Bayesian Optimization of Personalized Models for Patient Vital-Sign Monitoring

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

***Abstract*—Gaussian process regression (GPR) provides a means to generate flexible personalized models of time series of patient vital signs. These models can perform useful clinical inference in ways that population-based models cannot. A challenge for the use of personalized models is that they must be amenable to a wide range of parameterizations, to accommodate the plausible physiology of any patient in the population. Additionally, optimal performance is typically achieved when models are regularized in light of the knowledge of the physiology of the individual patient. In this paper, we describe a method to build GP models with varying complexity (via covariance kernels) and regularization (via fixed priors over hyperparameters) on a patient-specific level, for the purpose of robust vital-sign forecasting. To this end, our results present evidence in support of two main hypotheses: 1) the use of patient-specific models can outperform population-based models for useful clinical tasks, such as vital-sign forecasting; and 2) the optimal values of (hyper)parameters of these models are best determined by sophisticated methods of optimization, due to high correlation between dimensions of the search space. The resulting models are sufficiently robust to inform clinicians of a patient's vital-sign trajectory and warn of imminent deterioration.**

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

I. Introduction

PERSONALISED inference is a promising source of improvement for the manner in which patient physiological data may be used to provide better patient outcomes. In many clinical settings, the "early-warning signs" of impending physiological deterioration may be missed by time- and resource-constrained clinical staff; this effect may be compounded by the "data deluge" caused by acquisition of ever more complex patient data during routine care. An important exemplar is the field of continuous vital-sign monitoring in critical care settings - a task that is impractical without robust means of analysing the data and providing reliable forecasts of patient condition.

It is not possible for members of clinical staff to monitor all vital-signs of all patients at all times. Whereas a critical care specialist is reliant on summary statistics and risk scores (e.g., EWS, MEWS, NEWS) at single points in time [1], computational inference can make full use of the second-by-second acquisition of vital-sign measurements. Computational inference can assess deterioration in many forms including, (i) abnormal joint distribution of vital-signs [2], (ii) step-changes in vital signs [3], (iii) abnormal vital-sign trajectories [4], and (iv) a number of clinically-relevant features, such as lab values and cognition tests [5].

Personalised models for vital-sign monitoring offer significant advantages over population-based models due to the heterogeneity of large clinical populations. For example, comparing a patient's current vital-signs to the joint distribution of vital-signs of a non-deteriorating patient population has demonstrated advantages over population-based thresholds (such as MEWS [1]) to identify deteriorating patients [2]. However, patient-specific time-series models, even when using only a single time-series, have shown evidence of stronger performance [3] and are the subject of the work described in this paper. Despite their advantages, the use of personalised models has its inherent challenges, since the parameterisations of such models must be sufficiently flexible to accommodate a diverse range of patients, and we must learn patient-specific parameterisations for the individual patient in a timely manner.

Our preliminary discussion [6] described a method to learn quickly the "optimal" parameterisations of Gaussian processes (GPs) for the personalised modeling of vital-sign time-series. These optimised models provide (i) forecasts of a patient's future vital-sign trajectory, as shown in Fig. 1(a); and (ii) a robust form of step-change detection, by notifying clinicians when a patient's vital signs deviate significantly from their anticipated trajectory, as shown in Fig. 1(b). Poor parameterisations, as shown in Fig. 1(c) and (d) will interfere with these tasks. The work described in this paper expands on our preliminary discussion [6] in several ways:

1) The comprehensive study described in the current work ($N = 169$ patients) expands that considered in the preliminary discussion ($N = 34$).
2) Multiple competitive methods are presented, and tuned to identify optimal patient-specific parameterisations.

Manuscript received May 15, 2017; revised August 22, 2017; accepted September 2, 2017. Date of publication December 19, 2017; date of current version March 5, 2018. The work of G. W. Colopy was supported in part by the Clarendon fund and in part by the Engineering and Physical Sciences Research Council (EPSRC). The work of S. J. Roberts was supported by the Royal Academy of Engineering/Man-AHL Research Chair. The work of D. A. Clifton was supported in part by the Royal Academy of Engineering Research Fellowship, in part by the Research Fellowship at Balliol College, Oxford, and in part by the EPSRC "Grand Challenge" Early-Career Fellowship. *(Corresponding author: Glen Wright Colopy.)*

The authors are with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: glen.colopy@eng.ox.ac.uk; sjrob@robots.ox.ac.uk; davidc@robots.ox.ac.uk).

Digital Object Identifier 10.1109/JBHI.2017.2751509

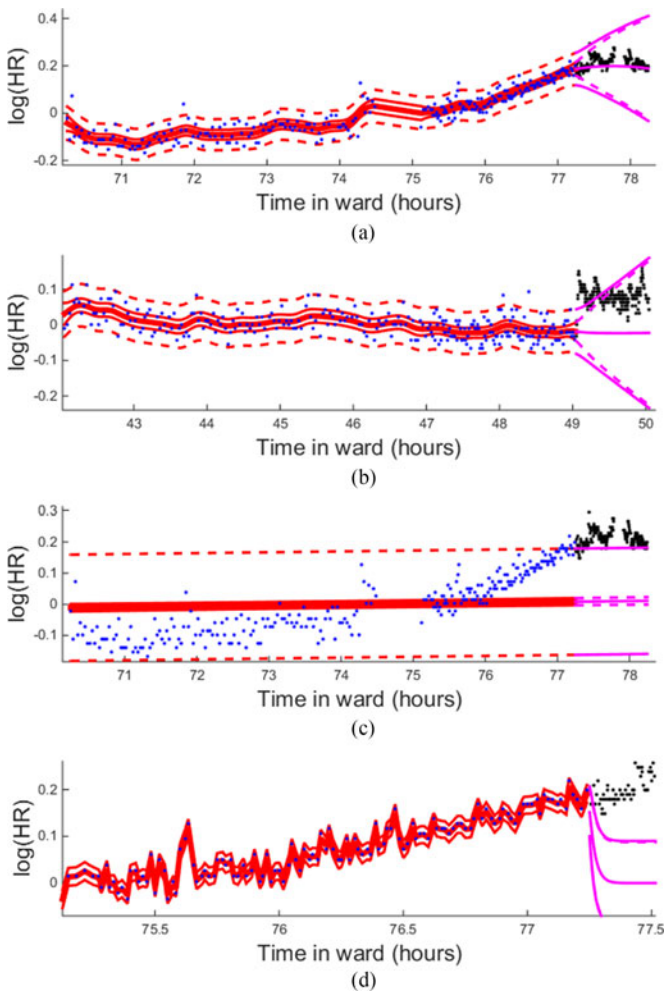

Fig. 1. Four time series with GPs fit to training points (blue) and forecasted to future points (black). Solid and dashed lines represent the 95% CI of the latent mean and observations, respectively. We wish to use GP fits inductively, to provide forecasts that will inform us of (a) future values, or (b) unforeseen deterioration. In both cases, correct parameterisations of GP models are necessary for robust clinical inference. Some poor forecasts, which may be mistaken for deterioration, could be foreseen, such as when the noise parameter, σ_n^2, is (c) over-estimated, or (d) under-estimated.

3) Optimisation is performed across all hyperparameters in the GP covariance function (our preliminary discussion considered only length-scale hyperparameters, for brevity).
4) Optimisation is described across a choice of 1, 2, or 3 kernels, to describe flexibly the complex nature of an individual patient's time-series.
5) A method of multi-objective optimisation is motivated and described, which we will show yields models that permit robust forecasting.

II. Data

A. Our Clinical Study

The data considered by the work described in this paper comprises 336 patients from the University of Pittsburgh Medical Center (UPMC) step-down ward. Continuous vital-sign time-series (heart rate, HR, respiratory rate, blood-oxygen saturation, SaO_2, blood pressure, and temperature) were collected as part of a clinical study. For each patient, each vital-sign time-series was retrospectively examined by clinical experts for perceived clinical emergencies, defined as being instances of abnormally high or low values of a single vital-sign's measurements that require medical intervention. Of the 336 patients, 59 patients were identified as having one or more clinical emergencies; 267 patients did not have any clinical emergencies identified by those criteria. The remaining $336 - 59 - 267 = 10$ patients had no recorded vital-sign data. For inclusion in the analysis, we consider those 169 patients for which there exists at least two hours of HR data in the first 24 hours of observation, and at least 1 hour of data between hours 24–72 of observation. This choice is motivated below.

B. Training, Validating, and Testing

The 169 patients under analysis were partitioned into subsets of 43 patients for training/validation purposes and 126 for testing[1]. As described later, we aim to learn patient-specific regularisers for complex, non-parametric models of the HR time-series from the first 24 hours of each patient's data, to improve forecasting of the HR time-series in hours 24–72. The regulariser aims to prevent overfitting of the model, thereby improving our ability to use the model subsequently, with previously-unseen data. This paper hypothesises that patient-specific regularisers will lead to optimal predictive capability. An example is shown in Fig. 2(a), which shows a time-series of HR data, and where the corresponding log marginal likelihood (LML) values of forecasts are shown in Fig. 2(c). In clinical practice, learning the optimal patient-specific regulariser would begin for each patient immediately upon admission to the ward. The ideal regulariser, according to currently-available data, would be updated as new data became available. For simplicity, we will examine the performance of these models after 24 hours of observation. This arbitrary cut-off could be any time $t > 0$, although an early cut-off reduces the quantity of data available from which to learn the regulariser, and a late cut-off reduces the quantity of subsequent data available to assess the performance of the regulariser. The range of previous data to select for training, and the frequency of updating the regularisers is considered as potential future work.

The number of HR measurements for each patient is shown in Fig. 3(a) for the 43 training/validation patients and 3(b) for the 126 (held-out, previously-unseen) test patients. Using 24 to 72 hours of continuous data from 169 patients results in a comparatively large study for continuous patient-specific modelling. Such a large study offers advantages concerning the validity of the final results. First, it allows us have a large number of held-out patients on which to test our proposed personalised modelling approach. Secondly, this allows us to examine our models on a wide range of patient physiologies, which vary

[1]N.B.: We use the term "validation" in its usual sense in the computer science literature, for providing an estimation of how our modelling choices will perform when using previously-unseen data (which is our "test" set). We note that the latter contrasts with the medical literature, which uses the term "validation set" to refer to held-out, previously-unseen data.

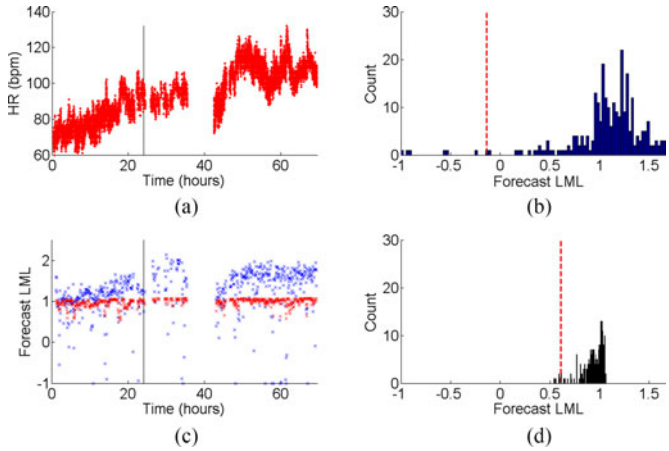

Fig. 2. (a) Continuously-acquired HR data for an exemplar patient, with (c) corresponding values of LML obtained from forecasting - using a GPR with an uninformative prior (blue) and regularised hyperparameters (red). Note that although the use of an uninformative prior (blue) tends to result in higher values of forecast LML for this patient, the regularised method (red) avoids the large number of poor estimates and is more robust overall. The distribution of the forecast LMLs from (c) is shown in (b) for uninformative priors and (d) for patient-specific regularisation to optimise the 2.5-percentile of forecasts (the vertical dashed line), which is objective function G_1, defined later. Although use of the regulariser certainly avoids worst-case scenario performance, we believe that the regulariser could be learned without such detriment to upper-end performance. This motivates the use of multi-objective optimisation, described later.

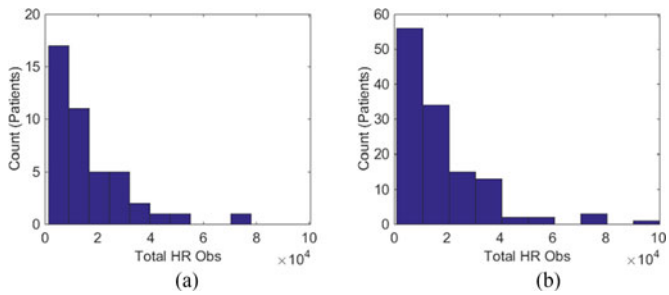

Fig. 3. The total number of HR measurements (a) for each of the 43 patients in the training/validation set, and (b) for each of the 126 patients in the test set. HR measurements are acquired at (a non-constant sampling rate) approximately $\frac{1}{3}$ Hz to $\frac{1}{5}$ Hz, and so patients with more than approximately 1.7×10^4 measurements correspond to those having more than 1 day of useable data. Patients with fewer measurements (which may be dispersed in time) will have fewer data from which to learn a personalised model (i.e., they may have fewer data in their first 24h), and fewer data with which to test the performance of a personalised model (i.e., they may have fewer data after their first 24h).

significantly even within the same ward. This means our proposed method will be evaluated across a wide-range of clinical scenarios, which is highly desirable for the goal of personalised modelling. Finally, the lengthy duration of monitoring for each patient within our study allows us to understand how far into the future a patient-specific model will remain useful, after optimised parameters are selected.

C. HR Pre-Processing

Continuous HR data are typically beset with artefactual corruption due to the measurement process, including partial attachment of the measurement probe (often an ECG electrode or finger-mounted pulse oximeter), or failures to identify the pulsatile complex of wave-forms (e.g., the QRS complex in the ECG) that subsequently confounds HR estimation.

Just as we advocate the modelling of a patient's time-series data using patient-specific (and not population-based) approaches, a principled approach to artefact removal evaluates potential artefacts in light of the patient's current measurements, and not using generic rules-of-thumb. A traditional approach is to set extreme thresholds based on population-derived values [2], but these ignore the majority of artefacts which are readily apparent in a patient-specific context.

As noted in [7], personalised artefact removal is preferable to common alternatives to handling artefactual measurements. For example, mean/median smoothing tends to reduce the variability of HR measurements which, in turn, will bias downward our estimates of HR variance. Setting thresholds to remove extreme-valued HR measurements, removes only a minority of all artefacts, leaving the remainder to hinder inference.

An artefact detection algorithm, previously validated on the UPMC data set [7], was used to pre-process the HR time-series for our current study. The method works by modelling HR variation within a short time window as draws from an assumedly-iid Gamma distribution. HR measurements with extreme deviation from measurements within the same window are identified by their low likelihood with respect to the density function of the Gamma distribution that has been fitted to the measurements within the window; these deviating data are subsequently removed as artefacts.

III. Personalised Vital-Sign Monitoring with Gaussian Processes

A. Motivation

The purpose of vital-sign monitoring in critical care is to assess the health status of the individual patient and to provide automated parsing of (often high-frequency) time-series data, to support clinical decisions. As shown in Fig. 1(a) and (b), robust forecasts of future vital-sign values can provide clinicians with a useful indication of the patient's health status. Additionally, vital signs can be directly assessed for signs of deterioration, thus providing early-warning to the clinician.

Common vital-sign monitoring algorithms in current clinical practice, such as [2], [8], [9], model individual vital-sign measurements as iid draws from a distribution obtained from a population of critical care patients. This distribution is therefore assumed to be unchanging with respect to time, thereby disregarding any dynamics in the data. Additionally, it assumes that all patients have the same distribution of vital-sign data. When a vital-sign measurement is abnormal with respect to this distribution, then the measurement is classified as being abnormal, and (if the subsequent alert is noticed by clinical staff) will precipitate clinical action.

The fundamental flaw to this population-based approach is that vital-sign measurements are not iid occurrences, but inextricably correlated with the patient's other measurements, and which change with time in a patient-specific manner. Vital signs

are contextual, for example, to the patient's characteristic baseline values or to values in the recent past. In particular, useful clinical inference can be gained by monitoring and modelling how a specific patient's vital signs change over time. Popular methods to model vital-signs include GPs [10]–[15]; Kalman filters and AR models [16], [17]; support vector machines; and recurrent neural networks [18]. This paper will focus on GPs.

Particularly good introductions to GP regression (GPR) as applied to time-series modelling and Bayesian optimisation are [19] and [20], respectively. GPR is an attractive method to model vital signs because GPs are non-parametric, thereby allowing (i) modelling without undesirable recourse to specific parametric forms, and (ii) complexity of the model to grow with the size of data. Furthermore, GP inference can be expressed in a Bayesian framework, thereby (iii) allowing the incorporation of prior knowledge of patient physiology into the model in several ways. GPR may be extended easily into the multivariate case to model several contemporaneous vital signs [4]. It is for these characteristics that GPs are amendable to the personalised patient models that motivate this work.

B. GPR Model

A helpful introduction to GPR and to the material described in this section can be found in [21]. Below we describe how GPR can be used to model a HR time-series, y, as a function of time, t. HR measurements are assumed to be generated from a latent time-varying function $f(t)$, corrupted by iid Gaussian noise, ϵ. That is, $y_i = f(t_i) + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_n^2)$. The correlation between two HR measurements y_i and y_j may be expressed as a function of their respective times $r = |t_i - t_j|$, which is modelled by a covariance function $k(t_i, t_j)$. The covariance function has an interpretable parametric form and can be used to enforce desirable properties, such as smoothness or periodicity on $f(t)$.

Modelling $\mathbf{y} = \{y_1, .., y_n\}$ jointly yields a multivariate Normal (MVN), $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with corresponding likelihood function

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

Without loss of generality, $\boldsymbol{\mu}$ is typically set to $\mathbf{0}$ after detrending, and $\boldsymbol{\Sigma}_{i,j} = k(t_i, t_j)$. As shown in [21], prediction of new values $\boldsymbol{y}^*$ at times $\boldsymbol{t}^*$ occurs via the conditional distribution $\boldsymbol{y}^*|\boldsymbol{y}$, which is also MVN with

$$\boldsymbol{\mu}_{\mathbf{y}}^* = \mathrm{E}[\mathbf{y}^*] = \boldsymbol{K}^*\boldsymbol{K}^{-1}\mathbf{y}.$$
$$\boldsymbol{s}_{\mathbf{y}}^* = \mathrm{Var}[\mathbf{y}^*] = \boldsymbol{K}^{**} - \boldsymbol{K}^*\boldsymbol{K}^{-1}\boldsymbol{K}^{*T}. \quad (2)$$

where $\boldsymbol{K}^{**}$ is the covariance of $\boldsymbol{y}^*$, at $\boldsymbol{t}^*$, and $\boldsymbol{K}^*$ is the covariance between points $\boldsymbol{y}$ and $\boldsymbol{y}^*$, at $\boldsymbol{t}$ and $\boldsymbol{t}^*$, respectively. This means that the LML of $\boldsymbol{y}^*$, given the predictive distribution, is

$$\log p(\mathbf{y}^*) = \log \text{MVN}\left(\mathbf{y}^*|\boldsymbol{\mu}_{\mathbf{y}}^*, \text{diag}\left(\boldsymbol{s}_{\mathbf{y}}^*\right)\right). \quad (3)$$

Optimising the LML in (3) for the purpose of robust forecasting is the goal of this work.

The covariance kernel $k(t_i, t_j)$ determines the elements of $\boldsymbol{K}$; an example covariance function is the sum of several Matern$\left(\frac{3}{2}\right)$ functions, with additive white noise, σ_n^2:

$$k_a = \sum_{i=1}^{a} h_i^2 \left(1 + \frac{\mathbf{r}\sqrt{3}}{\lambda_i}\right) \exp\left(-\frac{\mathbf{r}\sqrt{3}}{\lambda_i}\right) + \sigma_n^2\delta(t_i, t_j). \quad (4)$$

which is parameterised by output scales h, length-scales λ, and noise variance σ_n^2, described earlier. The matrix $\mathbf{r}$ encodes distances r, as described above. The Kronecker delta function, δ, is 1 when the inputs are identical, and 0 otherwise, modeling noise variance at each measurement. The collation of these hyperparameters is the set $\boldsymbol{\theta}_a = \{h_{i=1,..,a}, \lambda_{i=1,..,a}, \sigma_n^2\}$. This covariance function encodes a once-differentiable function (attractive for the typically erratic and quantised volatility that occurs with HR time-series), with a additive components accounting for variation in the signal, in addition to noise corruption.

The suitability of a choice of values for $\boldsymbol{\theta}_a$ is assessed by (1). The value of $\boldsymbol{\theta}_a$ may be learned by maximising (1), or alternatively by integrating across a range of values for $\boldsymbol{\theta}_a$ [22].

C. GPR for Patient-Specific Inference

We aim to learn which (compound) kernel k_a and hyperparameters $\boldsymbol{\theta}_a$ best reflect the generative process of an individual patient's HR measurements. The observed vital-sign data provides insight into the patient's true physiology, but the data alone may be insufficient for robust estimation. Solutions include (i) regularising (global) priors that add a further term to (1) and which thereby shift its modal value; (ii) MCMC integration [22], which offsets the risk of a single poor choice in $\boldsymbol{\theta}_a$; and (iii) fixing the value of $\boldsymbol{\theta}_a$ to those values that have successfully modelled the patients values in the past. We propose the use of (iii), which requires us to select a patient-specific k_a and optimise the respective $\boldsymbol{\theta}_a$; this task represents a two-fold challenge.

IV. Optimisation of Patient-Specific Models

A. The Optimisation Task: Objective Function and Constraints

We seek to learn, for each patient, an optimal k_a and $\boldsymbol{\theta}_a$ such that the proportion of extremely low forecast performances is minimised. That is, we require a "worst-case" forecast performance that is at least robust. This represents the clinical reality that a system that has extremely poor worst-case performance will not be adopted by clinicians [23].

This effect is illustrated by examining the lower tails of the distributions in Fig. 2(b) and (d). In Fig. 2(b), uninformative priors result in a high best-case performance (sometimes reaching LML values above 1.5), but which have often-poor performance (LML < 0.5). The use of patient-specific priors, acting as a regulariser, yield results that have a lower maximum LML (of approximately 1.1) but which have much better worst-case performance (above 0.5).

We focus on the lower tail of forecast performance because either (i) such forecasts, if presented to clinical staff, are an extreme mischaracterisation of a patient's future possible trajectory (e.g., the estimate of σ_n^2 was too large); or (ii) such forecasts represent an instance of the model being over-certain (e.g., the

estimate of σ_n^2 was too small). We note that alarms subsequently provided to a clinician will be likely to occur for test data that have lower values of LML with respect to the GPR model, and we therefore adopt the approach that the system should be maximally robust when handling data that correspond to lower values of LML.

We define $\mathcal{L}_{\boldsymbol{\theta}}$ as being the set of all forecast log-likelihoods from sequential forecasting for an individual patient under parameter settings $\boldsymbol{\theta}$; for example, those values shown in Fig. 2(b) and (d). We define $G_1(\boldsymbol{\theta})$ as being the 2.5 percentile of $\mathcal{L}$, as is marked by the vertical dashed line in Fig. 2(b) and (d). We formalise our learning goal as an optimisation problem:

$$\begin{aligned} \max \quad & G_1(\boldsymbol{\theta}) \\ \text{s.t.} \quad & l_d \leq \boldsymbol{\theta}_d \leq u_d. \end{aligned} \tag{5}$$

The optimisation parameters u_d and l_d represent, respectively, the upper and lower bounds placed on the dth element of $\boldsymbol{\theta}$. For example, for $\boldsymbol{\theta}_2 = [h_1, \lambda_1, h_2, \lambda_2, \sigma_n]$, then $[l_2, u_2, l_4, u_4] = [2.5, 45, 60, 600]$ would require that the length scale of the first kernel fall between 2.5 and 45 minutes and that of the second kernel fall between 60 and 600 minutes.

Although the optimal solution of the bounded problem necessarily lies within the set of solutions of the unbounded problem, the constraints l and u are valuable to (i) reduce the search space to the most plausible locations of an optimal solution; (ii) prevent overlapping length scales, which reduces kernel complexity; and (iii) ensures only computationally-stable values of $\boldsymbol{\theta}$ are selected (i.e., to avoid computational singularity in $\mathbf{K}$).

The objective function $G_1(\boldsymbol{\theta})$ is non-analytic and requires sampling to learn its properties. Each evaluation of $G_1(\boldsymbol{\theta})$ is expensive (requiring a pass of sequential GP fitting/forecasting through the patient's available data, for each queried $\boldsymbol{\theta}$), making methods based on gradient- and line-search undesirable due to their extremely expensive use of multiple function evaluations at each step for each patient. This expense would prohibit such approaches being used in practice, where the latter is an important goal of our work. We will therefore propose several candidate methods to locate an optimal value of $\boldsymbol{\theta}$ for each patient in a computationally-tractable manner.

B. Optimisation of Multiple Clinical Objectives

Objective function G_1 is certainly not the only reasonable clinical goal: as seen in Fig. 4(b) and (c), optimising a lower quantile of forecast performance frequently reduces the predictive performance of larger quantiles, thereby giving clinicians less precision on trajectories that are easier to predict. Alternatively, there is no reason to recommend only a single duration over which to provide forecast, since clinical staff may be interested in short, medium, and long-term estimates. It is plausible that a value for $\boldsymbol{\theta}$ that optimises G_1 may be less desireable than a nearby alternative value of $\boldsymbol{\theta}$, which optimises other objectives, with little cost to G_1.

To optimise further objectives, we propose an alternative objective function, G_2, which is motivated by vectorisation

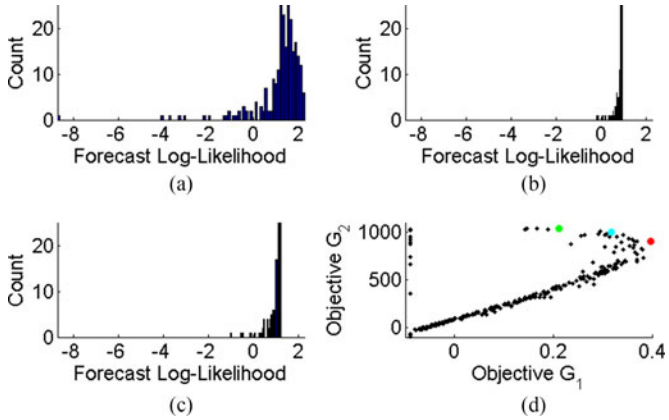

Fig. 4. Forecast LML within the first 24 hours for a single patient is shown for (a) uninformative priors, (b) finding $\boldsymbol{\theta}$ by optimising with objective G_1, and (c) finding $\boldsymbol{\theta}$ by optimising with objective G_2. In (d), the performance in G_1 and G_2 of 250 values $\boldsymbol{\theta}$ are shown. While most $\boldsymbol{\theta}$ with high values of G_2 have high values of G_1, a few (in the upper left) have very poor performance in G_1. The $\boldsymbol{\theta}$ value that optimises G_1 (red) and G_2 (green) show a trade-off in performance from selecting one value of $\boldsymbol{\theta}$ over another. The $\boldsymbol{\theta}$ that optimises G_3 (cyan) may be a reasonable trade-off between the objectives.

approaches to multi-objective opimisation.

$$G_2 = \sum_{\rho=2}^{50} (51 - \rho) \mathcal{L}_{\boldsymbol{\theta},\rho} \tag{6}$$

where $\mathcal{L}_{\boldsymbol{\theta},\rho}$ represents the ρth percentile of $\mathcal{L}_{\boldsymbol{\theta}}$. This formulation encodes a preference for optimising lower quantiles over higher quantiles, up to the median forecast performance. Values of $\boldsymbol{\theta}$ which perform strongly at the lower quantiles will dominate G_2, but if alternatives are found at little cost to the lower quantiles, then they will be preferred over more "myopic" values of $\boldsymbol{\theta}$.

Since performance at low quantiles is subjectively preferable, due to the clinical goals of personalised modeling, a potential pitfall of using objective G_2 is that performance at the higher quantiles (i.e., $\mathcal{L}_{\boldsymbol{\theta},\rho}$ at high values of ρ) may be large enough to outweigh the higher weighting of the performance at lower quantiles (i.e., $\mathcal{L}_{\boldsymbol{\theta},\rho}$ at low values of ρ), as seen in Fig. 4(d). To address this (rare) possibility, a final optimisation problem, motivated by work in the field of lexicographic multiobjective optimisation, is suggested:

$$\begin{aligned} \max \quad & G_3(\boldsymbol{\theta}) \\ \text{s.t.} \quad & G_3(\boldsymbol{\theta}) := G_2(\boldsymbol{\theta}) \\ & l_d \leq \theta_d \leq u_d \\ & G_1(\boldsymbol{\theta})^* - 0.1 \leq G_1(\boldsymbol{\theta}). \end{aligned} \tag{7}$$

where $G_1(\boldsymbol{\theta})^*$ is the best-found value of G_1. This encodes the preference that we are willing to allow G_1 performance to degrade up until the point at which the performance is less than 0.1 of its best-found value. Practically, the optimisation of G_3 will entail a search to optimise G_2, followed by a post-processing removal of any queries that do not satisfy the second constraint of G_3.

These modifications to the objective function are not superfluous addenda, but quantify valuable information when it comes

TABLE I
UNIFORMATIVE PRIORS FOR HYPERPARAMETER REGULARISATION

Prior	Parameterisation
Uniform	$p(x) \propto 1$
Square root uniform	$p(x^{1/2}) \propto 1$
Log-uniform	$p(\log x) \propto 1$
Log-log-uniform	$p(\log \log x) \propto 1$

to selecting a final patient-specific parameterisation. G_2 and G_3 may collate and regularise multiple performance objectives. When commentary is applicable to any of the three objectives, G or $G(\boldsymbol{\theta})$ with subscript will be used to denote a generic objective function.

V. BASELINE COMPARATOR METHODS

We aim to test two hypotheses: (i) that patient-specific models are superior for vital-sign forecasting, and (ii) the complexity of the search domain (noting that we have highly correlated GP hyperparameters) necessitates the use of appropriate optimisers, such as Bayesian optimisation (BO). To address (i) we implemented a comparator method with uninformative priors, which accommodate any plausible parameterisation in the patient population. To address (ii), we implemented a comparator well-tuned random search (RS) algorithm, and several related methods to search for patient-specific parameterisations.

A common critique of publications comparing the performance of different optimisation algorithms is that only the novel method (that proposed by the authors) is tuned for best performance on the optimisation problem at hand. This is a legitimate critique, which we address by describing how to tune each baseline method so as to provide strong comparators for proposed methods.

The training/validation set comprises 43 patients as described previously. The purpose of the training set is two-fold: (i) to learn the best optimiser to identify suitable values for $\boldsymbol{\theta}$ using data from hours 1–24, and (ii) to learn which high-performing $\boldsymbol{\theta}$ is best to perform forecasting in hours 24–72.

A. Combinatorial Search Over Uninformative Priors

1) Method: The most common response to modelling the heterogeneous physiology of patients is to assign an uninformative prior over each hyperparameter in $\boldsymbol{\theta}$. Common uniformative priors, which are available in [24], are shown in Table I. Further information on uninformative priors can be found in Section 2.4.3 of [25]. As seen by the parameterisations in Table I, the uninformative priors regularise the fit of the hyperparameters by placing increasingly-smaller probability mass on high values of that hyperparameter. For example, the log-uniform distribution will regularise against large values more stringently than the square-root uniform, but less stringently than the log-log-uniform distribution. These relations may be used to describe preferences for certain hyperparameters to have higher magnitudes than others without an explicit statement of what those values might be. For example, by placing a log-uniform prior on h and a log-log-uniform prior on σ_n^2, we may encode the belief that the time series' variance contains more signal than noise. Following this intuition, uninformative priors for kernels $k_{a=1,..,3}$ were selected, via extensive search over the various combinations of uninformative priors over each hyperparameer, to optimise the performance in G for that kernel.

Although only limited tuning of this baseline comparator is possible, the models that incorporate uninformative priors have a particular advantage: these models are free to learn a new value of $\boldsymbol{\theta}$ at every time point (after 24 hours) as more data is acquired, with high likelihood with respect to the patient's most current data (the optimisation methods, proposed below, will not update $\boldsymbol{\theta}$ after 24 hours).

2) Tuning Uninformative Priors: Although patient-specific tuning of hyperparameters is not possible in this context, patient-specific kernel selection (between k_1, k_2, and k_3) is possible, and represents the most sophisticated approach to model selection while still being a population-based modelling approach. For each patient, tuning included the selection the best GP kernel ($k_{a=1,..,3}$) with respect to performance in G over the first 24 hours, to perform forecasting in the subsequent days. Results were similar between (i) selecting the best kernel from $k_{1,..,3}$ after 24 hours for each patient, and (ii) simply choosing k_2 each time. This suggests that uninformative priors have difficulty making full use of more complex kernels, although this effect is not as dramatic as will be seen with random search, below. Approach (ii) was selected for the test set.

B. Random Search

1) Method: Random search (RS) is a popular and effective baseline comparator for global optimisation techniques. RS in most commonly defined to be a random uniform sampling of a hyper-rectangle in search of a global optimum. Common variants are discussed in [26]. As described in [27], strengths of RS include (i) being trivial to program, and (ii) exploiting the low effective dimensionality of many optimisation problems.

2) Tuning Random Search: RS may be tuned to the problem at hand by placing a higher sampling density where optima are likely to be found, making better use of each expensive evaluation. This can be achieved a priori, or adaptively, as new observations are learned. The bounds of the hyper-rectangle, l_d and u_d, from which random values of $\boldsymbol{\theta}$ are drawn, are apt for tuning. Tuning l_d and u_d involved running several long (1000 samples) random searches with large sampling bounds. The sampling bounds were tightened to include only regions in which patient-specific optima occurred. Further tuning of random search was attempted by using cross-validation techniques: for each training set patient under consideration, the best-found $\boldsymbol{\theta}$ values of the $43 - 1 = 42$ training-set patients not under consideration, were used to populate high-density regions for a more focused search. The best-found values of $\boldsymbol{\theta}$ of the 42 other patients were values found in previous runs. Samples around these points were then used to investigate further refinements to RS to either (i) form the centroids of more focussed sampling, or (ii) populate the seed generation of a genetic algorithm to optimise $\boldsymbol{\theta}$ (details of which are omitted here) These methods, trading exploitation for exploration, tended to be no better or worse than the original

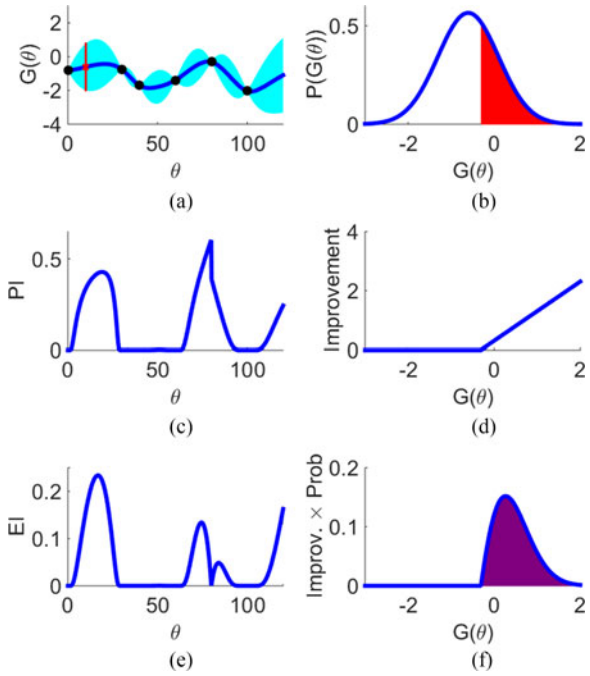

Fig. 5. The BO algorithm in (a) has fitted a posterior GP to the 6 observed queries of $G(\boldsymbol{\theta})$, with $G(80) = -0.3$ as the current best-found (i.e., maximum) value. The next query may be selected such that it maximises an acquisition function such as (c) the probability of improvement (PI) at a new value of $\boldsymbol{\theta}$ over the current best value of $\boldsymbol{\theta}$, or (e) the expected improvement (EI) of querying a new value of $\boldsymbol{\theta}$. The components of each acquisition function, assessed at $\boldsymbol{\theta} = 10$ are shown in ((b)(d)(f). The posterior marginal distribution of $G(\boldsymbol{\theta} = 10)$ is marked in red in (a) and plotted in (b). The probability that the query will exceed the current best value (as estimated by the GP) is the area under the curve (AUC) in red (which is therefore the integral above $\boldsymbol{\theta} = -0.3$). The improvement over the current best value, for any possible realisation of $G(\boldsymbol{\theta} = 10)$ is shown in (d). Note that if the true value is less than the current best, then the improvement is 0. Taking improvement to be a random variable emitting values in (d) with probability density (b), the expected value of improvement is in (f) the AUC of their product, in purple.

random search and thus we proceed with using the original RS, described above.

VI. Bayesian Optimisation

A. Overview of a Bayesian Optimisation Algorithm

BO formalises the probabilistic decision process of selecting new queries in light of the values from previous queries. Using the notation of Section III for GP modelling, Algorithm 1 presents the steps of BO, further illustrated in Fig. 5, and which is described in detail below.

For a set of known queries of the objective function, $G(\boldsymbol{\theta}_{\text{prev}})$ at queries $\boldsymbol{\theta}_{\text{prev}}$, BO uses a GP to estimate a posterior distribution in unqueried regions, $G(\boldsymbol{\theta}^*)$, where the values are uncertain, see Algorithm 1, line 3i, and Fig. 5(a). The desirability of the next query point, $\boldsymbol{\theta}_{\text{new}}$, is estimated by an acquisition function, $A(\boldsymbol{\theta}_{\text{new}})$, using the marginal posterior distribution, see Algorithm 1, line 3ii, and Fig. 5(c) and (e). The next query is the value of $\boldsymbol{\theta}$ that maximises $A(\boldsymbol{\theta})$. New queries continually update the available data, which refine the GP estimate of $G(\boldsymbol{\theta})$ in the remaining unexplored space. Ultimately, the best-found query is selected as the optimal solution encountered overall.

Algorithm 1: Bayesian optimisation algorithm.

1: **query** $G(\boldsymbol{\theta})$ at initial points, $\boldsymbol{\theta}_{\text{init}}$.
2: $\{G(\boldsymbol{\theta}_{\text{prev}}), \boldsymbol{\theta}_{\text{prev}}\} := \{G(\boldsymbol{\theta}_{\text{init}}), \boldsymbol{\theta}_{\text{init}}\}$
3: **while** Iter $<$ ComputationalBudget
  i **estimate** $\boldsymbol{\mu}_G^*$ and $\boldsymbol{s}_G^*$ from data $\{G(\boldsymbol{\theta}_{\text{prev}}), \boldsymbol{\theta}_{\text{prev}}\}$.
  ii **estimate** $A(\boldsymbol{\theta})$, from $\boldsymbol{\mu}_G^*$ and $\boldsymbol{s}_G^*$
  iii **query** $G(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{\text{new}} := \arg\max_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$.
  iv $\{G(\boldsymbol{\theta}_{\text{prev}}), \boldsymbol{\theta}_{\text{prev}}\} :=$
    $\{G(\boldsymbol{\theta}_{\text{prev}}) \bigcup G(\boldsymbol{\theta}_{\text{new}}), \boldsymbol{\theta}_{\text{prev}} \bigcup \boldsymbol{\theta}_{\text{new}}\}$
4: optimal solution $\boldsymbol{\theta} := \arg\max_{\boldsymbol{\theta}_{\text{prev}}} G(\boldsymbol{\theta}_{\text{prev}})$

B. Bayesian Optimisation Gaussian Process Prior

The approach to constructing a GP prior over $G(\boldsymbol{\theta})$ is the same as that described in Section III. We aim to incorporate a prior mean function and components of covariance function $C(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that reflect the underlying generative process of $G(\boldsymbol{\theta})$. For example, our queries, $G(\boldsymbol{\theta}_{\text{prev}})$, may be considered free of measurement error and therefore $C(\boldsymbol{\theta}, \boldsymbol{\theta}')$ will not contain a noise variance term σ_n, whereas $K(\boldsymbol{y}, \boldsymbol{y}')$ included σ_n. The posterior predictive distribution of $G(\boldsymbol{\theta})$ is calculated identically to that in (2), but with the hyperparameter values of $\mathbf{C}$ estimated using the queried data $\{G(\boldsymbol{\theta}_{\text{prev}}), \boldsymbol{\theta}_{\text{prev}}\}$. Equivalent to (2), $G(\boldsymbol{\theta}^*)|G(\boldsymbol{\theta}_{\text{prev}})$ is MVN such that

$$\begin{aligned} \boldsymbol{\mu}_G^* &= \mathrm{E}[G(\boldsymbol{\theta}^*)] = \boldsymbol{C}^* \boldsymbol{C}^{-1} G(\boldsymbol{\theta}^*). \\ \boldsymbol{s}_G^* &= \mathrm{Var}[G(\boldsymbol{\theta}^*)] = \boldsymbol{C}^{**} - \boldsymbol{C}^* \boldsymbol{C}^{-1} \boldsymbol{C}^{*T}. \end{aligned} \tag{8}$$

The predictive distribution of (8) is shown in Fig. 5(a), evaluated across the whole search domain, and in Fig. 5(b) at a single query point.

We seek a covariance function $c(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that is representative of the underlying relation between $\boldsymbol{\theta}$ and $G(\boldsymbol{\theta})$. This may seem daunting, considering that the salient aspects of $G(\boldsymbol{\theta})$ include that it is unknown, non-analytic, and expensive to sample. At the same time, BO modelling avoids the common pitfalls of other techniques in modelling $G(\boldsymbol{\theta})$ [28], for example the imputation of a parametric form. Furthermore, just like any other statistical model, we can examine whether our posterior GP is successfully describing the unexplored regions of $G(\boldsymbol{\theta})$. Such a function is described, in detail, in a later section.

C. Bayesian Optimisation Acquisition Function

With the posterior estimate from (8), our preference between different future queries in the hyperparameter space $\boldsymbol{\theta}$ may be expressed as a trade-off between the value we expect at the query, $\boldsymbol{\mu}_G^*$, and the uncertainty around that expectation, $\boldsymbol{s}_G^*$. We formalise this preference via an acquisition function $A(\boldsymbol{\theta}^*)$, see Algorithm 1, line 3ii. Given the current best-found value, G_{best}, popular choices of acquisition functions include probability of improvement (PI) over G_{best} as shown in Fig. 5(c), or the expected improvement (EI) over G_{best}, as shown in Fig. 5(e). We select $A(\boldsymbol{\theta}^*)$ to be EI, since this incorporates the magnitude of improvement, compared to PI in which large and small improvements over G_{best} are weighted equally (where details can

be found in [20]). This means

$$A(\boldsymbol{\theta}^*) := EI(\boldsymbol{\theta}^*) = (G_{\text{best}} - \boldsymbol{\mu}^*_G)\Phi\left(G_{\text{best}}|\boldsymbol{\mu}^*_G, \boldsymbol{s}^*_G\right) + (\boldsymbol{s}^*_G)N\left(G_{\text{best}}|\boldsymbol{\mu}^*_G, \boldsymbol{s}^*_G\right), \quad (9)$$

where Φ and N are the Gaussian cumulative distribution and probability density, respectively. Using the above, BO queries new values if they are near high-performing previous-queries, or if there is considerable uncertainty, which leaves the opportunity for improvement over the current best. Thus, the $\boldsymbol{\theta}^*$ that maximises $A(\boldsymbol{\theta})$ is queried next (Algorithm 1, line 3iii). Note from Fig. 5(c) and 5(e) that $\boldsymbol{\theta}_{\text{new}}$ will differ according to the acquisition function. The result is added as a data point to all queried points (Algorithm 1, line 3iv). The best-found query is then used as the optimial solution (Algorithm 1, line 4), once a stopping criterion is met (typically a computational budget when evaluations are expensive).

D. Tuning Bayesian Optimisation

Bayesian optimisation has a variety of tuneable components, which can be divided between tuning (i) the GP prior over $G(\boldsymbol{\theta})$, and (ii) the acquisition function. The motivation of tuning the GP statistical model is to satisfy the question, "Given the observed data, and my prior knowledge, is the GP over $G(\boldsymbol{\theta})$ appropriately representing my uncertainty about the unexplored areas of the hyperparameter space $\boldsymbol{\theta}$?". The motivation for tuning the acquisition function is to satisfy the question, "Assuming that the GP model $G(\boldsymbol{\theta})$ is correct, what is the wisest next choice to query in $\boldsymbol{\theta}$, in light of my remaining computational budget?" Between (i) and (ii), BO provides many venues to incorporate useful knowledge that may ultimately result in a more efficient sampling of the search domain.

To appropriately tune the GP model of $G(\boldsymbol{\theta})$, we first note that each dimension of the hyperparameter search space $\boldsymbol{\theta}$ varies over different ranges of magnitude. For example, the search space, confined by upper and lower bounds l_d and u_d (described earlier) differ by orders of magnitude, e.g., the l_d and u_d bounds constraining output-scales, h, and those confining length-scales, λ. More importantly, the hyperparameters (of the patient-monitoring GP) are in different units, e.g., h and σ_n are measured in log-HR bpm, whereas λ are measured in minutes. This means that the variation in $G(\boldsymbol{\theta})$ varies by the dimension in which $\boldsymbol{\theta}$ is changing: a change in σ_n of 0.01 log-HR bpm may induce a substantial change in $G(\boldsymbol{\theta})$, whereas a change in λ of 1 minute is almost certain to induce no change in $G(\boldsymbol{\theta})$. A Matérn Automatic Relevance Determination (ARD) kernel

$$c(\boldsymbol{\theta}, \boldsymbol{\theta}') = \eta^2 \left(1 + \sqrt{3r}\right) \exp\left(-\sqrt{3r}\right)$$
$$\text{s.t. } r = \sum_{d=1}^{D} \frac{(\boldsymbol{\theta}_d - \boldsymbol{\theta}'_d)^2}{\nu_d} \quad (10)$$

was selected (and implemented using [29]). Hyperparameters η and ν of (10) may be recognised as the respective analogues to h and λ of (4). This $c(\boldsymbol{\theta}, \boldsymbol{\theta}')$ allows a unique length scale for each dimension of the search domain while remaining enough that the posterior GP can be fit quickly (<1 second for up to 250 observations), which makes the decision time for the next query equally negligible as for RS. The acquisition function was selected to be EI.

To improve data stationarity, values of $G(\boldsymbol{\theta}) < -3$ were left-censored to be -3. Such values represent very rare, extremely poor forecasting performances, usually located at low values of σ_n. By censoring these low values, the BO algorithm remains informed that forecasting performance was low at this point, without hindering GP inference by fitting a stationary GP (such as those described) to highly non-stationary data points. Note that this censoring is only performed for fitting the GP, not for evaluation of performance at that point. Since "it seems strange to first make an ad-hoc transformation, and then use a principled Bayesian probabilistic model" [30], we suggested warped GPs for future work to provide transformations that are more tailored to the data and GP model at hand.

Performance (in finding high-performing $\boldsymbol{\theta}$ for patient monitoring) was further improved by tuning only a subset of the dimensions of $\boldsymbol{\theta}$ at each BO iteration. The sequential BO (SBO) of parameters, optimises only a subset, $\mathcal{D}$, of $\boldsymbol{\theta}$'s D total dimensions by holding the remaining values fixed at those of the current best found $\boldsymbol{\theta}$. That is, SBO can be achieved simply by changing line 3iii of Algorithm 1 to

$$\textbf{query } G(\boldsymbol{\theta}) \text{ at } \boldsymbol{\theta}_{\text{new}} := \arg\max_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$$
$$\text{s.t. } \boldsymbol{\theta}_d = \boldsymbol{\theta}_d^{\text{best}} \; \forall d \notin \mathcal{D}. \quad (11)$$

Subset $\mathcal{D}$ can be chosen, e.g., to simultaneously tune length-scales ($\lambda_{i=1,..,a}$), variances ($h_{i=1,..,a}$ and σ_n), hyperparameters of a specific kernel (ex. h_i and λ_i), at random, or any other selection criterion.

VII. Results with the Training Set

The primary purpose of using the training set was to identify which optimisation method was most successful in identifying high-performing values of $\boldsymbol{\theta}$. As shown in Fig. 6, RS was unable to take advantage of increasingly complex GP kernels. As more kernel hyperparameters needed to be considered, the effective dimensionality of the search space increased (from a 3-dimensional search for k_1 to a 7-dimensional search for k_3). Although, RS successfully optimised k_1, its performance rapidly deteriorated in k_2 and k_3. Since k_1 can, effectively, be achieved in k_2 and k_3 by setting $h_2 = h_3 = 0$, it is clear that RS has difficulty finding the best values merely by enumerating random possibilities. In other words, for any patient for whom k_1 truly is the best kernel choice, an effective optimiser would learn to remove the additional kernels by setting the appropriate hyperparameters to 0. In this case, random enumeration was insufficient to learn this option while simultaneously learning good parameterisations for k_1. As the performance of SBO demonstrates, the additional kernels could be used to further optimise the objective function, so long as the hyperparameter space is properly explored.

SBO improved performance as the complexity of the search space increased. Optimising over objective G_1 was sufficient to identify good $\boldsymbol{\theta}$ for G_1, G_2, and G_3, suggesting that the

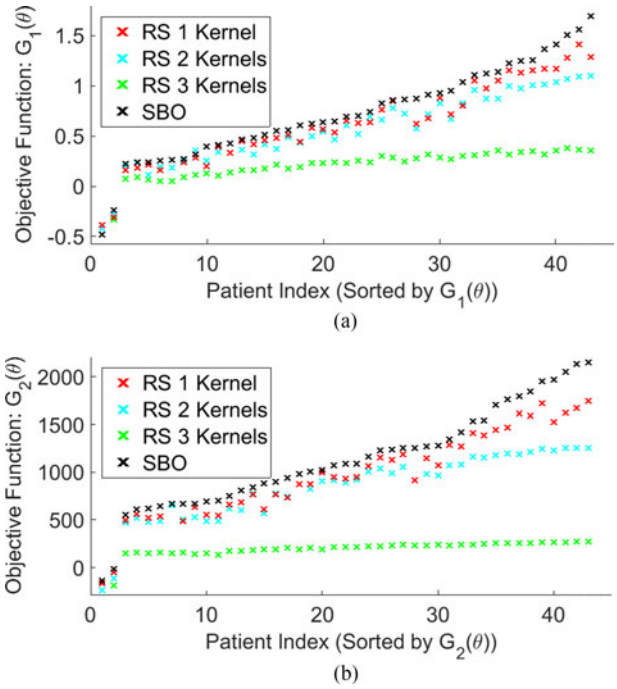

Fig. 6. Best-found values to optimise (a) G_1 and (b) G_2 using data from the training set patients. The SBO method optimising G_2 not only finds superior values in its specific task, maximising G_2 for 30-minute forecast, but also for tasks it is not optimising, such as (b) G_2 for 60 minute look-ahead. This performance reflects both G_2's regularising properties of vectorising multiple objectives, and SBO's ability to sample in promising regions of the parameter space $\boldsymbol{\theta}$.

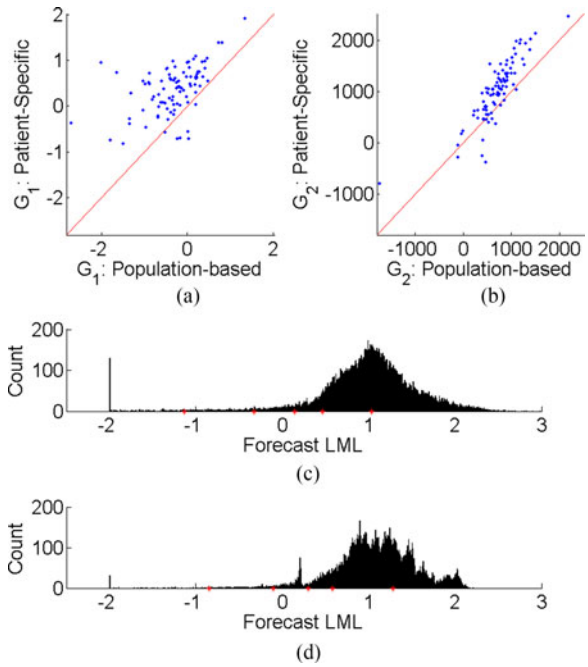

Fig. 7. Forecast performance for test set patients during hours 24–72. Improvement is shown on a per-patient basis for (a) G_1 and (b) G_2. The aggregated forecast performance of (c) uninformative priors can be compared to (d) patient-specific $\boldsymbol{\theta}$'s. Forecast LMLs were left-censored at -2 for visual-clarity. The forecasts LML values in (c) have a significant mode at -2 due to this censoring, where the forecast LML values in (d) do not. For (c) and (d), red points mark the 1, 2.5, 5, 10, and 50-percentiles.

weighting scheme for G_2, while simple, successfully emphasised performance at lower quantiles of forecast likelihood. The SBO search not only discovered values of $\boldsymbol{\theta}$ that optimised 30-minute forecasts but also forecast looking ahead 5–60 minutes, suggesting that the solutions were not myopic to a single physiological time-scale. Results were similar for optimisation using objective G_2, but less-pronounced than those from G_1. From the results of the training set, an SBO optimisation search over G_1 was selected to compete against uniformative priors in the test set, as described below.

VIII. Results with the Test Set

Forecasts via SBO-derived values of $\boldsymbol{\theta}$ were then compared to forecasts via values of $\boldsymbol{\theta}$ regularised via uninformative priors on a test set of 126 patients. SBO learned from the first 24 hours of the patient's stay on the ward, with 250 queries of $G_1(\boldsymbol{\theta})$ to identify an optimal $\boldsymbol{\theta}$, for use when modelling new data from subsequent days. Fig. 7 shows forecast performance of the SBO-found optimal $\boldsymbol{\theta}$, subsequent to the first 24 hours.

Compared to the population-based uninformative priors, SBO-tuned GPs demonstrated superior performance, both in G_1, in which they were trained, and the G_2 objective function. As seen in Fig. 7(a), for G_1, on a patient-by-patient basis, only 6 of 126 patients did not benefit from personalised SBO parameter tuning. The 6 without improvement tended to have only small amounts of data on which to learn and test SBO's efficacy. Furthermore, the performance reduction in 2 of the 6 cases was negligible. In contrast, SBO training resulted in significant gains for most patients. Most importantly, as seen by the left-side of Fig. 7(a), the largest gains were made for patients with the worst forecasts under population-based regularisation. These would be those patients most likely to generate alarms. Similar trends can be seen in Fig. 7(b). Across forecast depths of 1–45 minutes, patient-specific improvement was as good or better than those shown in Fig. 7. Results were slightly worse at a forecast depth of 60. On aggregate, Fig. 7(c) and (d), the lowest forecast LMLs were improved when using SBO. The population-based regulariser had hundreds of worst-case forecasts (left-censored at -2), whereas the patient-specific models had tens of instances of forecasts at that low level of LML.

IX. Conclusion

We have shown that patient-specific regularisation improves the robust forecasting of patient vital signs, using heart rate as an exemplar. Several optimisation methods to learn these regularisers were presented, along with ideas concerning how such methods might be tuned. A useful vectorisation of multiple objectives was also learned, and we have demonstrated its robustness at optimising different objectives at different forecast lengths.

Future work for personalised regularisation will first need to confront the challenges of optimising over a space with high ef-

fective dimensionality, due to the correlation of hyperparameter values within and between kernels. In particular, it is desirable to learn a way for simple-to-code methods, such as RS, to be competitive (if at all possible) in the optimisation environment, which they currently are not when implemented naively. The benefits of using more sophisticated, BO-based, optimisation methods were significant, suggesting that the most promising future work will build on these methods. Particular priorities may include (i) warped BO to handle non-stationary data without resort to heuristics [30], (ii) dimensionality reduction techniques to better search across highly correlated dimensions, and (iii) kernels that will encode effectively-identical domain coordinates (e.g., and value of λ_i will perform the same when $h_i = 0$). These steps will be invaluable when optimising over hyperprior distributions, which would improve the flexibility of the patient vital sign model, but increase the size and complexity of the search domain. Future work could also consider multitask GPR, both for patient monitoring across vital signs [14] (e.g., optimising models learning the correlation between HR, BR, and SpO_2), and to facilitate BO across multiple objectives [31] (e.g., so that G_1 and G_2 are not optimised independently).

Finally, these methods must be validated for their most important clinical uses, which are (i) the detection of deteriorating patients and (ii) physiologically-interpretable regularisers.

References

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

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

[3] G. W. Colopy, M. A. F. Pimentel, S. J. Roberts, and D. A. Clifton, "Bayesian Gaussian processes for identifying the deteriorating patient," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2016, pp. 5311–5314.

[4] M. Pimentel, "Modelling of vital-sign data from post-operative patients," Ph.D. dissertation, Univ. Oxford, Oxford, U.K., 2015.

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

[6] G. W. Colopy, M. A. F. Pimentel, S. J. Roberts, and D. A. Clifton, "Bayesian optimisation of Gaussian processes for identifying the deteriorating patient," in *Proc. 2017 IEEE EMBS Int. Conf. Biomed. Health Informat.*, Feb. 2017, pp. 85–88.

[7] G. W. Colopy, T. Zhu, L. Clifton, S. J. Roberts, and D. Clifton, "Likelihood-based artefact detection in continuously-acquired patient vital signs," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jul. 2017, pp. 2146–2149.

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

[9] D. Clifton, D. Wong, L. Clifton, R. Pullinger, and L. Tarassenko, "A large-scale clinical validation of an integrated monitoring system in the emergency department," *IEEE Trans. Inf. Technol. Biomed.*, vol. 17, no. 4, pp. 835–877, Jul. 2013.

[10] L. Clifton, D. Clifton, M. A. Pimentel, P. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.

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

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

[13] L. Clifton, D. Clifton, M. Pimentel, P. Watkinson, and L. Tarassenko, "Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 3, pp. 722–730, May 2014.

[14] R. Duerichen, M. Pimentel, L. Clifton, A. Schweikard, and D. Clifton, "Multi-task Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, Jan. 2015.

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

[16] C. Williams, J. Quinn, and N. Mcintosh, "Factorial switching Kalman filters for condition monitoring in neonatal intensive care," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1513–1520.

[17] J. Quinn, C. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1537–1551, Sep. 2009.

[18] E. Choi, M. T. Bahadori, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. 2016 Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.

[19] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Philos. Trans. Roy. Soc. Lond. A, Math., Phys. Eng. Sci.*, vol. 371, no. 1984, 2012, Art. no. 20110550.

[20] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[21] M. Ebden, "Gaussian processes: A quick introduction," arXiv, May 12, 2015. [Online]. Available: http://arxiv.org/abs/1505.02965v2

[22] I. Murray, R. Adams, and D. MacKay, "Elliptical slice sampling," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9, May 2010, pp. 541–548.

[23] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[24] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "Bayesian modeling with Gaussian processes using the GPstuff toolbox," arXiv, Jun. 25, 2012. [Online]. Available: http://arxiv.org/abs/1206.5754v6

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[26] F. J. Solis and R. J.-B. Wets, "Minimization by random search techniques," *Math. Oper. Res.*, vol. 6, no. 1, pp. 19–30, 1981.

[27] J. Bergstra and Y. Benjio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[28] D. Jones, "A taxonomy of global optimization methods based on response surfaces," *J. Global Optim.*, vol. 21, pp. 345–383, 2001.

[29] C. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn. Res.*, vol. 11, pp. 3011–3015, Nov. 2010.

[30] E. Snelson, Z. Ghahramani, and C. E. Rasmussen, "Warped Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 337–344.

[31] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task bayesian optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2004–2012.