth function with variance at three length-scales. The idea is to allow $\lambda^{\theta_4}_1$ to fit minute-by-minute detail, while still accounting for larger scale trends. In this, $k_4$ may handle short perturbations, without losing the capacity to track of important long-term trends (such as those that might be indicative of abnormality). Kernels $k_2$ and $k_3$ make the same assumptions as $k_4$ regarding the length-scales of the sources of variance, but instead hypothesises that the minute-by-minute variance is discontinuous.

An attractive aspect of the GPR approach is that, without explicit specification of the functional form of an HR time-series, we may encode plausible dynamics of the process assumed to generate HR measurements. Fig. 7 illustrates such dynamics plotting the $\theta^{MAP}$ fits of 4 kernels to the same HR time-series.

Two more baseline predictors are also included: $k_{-1}$ a constant prediction, being the training mean, and $k_{-2}$, a constant prediction, being the last observed value. Both will predict constant posterior variance (at the training set variance).

**Fitting a Gaussian Process’s hyperparameters** To estimate the values of $\theta_a$ for any kernel $k_a$, two iterative random samples of the posterior space were taken, as described in Fig. 6 and Table 5. The first estimate for $\theta_a$ is the maximum a posteriori (MAP) estimate, $\theta_a^{MAP}$, which is the mode of the posterior marginal likelihood. The second estimate for $\theta_a$ is the mean of the posterior marginal likelihood, $\theta_a^{Mean}$. This estimation approach was used for all GPs.
Figure 6: Estimation of $\theta^{MAP}$ and $\theta^{Mean}$ given training points $(y, x)$. An initial random sample of 5000 values for $\theta$ were drawn (according to Table 5), and their respective posterior marginal likelihoods, $p(y|x, \theta)$, were calculated (far left). From the initial sample, the 100 $\theta$ values with the largest associated marginal likelihood were identified (second from left). These 100 $\theta$ values each formed the centre for 50 additional samples (for 5000 total) focussed around the specific $\theta$ (according to Table 6). The marginal likelihoods of these focussed samples were calculated (second from right). The combined sample of 10000 values of $\theta$ comprised the initial and focussed samples, and their respective marginal likelihoods (right columns). $\theta^{MAP}$ was estimated to be the $\theta$ within the combined sample with the highest marginal likelihood. In practice, $\theta^{MAP}$ was always a member of the focussed sample. $\theta^{Mean}$ was estimated by scaling the combined sample’s marginal likelihoods to sum to 1, and taking the dot product of the combined sample $\theta$ values and the scaled marginal likelihoods.

discussed in the report.

HR Forecasting Loss Functions Three loss functions will be used to quantify the loss associated with using $p(y_*|y) \sim N(\bar{y}_*, \Sigma)$ (from Eq. 1) to predict unseen testing points, $y_*:

\[ L_{RMSE}\{p(y_*|y)\} = \sqrt{(y_* - \bar{y}_*)^T(y_* - \bar{y}_*)/n}. \]  

(15)

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$h_1$, $h_2$, $h_3$</th>
<th>$\sigma_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_0$</td>
<td>$U(1, 150)$</td>
<td>-</td>
<td>-</td>
<td>$U(0, 0.25)$</td>
<td>$U(0, 0.125)$</td>
</tr>
<tr>
<td>$k_1$, $k_2$, $k_3$</td>
<td>$U(1, 60)$</td>
<td>$U(40, 150)$</td>
<td>-</td>
<td>$U(0, 0.25)$</td>
<td>$U(0, 0.125)$</td>
</tr>
<tr>
<td>$k_4$</td>
<td>$U(1, 15)$</td>
<td>$U(15, 60)$</td>
<td>$U(60, 150)$</td>
<td>$U(0, 0.25)$</td>
<td>$U(0, 0.125)$</td>
</tr>
</tbody>
</table>

Table 5: Distributions for the initial sample of $\theta$. The initial sample of $\theta$, described in Fig. 6, was drawn from the above uniform distributions. With a larger number of constituent kernels, the length-scales ($\lambda$) become more precisely defined. Each signal-variance ($h$) and noise-variance ($\sigma_n$) was sampled from an identical initial distribution, regardless of the kernel.
Table 6: Distributions for the focused sample of $\theta$. For each of the 100 highest-likelihood $\theta$ of the initial sample (Fig. 6), 50 focused samples were drawn by adding the above uniformly distributed perturbations to $\theta$. For example, for $\theta_0 = [h, \lambda, \sigma_n]$, each of the 50 focused samples would be drawn from $[U(h - 0.01, h + 0.01), U(\lambda - 2.5, \lambda + 2.5), U(\sigma_n - 0.01, \sigma_n + 0.01)]$. If a sampled hyperparameter value was less than 0, the hyperparameter was assigned its absolute value.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$h_1$, $h_2$, $h_3$</th>
<th>$\sigma_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_0$</td>
<td>$U(-2.5, 2.5)$</td>
<td>-</td>
<td>-</td>
<td>$U(-0.01, 0.01)$</td>
<td>$U(-0.01, 0.01)$</td>
</tr>
<tr>
<td>$k_1$, $k_2$, $k_3$</td>
<td>$U(-1, 1)$</td>
<td>$U(-5.5)$</td>
<td>-</td>
<td>$U(-0.01, 0.01)$</td>
<td>$U(-0.01, 0.01)$</td>
</tr>
<tr>
<td>$k_4$</td>
<td>$U(-1, 1)$</td>
<td>$U(-2.5, 2.5)$</td>
<td>$U(-5.5)$</td>
<td>$U(-0.01, 0.01)$</td>
<td>$U(-0.01, 0.01)$</td>
</tr>
</tbody>
</table>

Figure 7: Four kernels fit to the same HR time-series. The rational quadratic kernel is not used in this project, except to illustrate the effect of smoothing across the length-scales of many SE kernels. Each kernel used a $\theta^{MAP}$ estimate of its respective hyperparameters. At approximately 9.1 hours, the patient has a (potentially artefactual) elevation in HR. The rational quadratic and squared-exponential kernels’ $\theta^{MAP}$ estimates are the most responsive to these elevated values.
Candidate $\theta_1$ Fits and Forecasts

$\theta_1^{\text{Mean}}$ Fit and Forecast

Figure 8: Fits and Forecasts using random $\theta_1$’s, $\theta_1^{\text{MAP}}$, or $\theta_1^{\text{Mean}}$. The low likelihood red GP fits detail to excess and is unable to forecast because the posterior mean immediately returns to the prior mean. The green and blue GPs correctly identify much of the HR volatility as signal-noise and, accordingly, are much smoother. The blue GP has higher LML as the result of better fitting the (potentially artefactual) elevated HR around 9.1 hours. Had the blue GP been used as the $\theta_1^{\text{MAP}}$ fit, it would have been less influenced by the elevated HR measurements than the $\theta_1^{\text{Mean}}$ GP in black. This suggests that the posterior likelihood of $\theta_1$ is more heavily weighted toward a small $\lambda_1$, and fitting detail. This preference for fitting detail is why the $\theta_1^{\text{Mean}}$ forecast has a modest dip at 10 hours, despite the general upward trend of the HR measurements.

\[
L_{\log \text{Lik}} \{ p(y_*|y) \} = \left[ -\frac{1}{2} \log(|\Sigma|) - (y_* - \bar{y}_*)^T \Sigma^{-1} (y_* - \bar{y}_*) - n \log(\pi) \right] / n. \quad (16)
\]

\[
L_{\text{MAPE}} \{ p(y_*|y) \} = 1 \cdot \frac{| \bar{y}_* - y_* |}{y_*}^T / n, \quad (17)
\]

where $1$ is a vector of ones and “•” denotes a dot product. The motivation of using $L_{\text{RMSE}}$ is to capture the difference between what the GPR forecast and the actual measurement values (of the bedside monitor). The motivation of using $L_{\text{MAPE}}$ is similar, but accounts for the magnitude of the observed values. Notably, neither $L_{\text{RMSE}}$ nor $L_{\text{MAPE}}$ incorporate uncertainty around the predicted latent function. $L_{\log \text{Lik}}$ incorporates this uncertainty, and accounts for how well the posterior distribution describes the unseen points. $L_{\log \text{Lik}}$ penalises a prediction for being overly certain when it is inaccurate as well as being under-uncertain where the prediction is accurate.

Experimental Design Decisions for Comparing the Candidate Covariance Kernels

This section describes the process for selecting patients and time-series segments that will be
used to test and validate the forecasting performance of each candidate kernel.

UPMC patients were allocated to one of four mutually-exclusive groups, named “Training Group 1” ($G^1$), “Training Group 2” ($G^2$), “Non-C-event Testing Group” ($G^3$), and “C-event Testing Group” ($G^4$). The purpose of $G^1$ and $G^2$ is to allow a large amount of exploration in the UPMC data without over-fitting to negative and positive examples. Group $G^4$ comprised each of the 59 UPMC patients with an identified C-event. Group $G^3$ was chosen by iteratively taking each C-patient (ordered by patient index) and selecting the Non-C-patient whose LOS was closest in absolute value. To create $G^1$ and $G^2$, the remaining 206 Non-C-event patients (with more than 2 hours of HR data) were sorted according to LOS. The patients with odd patient identification numbers were assigned to $G^1$ and the patients with even numbers were assigned to $G^2$. Visual inspection showed that the LOS distributions of $G^1$ and $G^2$ were very similar.

Viable Training Windows for Sequential Prediction $G^1$ provided the patients to see which kernels would outperform $k_0$. To ensure that training and testing windows were sufficiently comparable, a several rules were devised requiring that (1) training windows had sufficient data to make a reliable prediction, (2) testing windows had sufficient data to make a reliable statement about forecasting error, and (3) testing points were sufficiently close to the final training points so that no windows were at a systematic disadvantage. Training windows were sampled from patients as described in Fig. 9, subject to criteria described in Table 7.

Identification of Optimal Kernel Choice in Forecasting of Patient HR Data The seven GPR candidate kernels ($k_0, ..., k_6$), and the two simple predictors ($k_{-1}$ and $k_{-2}$) were assessed on data from 30 patients in group $G^1$. As described in Table 7, 300 training windows were assessed, with no more than 15 windows drawn from a single patient. To assess the value of accounting for the uncertainty in $\theta_a$, both $\theta^\text{MAP}_a$ and $\theta^\text{Mean}_a$ forecasts were included. Testing windows of length 15 minutes to 120 minutes were examined.

Fig. 10(a) is an example of the plots comparing kernel performance, in this case $L_{\text{logLik}}$ for a forecast window of 75 minutes. For each candidate kernel $k_a$, the performance of $\theta^\text{MAP}_a$ is shown the the left, followed by the performance of $\theta^\text{Mean}_a$. The most pronounced general trend is that $\theta^\text{Mean}_a$ forecasts are much more robust, as shown by their narrower 1st-3rd quartile
### Table 7: Training window requirements for each project task.

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients in $G^1$</td>
<td>30</td>
<td>0</td>
<td>103</td>
<td>0</td>
<td>103</td>
</tr>
<tr>
<td>Number of patients in $G^2$</td>
<td>0</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>103</td>
</tr>
<tr>
<td>Number of patients in $G^3$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>Number of patients in $G^4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>Training windows per patient</td>
<td>≤ 15</td>
<td>≤ 60</td>
<td>≤ 60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total training windows</td>
<td>300</td>
<td>900</td>
<td>~4k</td>
<td>~3k</td>
<td>~13k</td>
</tr>
<tr>
<td>Training window length</td>
<td>1-5 hrs</td>
<td>1-5 hrs</td>
<td>1-5 hrs</td>
<td>0.25-5 hrs</td>
<td>0.25-5 hrs</td>
</tr>
<tr>
<td>Last hour of training window</td>
<td>≥ 15 obs</td>
<td>≥ 15 obs</td>
<td>≥ 15 obs</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Last 15 min of training window</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>≥ 3 obs</td>
<td>≥ 3 obs</td>
</tr>
<tr>
<td>Time between training GP fits</td>
<td>1 hour</td>
<td>1 hour</td>
<td>1 hour</td>
<td>15 min</td>
<td>15 min</td>
</tr>
<tr>
<td>Time between last $x_<em>$ and first $x_</em>$</td>
<td>≤ 15 min</td>
<td>≤ 15 min</td>
<td>≤ 15 min</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Test window length</td>
<td>≤ 2 hrs</td>
<td>≤ 2 hrs</td>
<td>≤ 2 hrs</td>
<td>≤ 2 hrs</td>
<td>-</td>
</tr>
<tr>
<td>Test window</td>
<td>≥ 15 obs</td>
<td>≥ 15 obs</td>
<td>≥ 15 obs</td>
<td>≥ 2 obs</td>
<td>-</td>
</tr>
</tbody>
</table>

Task 1 is to find the optimal kernel for robust forecasting. Task 2 is the validation of the chosen optimal kernel over $k_0$. Task 3 is to find an optimal prior over $\theta_0$. Task 4 is the step-change detection GPR (described in chapter 5). Task 5 is the time-series matching GPR. The viability of a testing window differed according to the task under consideration. For example, tasks 1-3 (to assess forecast accuracy) were more stringent in selecting observation-rich training and test windows to maximize comparability of performance metrics between windows. Task 2 is intended to validate the results of task 1, so an similar number of patients was chosen, but with more many training windows per patient to improve the precision of performance difference between $k_0$ and the proposed $k_{opt}$. Tasks 4 and 5 place a premium on fitting GPs frequently to avoid missing potential alarms. Tasks 4 and 5 were therefore less stringent in their standards. Task 5 requires a no testing windows to generate alarms and therefore does not take those attributes into account.
Figure 9: Sequential training and testing windows for patient time-series. The viable training and testing windows are shown for four hypothetical patients, according to the selection criteria of Task 1, 2, and 3 in Table 7. Each block represents a 1-hour window of patient data. Inside each block is the number of measurements within that 1-hour window. Patient A represents the "ideal" time-series with no missing measurements across all 9 hours. Each new training window is shifted forward one hour from the previous training window, encompassing up to 5 hours of training data. The testing window comprises any measurements up to 2 hours following the training window. In Patient B’s data, the sparsity of hours 3-5 disallows those windows from being used to assess forecast accuracy, as does the sparsity of hours 1, 3, and 5 of Patient C. For patient D, only hours 4 and 9 contain sufficient data for a training set, but the subsequent testing windows are too sparse. Therefore, Patient D would not contribute to the performance metrics in the results section. For Task 1, a patient with more than 15 viable training windows would have 15 selected at random for use in the results.

range, and less heavy tails. The added robustness from choosing $\theta_a^{Mean}$ over $\theta_a^{MAP}$ indicates that more principled methods of estimating the full posterior mean is likely to further improve the robustness of forecasts. Each candidate kernel $k_1, ..., k_6$ appears to produce preferable forecasts over baseline comparator $k_0$, including the simple predictors $k_{-1}$ and $k_{-2}$. In particular, $k_6(\theta_6^{Mean})$ had a median and first quartile $L_{logLik}$ of 1.19, and 0.57, respectively. In comparison, $k_0(\theta_0^{MAP})$’s median and first quartile were 0.74 and -3.39, respectively, and $k_0(\theta_0^{Mean})$’s median and first quartile were 0.85 and -1.65, respectively. So the additional complexity of $k_6$ is warranted for a more robust forecast. Predictors $k_{-1}$ and $k_{-2}$ also tended to outperform $k_0$, with respective medians and first quartiles of 1.18 and 0.10, and 1.21 and -0.09.

There was much less distinction between the candidate GPR kernels for $L_{RMSE}$ and $L_{MAPE}$ for a 75 minute training window. For example, median RMSE performances of $\theta_a^{MAP}$ and $\theta_a^{Mean}$ were about 5.1 for $k_{-1}, k_1, ..., k_6$, 5.5 for $k_0$, and 5.7 for $k_{-2}$. The relative performances of the candidate kernels varied slightly depending on the loss function and testing window. $k_6$ demonstrated robust and consistent performance or a variety of window lengths and therefore was chosen to be the candidate for $k_{opt}$ for validation in $G^2$. Since $k_1$ was the most robust and consistent performer amongst the candidates with three constituent kernels, and it was assessed
Figure 10: Optimal kernel $k_a$ for robust HR forecasts over $k_0$. The performance of 7 candidate GP kernels ($k_0, ..., k_6$) were assessed on 300 training windows in group $G^1$. $k_1$ was determined to be the most robust candidate with three constituent kernels. $k_6$ was determined to be the most robust candidate overall. $k_1$ and $k_6$ were assessed on a further 900 training windows in group $G^2$, along with $k_0$, $k_{-1}$, and $k_{-2}$. The results demonstrated that the added complexity of $k_1$ and $k_6$, along with the use of $\theta_a^{\text{Mean}}$ over $\theta_a^{\text{MAP}}$ could significantly improve the robustness of a GPs forecast of HR measurements.

Validation of Optimal Kernel Choice in Forecasting of Patient HR Data  To validate the choice of $k_6$ and $k_1$ as superior kernels, a new set of 900 training windows was drawn from 35 patients in $G^2$ (as described in Table 7 and Fig. 9). The number of training windows was increased in order to more precisely quantify the difference in performance over baseline comparator $k_0$. The results in $G^2$, shown in Fig.10(b) suggested that $k_6(\theta_6^{\text{Mean}})$ is more robust (compared to $k_{-2}$, $k_{-1}$, and $k_0$) than originally suggested by the performance in $G^1$: $k_6(\theta_6^{\text{Mean}})$ and $k_1(\theta_1^{\text{Mean}})$ both had a median $L_{\log\text{Lik}}$ of 1.07, compared to 0.80, 1.16, and 1.10 of $k_0(\theta_0^{\text{MAP}})$, $k_{-1}$, and $k_{-2}$ respectively. The first decile performance was 0.02 for $k_6(\theta_6^{\text{Mean}})$, compared to -0.14, -4.11, -1.10, and -0.76 for $k_1(\theta_1^{\text{Mean}})$, $k_0(\theta_0^{\text{MAP}})$, $k_{-1}$, and $k_{-2}$ respectively. $k_6(\theta_6^{\text{Mean}})$ demonstrated that its usual performance was comparable or superior to other forecasting methods, but was much more robust against extremely poor forecasts. For $L_{\text{RMSE}}$ and $L_{\text{MAPE}}$, the performance advantage of $k_1$ and $k_6$ appeared about the same (i.e. a relatively small improvement over $k_0$) as already seen in group $G^1$.

The superior performance of the more complex covariance kernels demonstrates the importance (and benefit) of building a kernel around an understanding of the patient’s physiology. The drastic improvement to robustness caused by using $\theta_a^{\text{Mean}}$ over $\theta_a^{\text{MAP}}$ also shows the importance of managing the uncertainty around the estimates for the hyperparameters.
4.4 Regularising Priors in Forecasting of Patient HR Data

A GP's forecasting performance is heavily influenced by the length-scale(s) describing the sources of variance. This effect is especially acute for kernels with a small number of kernels, such as $k_0$: if $\lambda_1^{\theta_0}$ is too short, forecasts will quickly regress to the training mean, missing clearly-visible trends. If $\lambda_1^{\theta_0}$ is too long, it will be insensitive to any short-term dynamics immediately prior to the beginning of the forecast interval. Although some regularisation is provided by optimisation of LML, the encoding of prior knowledge of the values of the hyperparameters may provide improved forecasting performance. We propose regularising priors over two aspects of $\theta_a$: (i) the characteristic length-scales, and (ii) the signal-to-noise ratio (SNR). The latter is $\frac{\sum_{i=1}^{I} h_i}{\sigma_n}$ of $\theta_a$ where $I$ is the number of constituent kernels in $k_a$. Regularised the length-scales is physiologically intuitive, since the priors relate directly to concepts of the rapidity and persistence of HR trends.

A log-normal$(\mu, \sigma)$ prior\textsuperscript{d} is proposed for lengths-scales and SNR. The log-normal has support on on the real positive line, and there are easy ways to modulate its mode, mean, and spread. The distribution tends to have a heavy right-tail, which may be helpful in describing our lack of preference between, for example, models with a high SNR. The plausible range of “desirable” length-scales and SNRs is very diffuse and confounded by several factors. Accordingly, a wide range of regularising priors could be hypothesised. These different hypothesis-priors can have vastly different areas of high prior-likelihood. For an initial investigation, we will begin with a single prior applied to a single candidate kernel. $k_0(\theta_0^{MAP})$ is a good starting point because its forecasting robustness is the most dependent on the choice of $\lambda_1$. The forecasting improvement of $k_0(\theta_0^{Mean})$ over $k_0(\theta_0^{MAP})$ suggests that additional kernels is not strictly necessary to make $k_0$ more robust.

To assess where a prior may be helpful, the performance of $k_0$ in $G^1$ (described above) were examined as functions of the length-scale ($\lambda_1$), and the SNR ($\frac{h_1}{\sigma_n}$) as seen in Fig. 11. An improvement was seen by a tighter clustering of estimated $\lambda_1$ values in the 20-30 minute range. So we hypothesise that priors that shift posterior likelihood towards $\lambda_1 \in [20, 30]$ will outperform uninformative priors, or priors that shift likelihood towards other regions.

\textsuperscript{d}This hyperparameter $\sigma$ is distinct from the $\sigma_n$ hyperparameter of the white-noise kernel and the candidate kernels.
Experimental Design Decisions for Comparing the Candidate Regularising Priors

To investigate the performance of a large number of priors, viable training windows were selected according to Table 7 and Fig. 9. For each viable training window, the sampling scheme to fit a GP (described in Fig. 6) was modified to provide a single, thorough sampling scheme to assess the forecasting capacity of all candidate priors \( p_a(\theta_0) \). The modified sampling scheme is described in Fig. 12.

The proposed family of log-normal priors had a mean and median between 2.5 minutes and 45 minutes. Fig. 13 shows the iso-mean, iso-mode, and entropy lines of the proposed family of priors. As the mean and median of the priors approach each other, the prior becomes more peaked and therefore, more influential over the posterior likelihood. The yellow lines on Fig. 13 mark the entropy bounds of 2.5 and 3.5. This is approximately where the priors are neither “too peaked” nor “too flat.” A prior that was “too peaked” would tend to ignore the information conveyed by the data through the model-likelihood, whereas a prior that was too flat, would convey little prior knowledge. We hypothesise that priors that peak in the 20-30 minute range will demonstrate superior performances, in terms of the loss functions. As the priors diverge from this range, the performance will diminish.

Results Optimal Prior over \( \theta_0^{MAP} \) The grid of priors forecasting performances in Fig. 14 suggest that there is an area amongst the log-normal priors with a small, but consistent improvement in median \( L_{RMSE} \) across the training windows. Toward the centre of this area, \((\mu, \sigma) \approx (3.5, 0.27)\), the standard deviation of \( L_{RMSE} \) also small, suggesting that \( L_{RMSE} \) is consistently lower as well. In this region, \( L_{logLik} \) of the forecasts is higher, but not at a local
Figure 12: Sampling scheme to simultaneously assess all candidate priors $p_a(\theta_0)$. The goal of the revised sampling scheme was to avoid redundant calculations when fitting GPs to the same training window using different priors. In particular, the computation burden of inverting the covariance matrix $K$ for each sampled value of $\theta_0$. The modified sampling scheme began as in Fig. 6: an initial random sample of 5000 values for $\theta_0$ were drawn (according to Table 5), and their respective posterior marginal likelihoods, $p(y|x, \theta_0)$, were calculated (far left). The initial sample was partitioned into three subsets according to the value of $\lambda_{\theta_0}$. From each partition the 100 $\theta_0$ values with the largest associated marginal likelihood were identified (second from left). Each of these 300 $\theta$ values formed the centre for 50 additional samples (for 15000 total) focussed around the specific $\theta$ (according to Table 6). The marginal likelihoods of these focussed samples were calculated (second from right). The purpose of this partitioning was to ensure that areas of high prior likelihood, $p_a(\theta_0)$, would be thoroughly sampled, regardless of which areas had high model likelihood. The combined sample of 20000 values of $\theta_0$ comprised the initial and focussed samples, and their respective marginal likelihoods (second from right columns). For any candidate prior, $p_a(\theta_0)$, the prior likelihood was calculated for each in the $\theta_0$ combined sample (orange column, second from right). $\theta_0^{MAP}$ was the value of $\theta_0$ with the largest posterior likelihood, $p(y|x, \theta)p_a(\theta)$. The associated forecast $p(y_*|y, \theta^{MAP})$ was $p_a(\theta_0)$’s forecast for that training window.
optimum as is the case for $L_{RMSE}$. The coordinate of $(\mu, \sigma) \approx (3.5, 0.27)$ for a log-normal distribution roughly corresponds to a mode of 30.8 minutes, and a mean of 34.3 minutes. This is close to the “preferable” region for $\lambda_1$, as estimated from the plot of $\lambda_1$ against log-likelihood performance. This is initial evidence that encouraging fitted values towards regions known to have successful forecasts may be a fruitful approach to improving forecasts. Further work will be needed before determining if this improvement is robust, ultimately, in the presence of both Non-C”-patients and C”-patients.

5 Methods and Results for Deterioration Detection

(1) Univariate Thresholding Univariate thresholding is a method currently in clinical use. Thresholding is the approach of setting an explicit threshold value that alarms if a patient’s
vital sign exceeds the threshold. Crossing this threshold is called an “exceedance” and can capture deviation in either a negative or positive direction. The univariate threshold presented here will be the upper-threshold on HR, which accounts for more episodes of patient patient deterioration than a lower-bound threshold [18].

(2) Kernel Density Estimate The KDE approach [18] represents the current state-of-the-art in patient monitoring. The model of normality was created using all data for $G^1$ and $G^2$ patients. Missing values were imputed using a capture-and-hold method that imputed missing measurements to be the most recent measurement, holding 30 minutes for SBP and DBP, and 5 minutes for HR, BR, and SpO$_2$. Following the design process for “Visensia”, K-means clustering found the location of 500 clusters in the training set. The mean location of the 500 cluster-centroids was denoted the “centroid mean.” The 400 centroids closest to the centroid mean were kept to form the KDE. Denoting $x$ to be a data point of all 5 vital signs, and $x_i$ to be the centroid of cluster $i = 1...400$, the density estimate was given by

$$p(x) = \frac{1}{400(2\pi)^{5/2} \sigma^5} \sum_{i=1}^{400} \exp \left( -\frac{|x - x_i|^2}{2\sigma} \right),$$

(18)

where $\sigma$ is an estimate of the mean variance around each cluster-centroid. As described by Hann, who in turn referenced Bishop [? ], the variance around each cluster-centroid was calculated by finding the 10 closest observations (in Euclidean distance) to each cluster-centroid. The mean Euclidean distance between the centroid and the 10 closest observations was considered the variance around the centroid. The mean of all 400 centroid variances was $\sigma$. The novelty score for a newly observed 5-dimensional data point is the negative log-likelihood:

$$z(x) = -\ln (p(x)).$$

(19)

The Novelty Scores for data in $G^3$ and $G^4$ patient’s time-series was calculated. A threshold $\kappa$ on $z(x)$ the KDE Novelty could then be defined, with alerts raised if $z(x) \geq \kappa$.

(3) GPR step-change detection The term “step-change” describes a marked discontinuity of new observations from previous observations. This concept has a useful application in novelty detection: if previous observations are the product of a “normal” generative process, then a
departure from the expectations of the previous observations could be viewed as a step-change from “normality.” As illustrated in Fig. 15, the GPR step-change detector takes a patient’s current time-series, which is assumed to represent “normality” for that patient, and forecast future measurements. If the observed measurements deviate sufficiently from the forecast, then an alarm can be sounded. This method is a threshold on forecast-error, with the attractive hypothesis that a deteriorating patient’s time-series may be more erratic and, hence, more difficult to forecast accurately. This approach has several interesting challenges: no alarm would sound for any patient that is “predictably abnormal” (e.g. a patient steadily maintaining a HR of 200 bpm.) Furthermore, there is the trade-off between collecting more data (over a longer forecast window) to more precisely quantify a step-change, and a reduced response time (of a shorter forecast window.)

(4) GPR Time-Series Matching The GPR time-series matching method attempts to perform inference on the dynamics of a newly fit time-series by comparing it to time-series of patients determined to have been healthy. A GP using kernel \( k_1 \) is fitted to the training as illustrated in Fig. 15 and described in Table 7. Once \( \theta_1 \) is fit, the values of the previous hour, \( y_p \), are forecasted at sixty 1-minute intervals. Any forecasted time-stamp earlier than the first training observation is removed from the fit. The product is a multivariate Gaussian, \( p(y_p | y) \sim N(\bar{y}_p, \Sigma_p) \) of up-to 60 dimensions. The novelty of this multivariate Gaussian is the minimum KL-divergence between \( N(\bar{y}_p, \Sigma_p) \) and any multivariate Gaussian, \( N(\bar{y}_t, \Sigma_t) \), from a dictionary of 10,000 reference patterns. The reference patterns come from the time-series of patients in groups \( G^1 \) and \( G^2 \), fitted in an identical fashion. A difference in the dimensions of the multivariate Gaussians uses the longest possible alignment of each segment’s most recent values. A particularly useful refinement to the “novelty” \( \text{min-KL}(N(\bar{y}_p, \Sigma_p), N(\bar{y}_t, \Sigma_t)) \) is to restrict the training patterns to only be those within a few hours of “time since admission” as the testing segment. The \( k^{th} \) order statistic of KL-divergence can also replace the minimum to strengthen the signal against abnormal “normal” training patterns. By choosing an upper threshold on KL, a testing segment’s deviation from any of the observed healthy segments can be used to alarm deviation from normality.
Figure 15: Sequential training and testing windows for deterioration detection. The viable training and testing windows are shown hypothetical Patient E, according to the selection criteria of Tasks 4 and 5 in Table 7. Each block represents a 15-minute window of patient data. Inside each block is the number of measurements within that 15-minute window. For step-change detection, the viability of a training window depends on having at least 2 observations within the pre-defined test window. In this the step-change detector using a 60 minute forecast use of a prediction from Patient E’s fourth data block, whereas the step-change detector using a 15 minute forecast cannot. Once the forecast accuracy is known, a step-change alarm is made (bottom left) depending on whether there is little deviance (“No Step-change”) or significant deviance (“Likely Step-change”). Since time-series matching requires no testing window, it can make the most predictions using Patient E’s time-series: Only windows with 1 or 0 observations are prevented from generating a time-series matching decision. Any viable training window (orange) is fitted by GPR to create a new pattern. Each new pattern is compared each member in a dictionary of healthy patterns and the KL divergence is recorded. New patterns with high KL divergence from all healthy pattern are likely to represent deterioration, and an alarm should be sounded. The dictionary of healthy patterns is created by applying the criteria of task 5 in Table 7 to each patient in groups $G^1$ and $G^2$. 
Figure 16: Order Statistics 1-10 for C"-Patient and a Non-C"-Patient. The ten lowest KL values for the C"-patient are much higher and more tightly distributed. The KL values also become more divergent as the C"-event approaches.

Figure 17: Timeline for TEW of patient deterioration. The TEW was calculated for each C"-Patient in group $G^4$. The timestamp of the first C"-event is centred at 0, and the credible period for a clinical alarm was from 8 hours prior to the C"-event until 2 hours following the C"-event. Any clinical alarm outside of this credible region (e.g. Alarm A or D) was counted neither as an early warning nor a false positive. The TEW for a patient was the first alarm to occur within the credible period (e.g. If Alarms B and C both occurred, the TEW would be 4 hours). If no clinical alarm occurred in the 10-hour credible region, then the alarm time was right-censored at -2.

**Evaluations** The method of evaluation must characterise the trade-off between time of early warning (TEW) and false positive rate (FPR) of each early detection method. The calculation for TEW is described in Fig. 17. The FPR for each method was the total number of clinical alarms sounded on the Non-C"-Patients of group $G^3$, divided by the total number of predictions on $G^3$ patients.

**Results of GPR step-change detection** The step-change detector’s performance was robust for forecast windows of 15-45 minutes. Longer forecast windows were likely hindered by delayed identification of a step-change and the reduced accuracy of $k_1(\theta_1^{MAP})$ and $k_1(\theta_1^{Mean})$ at the end of the forecast window. An approach that combined thresholds at multiple test window lengths did not show significant improvement over the single window-length step-change de-
GPR Step-change with optimised forecast window

Figure 18: Sensitivity of Step-change Detection Performance to choice in forecast error metric forecast. If multiple forecast window lengths were used, the univariate GPR converged in performance towards the KDE detector.

Detector: $k_1(\theta_1^{MAP})$ forecasts detecting step-changes in log-likelihood showed a large upward-shift in its 25th quantile (i.e. the C4-patients who have the least amount of advanced warning now have an advanced warning more in-line with their “peers”). For most patients, this would have meant an early warning less than an hour later than the same patient would have received when monitored by the KDE. While promising (especially in light of being a univariable detector) it is unclear whether the results are sufficiently robust given the variety of discrete design decisions required to combine thresholds at different test window lengths. Figure 18 shows the performance of different step-change detectors when varying $k_1(\theta_1^{MAP})$ to threshold on $L_{RMSE}$ and $L_{LogLik}$ when attempting to optimise over multiple forecast window lengths.

These results are encouraging for step-change detection’s utility in deterioration detection. Step-change detection was also attempted using $k_6(\theta_6^{Mean})$ and $k_6(\theta_6^{MAP})$, but performance was only marginally better than univariate thresholding. A possible explanation for this is that $k_6$ and better able to model the non-smooth short-term noise that would generate an alarm in $k_1$. A potential avenue for improvement is to use $k_4(\theta_4^{Mean})$ or $k_4(\theta_4^{MAP})$, which has a demonstrably better forecasting ability and $k_1$, as seen in Fig. 10(a), but still stipulates a smooth latent function. The biggest improvement to step-change detection is likely to be the introduction of EVT metrics, which can incorporate the knowledge of extreme outliers, adjusted for the number of new points under observation. This could allow for a more principled manner by which to consider across multiple test window lengths simultaneously.
**Time-Series Matching** The proposed method of time-series matching showed improvement over both the Univariate Threshold and the KDE approach. As with step-change detection, time-series matching has a variety of design decisions to which performance may be sensitive, in particular, the choice of reference patterns included in the dictionary for each new patient. In the absence of any filtering of the reference patterns, performance was very poor. Only reference patterns with 60 or 61 values. The most important design decision seemed to be closely matching time of stay between the new segments to the reference pattern: patient vital signs are usually at their most unstable at the beginning of their stay. Over time, their vital signs should stabilise, as should the vital signs of the reference patterns. In this case, it was very important to perform some type of filtering based on the current time of stay. TEW vs. FPR was best when allowing a 1-hour-maximum difference between the new segments and the reference segments, but remained robust between 1-5 hours.

Performance was robust against the more metric-focused decisions, such as order of the KL divergence inputs, and whether to scale the KL divergence by the dimension of testing segment. A final consideration was the order statistic of the KL divergence. With over 10,600 reference patterns, some number will exhibit abnormal physiology. (Consider the Non-C"-patient from Fig. 1.) The choice in order statistics would encode the belief that there are a small number of these reference patterns.

One cause for scepticism is that many of the C"-patients had the most extreme KL values towards the beginning of their 8-hour credible warning window. The values then became less-extreme as they moved towards the C"-event, and then became more extreme again as they came very near to the C"-event. While it is reassuring that the KL metrics become more extreme when the testing segments are very near to the C"-event, a good alarm should be one that sounds consistently.

**Conclusions** A comparison of the approximate standings of each deterioration detection method is in Fig. 19. Each method has many areas for improvement. The most interesting aspect of the current results is that each GPR method is univariate, compared to the KDE, which has the advantage of 4 additional vital signs and the context of their interrelationship. The GPR’s advantage is that it can incorporate the interrelation of the HR measurements.

---

5The KL-divergence of two multivariate normals is additive in the case of independence.
with themselves, thereby extract fuller the information contained within the time-series. Ideas for more fully exploiting the advantages of the time-series approach will be described in the proposal for future work.

6 Proposed Future Work

6.1 Pertaining to Forecast Accuracy and Early Warning of Deterioration

The highest priority of future work is to improve the GPR models to more realistically describe the underlying physiological and measurement processes. Part of this improvement will be a more rigorous grasp of the uncertainty in the model itself. These improvements are useful to both clinical goals. The items described below are not a collection of disjoint tasks, but instead intended to build on each other for a more technically and clinically thorough DPhil project.

**Priority 1: Bayesian Monte Carlo**  The rudimentary estimate of posterior mean, \( \theta_{a_{\text{Mean}}} \), was at least as important (for robust forecasting) as the proper choice of kernel. A more sophisticated method to explore the full posterior likelihood is likely to improve robustness. MCMC approaches are popular but have well-documented drawbacks [39] that Bayesian Monte Carlo (BMC) [40] works to address. A particular example (which will only become more prevalent in fitting MTGPs) of putting BMC to use is to handle a multimodal posterior where the lower-likelihood mode is a heavy tail: whereas the current \( \theta_{a_{\text{Mean}}} \) might fall in a low-likelihood region between modes, BMC may help decide between highly-peaked regions and
Figure 20: Gantt chart of proposed future work. The next 4 months’ work will investigate each univariate vital sign, along with the external validation of any successful univariate methods. Following this, MTGPs will be developed for both forecasting and deterioration detection. Any successful MTGP approaches will be validated subsequently on external data sets.
less-peaked regions that cover more total area. It is possible that the additional computational
burden of BMC might equal or exceed the effort saved by a more efficient search/characterisation
of the hyper parameter space. This is the first priority shown in the Gantt chart because
performance gains from BMC (or MCMC) may have important implications for subsequent
modelling priorities (e.g. the necessity of priors over $\lambda$).

**Priority 2: Multi-task Gaussian Processes with Remaining vital signs** The large
automated framework to explore GP modelling of HR will facilitate a faster exploration of
preferable kernels and priors for the remaining vital signs. Following HR, the order of priority
is $\text{SpO}_2$, BR, SBP, and DBP. It is unrealistic, though, to model these as independent processes.
An MTGP approach is the obvious step to model these correlated time-series.

A thorough understanding of the dynamics of each univariate GP is essential to reliably
fitting MTGPs. Likely challenges of modelling the remaining vital signs will be to account for
different sampling rates (e.g. SBP and DBP are measured every 30 minutes, so a kernel for
short-term detail may be inappropriate). A particularly obvious omission from the candidate
kernels considered earlier is periodicity. Many patients had insufficient data to warrant the
use of a periodic kernel. The value of additional vital signs is several-fold: first, the trends of
individual vital signs are contextualised with the preceding or contemporaneous trends among
the other signals. For example, $k_1$ can capture long-term multi-hour trends augmented
by shorter-term trends only if the long-term trends continue in an ascending or descending
direction. Additional vital signs may anticipate when HR is likely to experience a change in
the direction of a long-term trend. Due to the measurement noise inherent in HR, less volatile
vital signs might to help distinguish upward HR movements that are noise from those that
are significant trends. The correlation structure in physiological signals has implications for
novelty detection. In particular, both time-series matching and step-change detection would
both be sensitive to deviation from the established correlation structure of healthy patients. For
time-series matching, additional vital signs could also reduce the larger posterior uncertainty in
sparsely-measured regions. For step-change detection, deviant values may be easier to detect
due to greater precision and accuracy forecast in the forecasted latent function.
**Priority 3: Reducing Computational Burden** In practice, once an optimal method of fitting GP’s has been found, more frequent analysis will be managed by Cholesky decomposition to perform sequential updates between larger, complete updates. Further improvement could be made taking advantage of the diagonal-constant (band-diagonal) structure of Toeplitz matrices. In scenarios with frequent and constantly measured signals, this could be performed with few (or no) simplifying assumptions. Fitting to multiple hours of data will involve thousands of data points, so approaches such as “inducing points” might be necessary to handle the computational burden.

**Priority 4: Validation on External Data Sets** External validation is a sensible precaution, due to the inductive nature of putting retrospective research into clinical practice. Any methods that are proven successful with data from UPMC patients must be validated on new data sets whose patients are in a different, but related clinical setting. There are several related data sets available, notably the post-operative patients in CALMS-2 (and subsequent CALMS studies) and a data set of over 200,000 in-patients from Portsmouth Hospital, UK.

**6.2 Pertaining Only to Forecast Accuracy**

**Priority 1: Forecasting in the Presence of Normal and Abnormal Physiology** A key consideration in forecasting is robustness in the presence of both normal and abnormal data. The current forecasting work has been entirely performed using data from Non-C"-patients. Although some of these patients surely exhibited periods of unusual physiology there is no reason to believe that the most robust priors for Non-C"-patients and also the most robust priors for C"-patients. Future work must bridge the gap from forecasting for Non-C"-patients to forecasting for C"- and Non-C"-patients. These differences may provide further insight to distinguish between C" (unstable) and Non-C" (stable) physiology.

In light of the results described earlier, when working with MTGPs, priors over the within-task kernels might continue to show promise. Candidate priors over the hyperparameters of the between-task kernels are less apparent, and may need to be learned from data. This would require some thought as to whether there are specific physiological concepts to be encoded into these priors over hyperparameters (in addition to that which is encoded by the kernel itself.)
**Priority 2: Transition Away from Summary Statistics of Forecast Error**  A major oversight of the current work is that performance is evaluated as a summary across all patients and all testing windows. Previous work has demonstrated there were “top performers” that in certain patients had the most accurate forecasts. From this, it would be useful to identify priors (or families of priors) that perform well in different scenarios. Once these priors are found, the Bayesian inference framework can be used to decide on the most appropriate model in light of the data. (In the future, it would also be interesting to examine active learning of successful kernel and priors for individual patients.)

### 6.3 Pertaining Only to Early Warning of Deterioration

**Priority 1: EVT in Step-change Detection**  EVT could be a valuable tool to discern between true data deviation due to physiology and sporadic artefact within the same inference framework. The sophistication of measuring novelty can be meaningfully enhanced by examining the order statistics of deviations from the forecast (scaled by posterior uncertainty). A principled approach to artefact detection would also remove the need of heuristic approaches, such as thresholds and the “4-of-5 minute” rule in alarms, used by Hann [18].

**Priority 2: Improved Quality of Fitted Segments in Time-Series Matching**  The refinement of the fitted reference patterns based on their fitted values has not been explored. Within $G^1$ and $G^2$, otherwise “normal” patients have undoubtedly exhibited abnormal physiology. These “abnormal-normals” provide the motivation in [18], to remove the outermost 100 cluster centres from the KNN approach. KNN methods are already amenable to KL-Divergence metrics on multivariate normal distributions [34]. An initial attempt to improve the utility of reference patterns could be to cluster the reference patterns, and discard the patterns associated with the most outlying clusters. Alternatively, a least-outlying subset of the cluster centres could form the new references patterns. This second option is an analogue to the approach in [18] of comparing new measurements to the 400 cluster centres, but with only the single closest cluster centre under consideration.

A more heuristic approach to preprocessing the fitted reference segments could be to remove fitted segments according to low signal quality. For example, many fitted segments have intervals with very large standard deviations around the fitted mean. These sections are not usually
representative of underlying physiology, but instead occur due to the sparsity of measurements. Possible ways to avoid this include comparing only the non-sparse sections of the fitted segments. Something akin to a “signal-quality index” at each point could also be considered; for example, this could pertain to the time from the nearest training point.

**Priority 3: Additional Time-series Matching Metrics** New methods will need to be considered in order to better differentiate between the dynamics of fitted segments. Methods describing KL-divergence between GPs will be investigated first because they are a natural extension of the current work. Other methods such as Euclidean distance, longest common subsequence, and DTW have been examined in Pimentel in clustering GP. DTW is unlikely to be a priority for reasons discussed in chapter 2, but Pimentel demonstrated the clinical value of comparing several methods.

### 6.4 Risk Assessment

There are no considerable risks to the described plan, in terms of access to data: the majority of proposed work will continue to be on the UPMC data set, to which we already have access. One of my supervisors (David Clifton) already has access, and has performed analysis on several potential external data sets. Any delay to accessing these data sets is unlikely, and would not impede further development of GP techniques using the UPMC data available. A possible outcome is that the clinical benefit of GP deterioration detection methods will not surpass that of the KDE or MEWS in current clinical practice. This is is unlikely, in light of the results of univariate GPR methods. Even if unsuccessful in surpassing current clinical practice, GPR methods represent and important avenue to improve upon the i.i.d. assumptions of the current clinical models.

### References


Appendix: GSO.2 Questions

A. Please describe briefly any subject specific research skills that you have developed or improved in the course of your time as a Probationary Research Student. For example, these might include: research methodology; data analysis and management; record keeping; bibliographical skills; presentation of research.

My greatest source of development as a PRS is a greater grasp of the work required to progress my research project towards becoming the clinical state-of-the-art. Over the last year I have improved my understanding of the statistical tools at my disposal (Gaussian Processes) and the multitude of ways in which they may be used to address the clinical application. I have also learned how to better identify the most promising paths forward, with regards to my research, and how to remain focussed on work that addresses the most salient aspects of my DPhil project. Between conferences, workshops, CHI group presentations, and 1-on-1 meetings with my supervisors, I have also made progress in learning how to communicate my research to others. In particular, I have learned how to communicate better with researchers outside the sphere of biomedical engineering.

B. Please describe briefly any personal and professional skills in which you have received training or which you have enhanced during the course of your time as a Probationary Research Student. For example: time management, language skills, IT skills, team work, problem solving, presentation skills, teaching skills, career planning.

Between CHI group meeting and conferences, I have become much more experiences with discussing and defending my work. I have also gained experience in teaching and leading discussions on technical subjects related to my work such as convex optimization, Deep Gaussian processes, Gaussian processes using inducing points, and advanced parallel computing tasks. I have also worked on developing a greater range of professional relationships by becoming more involved with the American Statistical Association (ASA) and, in particular, the Statistical Learning and Data Mining (SL&DM) Section of the ASA. Over the last year I have written a recruitment newsletter for the Biopharmaceutical sections of the ASA, organized a section mixer for SL&DM, attended the SL&DM Annual Business Meeting, and am currently heading a committee to design a membership survey for the SL&DM. Dr. David Clifton has also facilitated my involvement with a study at the Diabetes Trial Unit in Oxford, and a trip to Guy’s and St. Thomas’ Hospital in London, both of which have been a great opportunity to experience work in a clinical setting.

C. Please identify any subject-specific or personal and professional skills in which you (and your supervisor(s)) foresee the need for further development or training.

It will be important for me to continue to develop as both a scientist and an engineer. As a scientist I will need to continue to ensure that my work remains on task to address the clinical goals of my DPhil. As an engineer, I will need to continue to develop my understanding of the statistical methods I am using, in particular, this would be Gaussian Processes, and more generally this would be Bayesian inference. An example of my future plans to continue to develop my knowledge of Gaussian Processes is my attendance of the Gaussian Process Summer School, at the University of Sheffield from 14-17 September, 2015. Both of my supervisors would like to see greater improvement in my time-management skills.

D. Please list any other activities which have contributed to the development of your work. For example, these might include courses attended, conference presentations given, publications, opportunities to undertake teaching, etc.

Over the last year I have attended two conferences: The 10th Conference on Bayesian Nonparametrics in Raleigh, NC (where I presented a poster on my DPhil work) and the Joint Statistical Meetings, in Seattle, WA (where I presented a poster on my DPhil work, chaired a contributed paper session on nonparametric regression, and chaired a round-table discussion on convex optimization.)

I have attended a workshop on environmental analytics, in Boulder, CO, to continue to develop my
understanding of Gaussian Processes, two workshops hosted by the Advanced Research Computing Lab in Oxford to become familiar with the parallel computing resources at the university, and workshop on “Preparation for Tutorial Teaching” within the Department of Engineering Science.

Each CDT student has had a week of training in video production, filming techniques, and editing.