Personalized Risk Scoring for Critical Care Prognosis using Mixtures of Gaussian Processes

Ahmed M. Alaa, *Member, IEEE*, Jinsung Yoon, Scott Hu, *MD*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—*Objective*: In this paper, we develop a personalized real-time risk scoring algorithm that provides timely and granular assessments for the clinical acuity of ward patients based on their (temporal) lab tests and vital signs; the proposed risk scoring system ensures timely intensive care unit (ICU) admissions for clinically deteriorating patients. *Methods*: The risk scoring system is based on the idea of *sequential hypothesis testing under an uncertain time horizon*. The system learns a set of latent patient *subtypes* from the offline electronic health record data, and trains a mixture of *Gaussian Process (GP) experts*, where each expert models the physiological data streams associated with a specific patient subtype. Transfer learning techniques are used to learn the relationship between a patient's latent subtype and her static admission information (e.g. age, gender, transfer status, ICD-9 codes, etc). *Results*: Experiments conducted on data from a heterogeneous cohort of 6,321 patients admitted to Ronald Reagan UCLA medical center show that our score significantly outperforms the currently deployed risk scores, such as the Rothman index, MEWS, APACHE and SOFA scores, in terms of timeliness, true positive rate (TPR), and positive predictive value (PPV). *Conclusion*: Our results reflect the importance of adopting the concepts of personalized medicine in critical care settings; significant accuracy and timeliness gains can be achieved by accounting for the patients' heterogeneity. *Significance*: The proposed risk scoring methodology can confer huge clinical and social benefits on a massive number of critically ill inpatients who exhibit adverse outcomes including, but not limited to, cardiac arrests, respiratory arrests, and septic shocks.**

***Index Terms*—Critical care medicine, Sequential Hypothesis testing, Personalized Medicine, Prognosis.**

I. Introduction

CRITICALLY ill patients who are hospitalized in regular wards with brain tumors, hematological malignancies, neutropenia, or those who are recipients of stem cell transplants, or upper-gastrointestinal surgeries, are vulnerable to a wide range of adverse outcomes, including neurologic conditions [1], septic shocks [2], post-operative complications [3]–[10], cardiopulmonary arrest [11], [12], and acute respiratory failure [13]. All these adverse events can lead to an unplanned ICU transfer [4], the timing of which plays a major role in determining clinical outcomes[1]; recent medical studies have confirmed that the efficacy of acute care interventions depends substantially on the timeliness of their application [8], [13], [15]. The problem of delayed ICU transfer is enormous: over 750,000 septic shocks and 200,000 cardiac arrests occur in the U.S. each year with mortality rates of 28.6% and 75% respectively [16], [17]. Fortunately, experts believe that much of these events could be prevented with accurate *prognosis* and early warning [18].

A. Summary of Contributions

To address the problem above, we develop a risk scoring algorithm that provides real-time, personalized assessments for the acuity of critical care patients in a hospital ward. The algorithm is trained using the electronic health record (EHR) data in an offline stage, and risk scores for a newly hospitalized patient are computed via the trained model in real-time using her temporal, irregularly sampled physiological measurements. The proposed risk scoring methodology is based on the idea of *sequential hypothesis testing under an uncertain time horizon*. That is, we view a patient's risk score as the optimal test statistic of a sequential hypothesis test that disentangles clinically stable patients from the clinically deteriorating ones as more physiological measurements are gathered over time. The sequential hypothesis test is based on a *non-stationary* model for the deteriorating patients' physiological time series. (Non-stationarity creates uncertainty in the latent time horizon of the patient's physiological time series.) Our conception of the risk score advances on the seminal work of Wald on sequential analysis [19], and is logically related to *optimal stopping* problems in the areas of finance and automatic control [20].

The underlying patient's physiological streams, based on which the sequential test is conducted, are modeled as multitask Gaussian Processes (GP) [21], [22], the hyper-parameters of which depend on the patient's (latent) clinical status, i.e. the true hypothesis of whether the patient is clinically stable or deteriorating. We capture the non-stationarity of the deteriorating patients' physiological streams by dividing every patient's stay in the ward into a sequence of temporal *epochs*, and allow the parameters of the multitask GP to vary across these epochs. Non-stationarity

A. Alaa, J. Yoon and M. van der Schaar are with the Department of Electrical Engineering, University of California, Los Angeles (UCLA), CA, 90095, USA (e-mail: ahmedmalaa@ucla.edu, jsyoon0823@ucla.edu, mihaela@ee.ucla.edu).

S. Hu is with the Division of Pulmonary and Critical Care Medicine, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles (UCLA), CA, 90095, USA (email: scotthu@mednet.ucla.edu).

[1] According to the *Joint Commission* (a nonprofit organization that accredits hospitals and gathers data related to adverse events), around 29% of (narcotic-related) bedside adverse events reported during the period from 2004 to 2011 were resulting from improper post-operative (or pre-operative) monitoring of patients [14]

is taken into account in the training phase by temporally aligning the physiological streams recorded in the EHR data, and is taken into account in the real-time deployment phase by repeatedly estimating the multitask GP epoch index over time.

The heterogeneity of the patients' population is captured by considering the patients' latent subtypes (or *phenotypes* [23]). The proposed algorithm discovers the number of patient subtypes from the training data, and learns a separate multitask GP model for the physiological streams associated with each subtype. Unsupervised discovery of the patients' latent subtypes is carried out using the expectation-maximization (EM) algorithm applied on the domain of clinically stable patients since these patients are dominant in the dataset, and are more likely to exhibit stationary physiological trajectories, thus their physiological streams are described with few hyper-parameters and can be efficiently estimated. The knowledge of the patients' latent subtypes –extracted from the domain of clinically stable patients– is then transferred to the domain of clinically deteriorating patients via *transfer learning*. Every GP model associated with (stable or deteriorating) patients who belong to a specific subtype is called a *GP expert*. Thus, every GP expert specialized in scoring the risk for one of the discovered patient subtypes. A patient's risk score is the optimal statistic of a sequential test, i.e. a weighted average of the posterior beliefs of all GP experts about the patient's clinical status given her physiological data stream, where the weights are computed based on the patient's hospital admission information (e.g. age, ICD-9 codes, etc), which we estimate using (transductive) *transfer learning*.

Experiments were conducted using a dataset for a heterogeneous cohort of 6,321 patients who were admitted during the years 2013-2016 to a general medicine floor in the Ronald Reagan UCLA medical center, a tertiary medical center. Results show that the proposed risk score consistently outperforms the currently deployed clinical scores in terms of timeliness and accuracy (i.e. the true positive rate (TPR) and the positive predictive value (PPV)), in addition to state-of-the-art machine learning algorithms that are based on *sliding-window* regression. Our results show that the proposed risk score boosts the AUC with 12% as compared to the Rothman index (the current technology deployed in our medical center), and can prompt alarms for ICU admission 12 hours before clinicians (on average) for a PPV of 25% and TPR of 50%, which provides the ward staff with a safety net for patient care by giving them sufficient time to intervene at an earlier time in order to prevent clinical deterioration. Moreover, the proposed risk score reduces the number of false alarms per number of true alarms for any setting of the TPR, which reduces the alarm fatigue and allows for better hospital resource management.

B. Related Works

Hospitals have been recently investing in prognostic risk scoring systems for critically ill patients in wards [3]–[9]. However, recent systematic reviews have shown that currently deployed expert-based risk scores, such as the MEWS score [24], provide only modest contributions to clinical outcomes [25]–[27]. To that end, a data-driven risk score, named the *Rothman index*, has been developed using regression analysis [5], and was shown to outperform the MEWS score and its variants [9]. Nevertheless, the Rothman index lacks a principled model for the hospitalized patient's physiological parameters, and is mainly constructed using a "one-size-fits-all" approach that leaves no room for personalized risk assessment that is tailored to the individual patient. A comprehensive, tabulated review of all the clinical scores used for ward patients is provided in Appendix A of the online supporting document in http://medianetlab.ee.ucla.edu/papers/Alaa_TBME_supp.pdf.

The problem of modeling multivariate physiological time series has been recently investigated by the machine learning community [2], [6], [10], [21], [22], [28]–[30]; some of the previous works have also adopted multitask GP models [6], [10], [21], [22], [28]. However, most of these works have focused on a *forecasting* problem in which the goal is to predict the future values of an observable bio-marker. For instance, [28] focuses on predicting the PFVC clinical marker (a measure of lung severity) for scleroderma patients, [6], [10], [21], [22] focus on predicting the future values of SOFA, APACHE and SAPS scores for ICU patients, and [30] focuses on predicting the GFR bio-marker values for patients with chronic kidney disease. Unfortunately, a major challenge encountered in our setting is that patients in regular wards have no such strongly indicative bio-markers; we face this challenge by resorting to a *latent class* modeling approach, in which different classes correspond to different severity states. Our model adopts two latent classes, which allows the risk scoring problem to be formulated as a *sequential hypothesis test* [19]. Consequently, our multitask GP model serves as a tool for computing the optimal test statistic, and not for performing GP regression as it is the case in the forecasting problems in [6], [10], [21], [22], [28]. To the best of our knowledge, our model is the first to conceptualize real-time risk scoring as a sequential testing procedure.

Our risk scoring model handles the heterogeneity of the patients' population via *subtyping*. Unlike previous works on subtyping in longitudinal disease progression models [28], [30], in which one set of subtypes is learned for the entire population of "sick" patients, the nature of the critical care setting (manifesting in our sequential testing framework) entails the need for learning different sets of subtypes for both clinical stability and deterioration. This imposes the challenge of learning a separate set of subtypes for the clinically deteriorating patients under class imbalance (ICU admission rate is less than 10%); we face this challenge via a novel learning algorithm that uses ideas from transfer learning to transfer the knowledge learned from the clinical stable population to the deteriorating population.

Most of the previous works on clinical risk prognosis used clinical endpoints (ICU admission or discharge) as "*surrogate labels*" for a patient's clinical deterioration, and hence used those labels to train a supervised (regression) model using the physiological data in a fixed-size time window before censoring. The supervised models used in the literature included logistic regression [31], [32] and SVMs [33]. We compare the performance of our model with these methods in Section IV. A detailed, tabulated comparisons with other risk scoring methodologies is provided in Appendix A in the supporting document. This paper builds on our previous work in [34] by adding deterioration and stability subtypes, and conducting experiments on a larger patient cohort. Our work has been presented in part in [35]; this paper extends on the model therein by incorporating model non-stationarity, developing new learning algorithms, and including more experimental details.

II. The Physiological Model

In this section, we present a comprehensive model for the patients' physiological data and develop a rigorous formulation for the risk scoring problem.

A. Modeling the Patients' Risks and Clinical Status

Two types of information are associated with every patient in the (surgical or medical) ward:

1- Physiological information $X(t)$: We define $X(t) = [X_1(t), X_2(t), \ldots, X_D(t)]^T$ as a D-dimensional stochastic process representing the patient's D physiological streams (lab tests and vital signs) as a function of time. The process $X_i(t)$ takes values from a space $\mathcal{X}_i$, and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots, \times \mathcal{X}_D$. Vital signs and lab tests are gathered at arbitrary time instances $\{t_{ij}\}_{i=1,j=1}^{D,M_i}$ (where $t = 0$ is the time at which the patient is admitted to the ward), where M_i is the total number of samples of vital sign (or lab test) i that where gathered during the patient's stay in the ward. Thus, the set of all observations of the physiological data that the ward staff has for a specific patient is given by $\{X_i(t_{ij})\}_{i=1,j=1}^{D,M_i}$, and we will refer to the realizations of these variables as $\{x_{ij}, t_{ij}\}_{ij}$.

2- Admission information Y: We define the S-dimensional random vector Y as the patient's static information obtained at admission (e.g. age, gender, ICD9 code, etc). The random vector Y is drawn from a space $\mathcal{Y}$, and we denote the realizations of the patient's static information as $Y = y$. Thus, the set of all (static and time-varying) information associated with a patient can be gathered in a set $\{y, \{x_{ij}, t_{ij}\}_{ij}\}$.

Let $V \in \{0, 1\}$ be a binary latent variable that corresponds to the patient's true clinical status; 0 standing for a stable clinical status, and 1 for a clinically deteriorating status. Since physiological streams manifest the patients' clinical statuses, it is natural to assume that the conditional distributions of $X^o(t) = X(t)\,|V = 0$ differ from that of $X^1(t) = X(t)\,|V = 1$. We assume that V is drawn randomly for every patient at admission time and stays fixed over the patient's stay in the ward, i.e. the value of V is revealed at the end of every physiological stream, where $V = 1$ if the patient is admitted to the ICU, and $V = 0$ if the patient is discharged home. During the patient's stay in the ward, the ward staff members are confronted with two hypotheses: the null hypothesis $\mathcal{H}_o$ corresponds to the hypothesis that the patient is clinically stable, whereas the alternative hypothesis $\mathcal{H}_1$ corresponds to the hypothesis that the patient is clinically deteriorating, i.e.

$$V = \begin{cases} 0: \ \mathcal{H}_o \text{ (clinically stable patient)}, \\ 1: \ \mathcal{H}_1 \text{ (clinically deteriorating patient)}. \end{cases} \quad (1)$$

Thus, the prognosis problem is a *sequential hypothesis test* [19], i.e. the clinicians need to reject one of the hypotheses at some point of time after observing a series of physiological measurements. Hence, following the seminal work of Wald in [19], we view the patient's risk score as the test statistic of the sequential hypothesis test. That is, the patient's risk score at time t, which we denote as $\bar{R}(t) \in [0, 1]$, is the posterior probability of hypothesis $\mathcal{H}_1$ given the observations $\{x_{ij}, t_{ij} \leq t\}_{ij}$, and we have that $\bar{R}(t) = \mathbb{P}\left(\mathcal{H}_1 \left|\{x_{ij}, t_{ij} \leq t\}_{ij}\right.\right)$, i.e.

$$\bar{R}(t) = \frac{\mathbb{P}\left(\{x_{ij}, t_{ij} \leq t\}_{ij} \left|\mathcal{H}_1\right.\right) \cdot \mathbb{P}\left(\mathcal{H}_1\right)}{\sum_{v \in \{0,1\}} \mathbb{P}\left(\{x_{ij}, t_{ij} \leq t\}_{ij} \left|\mathcal{H}_v\right.\right) \cdot \mathbb{P}\left(\mathcal{H}_v\right)}, \quad (2)$$

where $\mathbb{P}\left(\mathcal{H}_1\right)$ is the prior probability of a patient in the ward being admitted to the ICU (i.e. the rate of ICU admissions).

B. Modeling the Physiological Signals

Since the vital signs and lab tests are gathered at arbitrary, irregularly sampled time instances, it is convenient to adopt a continuous-time model for the patients' physiological stream using GPs [21], [22], [36]. We model the D (potentially correlated) physiological streams of a monitored patient as a multitask GP defined over $t \in \mathbb{R}_+$. The model parameters depend on the patient's latent clinical status V. Since clinically stable patients do not exhibit changes in their clinical status, we adopt a stationary model for $X^o(t)$. Contrarily, deteriorating patients pass through phases of clinical acuity, which invokes the need for a non-stationary model for $X^1(t)$. In the following, we present the physiological models for clinically stable and deteriorating patients, which we will then use as a proxy for risk scoring in the next Section.

Physiological Signals Model for Clinically Stable Patients

For clinically stable patients, i.e. $V = 0$, we adopt a multitask GP model for the physiological signal $X^o(t)$ as follows

$$X^o(t) \sim \mathcal{GP}(m_o(t), k_o(i, j, t, t^{'})), \quad (3)$$

where $m_o(t) : \mathbb{R}^+ \rightarrow \mathcal{X}$ is the *mean function*, and $k_o(i, j, t, t^{'}) : \mathcal{X}_i \times \mathcal{X}_j \times \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}_+$ is the *covariance kernel*. The mean function is assumed to be a constant vector, i.e. $m_o(t) = [m_o^1, m_o^2, \ldots, m_o^D]^T$, the entries of which represent the average value of the different physiological streams. We assume that the covariance kernel matrix $k_o(i, j, t, t^{'})$ has the following separable form [36]

$$k_o(i, j, t, t^{'}) = \mathbf{\Sigma}_o(i, j)\, k_o(t, t^{'}), \quad (4)$$

where $\mathbf{\Sigma}_o$ is a *stationary correlation matrix* that quantifies the correlations between the various physiological streams. The kernel function $k_o(t,t^{'})$ is *squared-exponential* kernel [21], [37], [38], defined as

$$k_o(t,t^{'}) = \omega_o^2 \, e^{-\frac{1}{2\ell_o^2}\,||t-t^{'}||^2}, \tag{5}$$

where ω_o and ℓ_o are hyper-parameters: ω_o^2 is the *variance hyper-parameter*, and ℓ_o is the *characteristic length-scale*. The parameter ω_o controls the dynamic range of the fluctuations of $X(t)$; the parameter ℓ_o controls the rate of such fluctuations. Note that (4) implies that we assume that all the physiological streams have the same temporal characteristics, i.e. the same variance and characteristic length-scale.

Since the correlation matrix $\mathbf{\Sigma}_o$ needs to be positive semi-definite, we adopt the "free-form" construction of the correlation matrix via the Cholesky decomposition as follows

$$\mathbf{\Sigma}_o = \mathbf{L}_o\,\mathbf{L}_o^T,\, \mathbf{L}_o = \begin{bmatrix} \sigma_{o,1} & 0 & \cdots & 0 \\ \sigma_{o,2} & \sigma_{o,3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{o,\bar{D}-m+1} & \sigma_{o,\bar{D}-m+2} & \cdots & \sigma_{o,\bar{D}} \end{bmatrix}, \tag{6}$$

where $\bar{D} = \frac{D(D+1)}{2}$ [36]. Since the variance of each stream is already captured by the entries of $\mathbf{\Sigma}_o$, we assume that $\omega_o = 1$ for all streams. Thus, the hyper-parameters that characterize a multi-task GP $\mathcal{GP}(m_o(t), k_o(i,j,t,t^{'}))$ are ℓ_o and the entries of $\mathbf{L}_o$, which we compactly write in a vector $\sigma_o = [\sigma_{o,j}]_{j=1}^{\bar{D}}$.

We summarize the parameters of the GP model capturing the physiological streams of clinically stable patients via the following parameter set

$$\mathbf{\Theta}_o = \{\{m_o^d\}_{d=1}^D,\ \ell_o,\ \sigma_o\}, \tag{7}$$

which aggregates the $\frac{D(D+1)}{2} + D + 1$ hyper-parameters of the multi-task GP. We write $X^o(t) \sim \mathcal{GP}(\mathbf{\Theta}_o)$ to denote an instance of a physiological stream of a clinically stable patient generated with a parameter set $\mathbf{\Theta}_o$.

Physiological Signals Model for Clinically Deteriorating Patients

For clinically deteriorating patients, i.e. patients with $V = 1$, we adopt a non-stationary model for $X^1(t)$ specified as follows

$$X^1(t) \sim \mathcal{GP}(\mathbf{\Theta}_1), \tag{8}$$

where $\mathbf{\Theta}_1$ is the parameter set for the physiological streams of deteriorating patients. Since deteriorating patients exhibit changes in their clinical status (e.g. progression from a more stable status to a less stable one), a stationary covariance kernel, such as the one defined in (5), and a constant mean function do not suffice to describe the physiological stream of a deteriorating patient. This motivates a non-stationary model for $X^1(t)$ that divides the time domain into a sequence of *epochs*, each is of duration T_1, and is associated with a distinct constant mean function and a distinct squared-exponential covariance kernel.

Let $T = K \cdot T_1$ be the maximum duration for a patient's stay in the ward. That is, the patient passes through K consecutive epochs, each of which has a mean function and a covariance kernel parametrized by $\mathbf{\Theta}_1^k = \{\{m_{1,k}^d\}_{d=1}^D,\ \ell_{1,k},\ \sigma_{1,k}\}, \forall k \in \{1,2,\ldots,K\}$. Since patients arrive at the hospital ward at random time instances, at which the clinical status is unknown, we define $\bar{k} \in \{1,2,\ldots,K\}$ as the unobservable, initial epoch index, which we assume to be drawn from an unknown distribution $\bar{k} \sim f_k(k)$. The physiological measurements gathered by the clinicians during the patient's are governed by a monotonically increasing sequence of epochs, i.e. the clinicians observe physiological measurements drawn from a process with the underlying epoch sequence $\{\bar{k},\bar{k}+1,\ldots,K\}$. For instance, if $K = 6$ and the realization of $\bar{k}$ is 3, then the (deteriorating) patient's physiological process $X^1(t)$ has its parameters changing over time according to the epoch sequence $\{3,4,5,6\}$. Note that the length of the patient's stay in the ward is given by $(K - \bar{k} + 1)\cdot T_1$, which is random since $\bar{k}$ is a random variable.

We assume that the physiological measurements across different epochs are independent, but measurements within the same epoch are correlated. Thus, the vital signs and lab tests are correlated within every interval in the set of intervals $\{[0,T_1),[T_1,2T_1),\ldots,[(K-\bar{k})\,T_1,(K-\bar{k}+1)\,T_1)\}$, but are uncorrelated across different time intervals. In other words, the covariance kernel for the process $X^1(t)$ is given by

$$k_1(i,j,t,t^{'}) = \begin{cases} \mathbf{\Sigma}_{1,k}(i,j)\,k_{1,k}(t,t^{'}),\ \forall t,t^{'} \in [t_1,t_2), \\ 0,\ \text{Otherwise} \end{cases} \tag{9}$$

where $[t_1,t_2) \in \{[0,T_1),\ldots,[(K-\bar{k})\,T_1,(K-\bar{k}+1)\,T_1)\}$, and

$$k_{1,k}(t,t^{'}) = \omega_{1,k}^2\, e^{-\frac{1}{2\ell_{1,k}^2}\,||t-t^{'}||^2}. \tag{10}$$

The parameters of the GP model for deteriorating patients can be summarized via the following parameter set

$$\mathbf{\Theta}_1 = \{\{m_{1,k}^d\}_{d=1}^D,\ \ell_{1,k},\ \sigma_{1,k}\}_{k=1}^K. \tag{11}$$

The parameter set $\mathbf{\Theta}_1$ encapsulates $K\left(\frac{D(D+1)}{2} + D + 1\right)$ hyper-parameters that describe the process $X^1(t)$. Note that the model $X^1(t)$ entails much more parameters than the model $X^o(t)$, which poses a significant challenge in learning the parameters of $X^1(t)$. We address this challenge elaborately in the next Section.

C. Modeling Patients' Subtypes

The model presented so far is constructed in a "one-size-fits-all" fashion. That is, the risk score computed in (2) considers the vital signs and lab tests for the monitored patient, without considering her baseline admission information (the vector Y). The interpretation of the manifest variables $\{x_{ij},t_{ij}\}_{ij}$ in terms of the risk for clinical deterioration may differ depending on the patient's age, gender, transfer status, or clinical history. Thus, a risk score that is tailored to the individual's admission feature would ensure a higher level

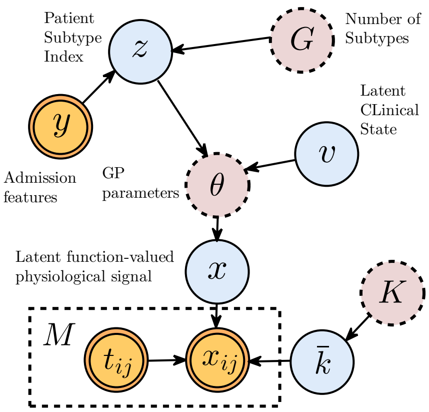


Fig. 1: Graphical model for the physiological signals.

of granularity in modeling the physiological signals, which would lead to a more accurate prognosis.

In order to ensure that our risk score is "personalized", we model the heterogeneity of the patients' population by incorporating a *subtype* variable $Z \in \mathcal{Z} = \{1, 2, \ldots, G\}$, which indicates the patient's latent *phenotype* which determines her physiological behavior, where G is the number of subtypes to which a patient may belong. That is, every patient has her physiological behavior being determined by both her clinical status and her latent subtype. We denote risk scores that take the patient's particular subtype into account as *"personalized risk scores*".

The influence of the patient's subtype Z on the patient's physiological model is captured by the following relations

$$Z \perp V \,|\, Y, \; V \perp Y \,|\, Z, \tag{12}$$

where $\perp$ denotes conditional independence. The relations in (12) imply that: (a) a patient's subtype is independent of her clinical status given her admission information, and (b) a patient's clinical status is independent of the admission information given her subtype. That is, knowledge of the patent's admission information suffices to infer her subtype (e.g. knowledge of age and gender, etc, is enough to know the subtype to which a patient belongs irrespective of the true clinical status), and knowledge of the patient's subtype is enough to infer the patient's vulnerability irrespective to the admission information. The first relation follows from the fact that the patient's subtype is an intrinsic feature of the patient that is independent of her clinical acuity, whereas the second relation follows from that fact that the information contained in Y is a subset of the information contained in the patient's intrinsic subtype Z.

The patient's subtype manifests in her physiological signals by manipulating the parameter sets for the multitask GPs representing both $X^o(t)$ and $X^1(t)$. In other words, the parameters of the multitask GP modeling the patient's physiological signal depends not only on her clinical status V, but also on her subtype Z. The parameter set for clinically deteriorating patients is denoted as $\mathbf{\Theta}^z_1$, and the parameter set for stable patients is denoted as $\mathbf{\Theta}^z_o$, where $Z = z$ is a realization for the patient's subtype. The construction of both parameter sets follows the description provided in the previous subsection. Therefore, the physiological signals for the patients in the ward are generated as follows

$$X^v(t) \,|Z = z \sim \mathcal{GP}(\mathbf{\Theta}^z_v). \tag{13}$$

Fig. 1 depicts a graphical model describing the generative process for the patients' physiological signals. The patient's subtype $Z = z$ is hidden, and affects both her clinical status $V = v$ and the physiological behavior that manifests in the vital signs and lab tests. The variable $\bar{V} \in \{0,1\} \times \mathcal{Z}, \bar{V} = [V \; Z]^T$ augments both the patient's subtype and clinical status; a realization of this variable $\bar{V} = \bar{v}$ determines the parameter set $\theta|\bar{v} = \mathbf{\Theta}^z_v$, which is used to generate a latent function-valued variable $X(t) = x \in \mathbb{R}^{\mathbb{R}^D}$. A plate model is then used to describe the sequence of measurements $\{x_{ij}\}_{i,j}$ gathered by the clinicians at time instances $\{t_{ij}\}_{i,j}$. The time instances $\{t_{ij}\}_{i,j}$ are assumed to be exogenously determined by the ward staff and are uninformative of the clinical status, hence they are modeled as parent nodes in the graphical models. Observations are influenced by the index of the first epoch, $\bar{k}$, which is also assumed to be exogenously determined by the patient's arrival to the ward. It can be seen that the probabilistic influences among the variables V, Z and Y in the graphical model in Fig. 1(c) capture the relations specified in (12).

Having defined the patients' subtypes, we refine the definition of the (non-personalized) risk score $\bar{R}(t)$, and incorporate the patient's individual static features in a personalized risk score $R(t, y)$ as follows

$$\begin{aligned} R(t, y) &= \mathbb{P}\left(\mathcal{H}_1 \,|\{x_{ij}, t_{ij}\}_{i,j}, Y = y\right) \\ &= \sum_{z \in \mathcal{Z}} \mathbb{P}\left(\mathcal{H}_1 \,|\{x_{ij}, t_{ij}\}_{i,j}, Z = z\right) \cdot \mathbb{P}(Z = z|Y = y) \\ &= \sum_{z \in \mathcal{Z}} \mathbb{P}\left(V = 1 \,|\{x_{ij}, t_{ij}\}_{i,j}, \mathbf{\Theta}^z_o, \mathbf{\Theta}^z_1\right) \cdot \mathbb{P}(Z = z|Y = y), \end{aligned} \tag{14}$$

where

$$\mathbb{P}\left(V = 1 \,|\{x_{ij}, t_{ij}\}_{i,j}, \mathbf{\Theta}^z_o, \mathbf{\Theta}^z_1\right) = \frac{\mathbb{P}\left(\{x_{ij}, t_{ij}\}_{i,j} \,|\mathbf{\Theta}^z_1\right) \cdot \mathbb{P}(V = 1|Z = z)}{\sum_{v \in \{0,1\}} \mathbb{P}\left(\{x_{ij}, t_{ij}\}_{i,j} \,|\mathbf{\Theta}^z_v\right) \cdot \mathbb{P}(V = v|Z = z)}, \tag{15}$$

where we have assumed in (14) and (15) that the epoch index $\bar{k}$ is observed and we dropped the conditioning on $\bar{k}$ for simplicity of exposition. In the next Section, we develop an algorithm that learns the patients' physiological model from offline data, and computes the monitored patients' personalized risk scores using (14) and (15).

III. A Personalized Risk Scoring Algorithm

In this Section, we propose an algorithm that learns the physiological model presented in the previous Section from offline data, and computes the risk score formulated in (14) and (15) for newly hospitalized patients in real-time.

A. Objectives

Given an offline training dataset $\mathcal{D}$ that comprises N *reference patients* whose physiological measurements were recorded in the electronic health record (EHR), we aim at learning a personalized risk scoring model, i.e. learning the parameters of the model presented in Section II, and applying the learned risk model for newly hospitalized patients.

The training dataset $\mathcal{D}$ is represented as a collection of tuples

$$\mathcal{D} = \left\{ \left(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}, y^{(n)}, v^{(n)} \right) \right\}_{n=1}^{N},$$

where each element in $\mathcal{D}$ corresponds to a reference patient; $\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}$ is the set of vital signs and lab tests measurements, $y^{(n)}$ is the admission information, and $v^{(n)}$ is the true clinical status (i.e. patient is admitted to the ICU or discharged home) of the n^{th} patient in $\mathcal{D}$. For $v \in \{0,1\}$, let

$$\mathcal{D}_v = \left\{ \left(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}, y^{(n)}, v^{(n)} \right) : v^{(n)} = v \right\},$$

where $\mathcal{D}_o$ is the set of data points for clinically stable patients, and $\mathcal{D}_1$ is the set of data points for clinically deteriorating patients, and $N_v = |\mathcal{D}_v|$ is the size of the dataset $\mathcal{D}_v$.

Our algorithm $\mathcal{A}$ operates in two modes: an *offline mode* $\mathcal{A}_{off}$, in which a risk scoring model is learned from the offline dataset $\mathcal{D}$, and an *online mode* $\mathcal{A}_{on}$, in which a risk score is sequentially computed for a newly hospitalized patient with a sequence of physiological measurements $\{x_{ij}, t_{ij}\}_{i,j}$, i.e.

$$(\hat{\mathbf{\Theta}}_o^1, \ldots, \hat{\mathbf{\Theta}}_o^G, \hat{\mathbf{\Theta}}_1^1, \ldots, \hat{\mathbf{\Theta}}_1^G) = \mathcal{A}_{off}(\mathcal{D}),$$

$$R(t,y) = \mathcal{A}_{on}(\{x_{ij}, t_{ij} \leq t\}_{i,j}, \hat{\mathbf{\Theta}}_o^1, \ldots, \hat{\mathbf{\Theta}}_o^G, \hat{\mathbf{\Theta}}_1^1, \ldots, \hat{\mathbf{\Theta}}_1^G).$$

That is, $\mathcal{A}_{off}$ estimates the parameter set for stable and deteriorating patients for all subtypes $(\hat{\mathbf{\Theta}}_o^1, \ldots, \hat{\mathbf{\Theta}}_o^G, \hat{\mathbf{\Theta}}_1^1, \ldots, \hat{\mathbf{\Theta}}_1^G)$, whereas $\mathcal{A}_{on}$ implements (14) and (15) to assign a risk score for the monitored patient in real-time.

In order to evaluate the predictive power of the algorithm $\mathcal{A}$, we set a threshold η on the computed risk score $R(t,y)$, and allow the algorithm to prompt an alarm (i.e. declare the hypothesis $\mathcal{H}_1$) whenever the risk score crosses that threshold. This resembles the structure of the optimal sequential hypothesis test, where the null hypothesis is rejected whenever the test statistic crosses a predefined threshold [19]. We define T_s as the *stopping time* at which the risk score computed by the algorithm $\mathcal{A}$ crosses the threshold η, i.e.

$$T_s(\eta) = \inf\{t \in \mathbb{R}_+ : R(t,y) \geq \eta\}.$$

The performance of the algorithm $\mathcal{A}$ is evaluated in terms of the positive predictive value (PPV), and the true positive rate (TPR) defined as follows

$$\text{PPV} = \frac{\mathbb{P}(T_s(\eta) \leq T_{end}|\mathcal{H}_1)}{\mathbb{P}(T_s(\eta) \leq T_{end}|\mathcal{H}_o) + \mathbb{P}(T_s(\eta) \leq T_{end}|\mathcal{H}_1)}, \quad (16)$$

and

$$\text{TPR} = \frac{\mathbb{P}(T_s(\eta) \leq T_{end}|\mathcal{H}_1)}{\mathbb{P}(T_s(\eta) \leq T_{end}|\mathcal{H}_1) + \mathbb{P}(T_s(\eta) > T_{end}|\mathcal{H}_1)}, \quad (17)$$

where T_{end} is the time at which observations of the patient's monitored physiological stream stops either because of an ICU admission or discharge (i.e. for a clinically deteriorating patient $T_{end} = (K - \bar{k} + 1) \cdot T_1$).

B. Algorithm

In this section, we propose an implementation for the algorithm $\mathcal{A}_{off}$ that learns the parameters of the physiological model presented in Section II from a dataset $\mathcal{D}$, and an implementation for the algorithm $\mathcal{A}_{on}$ which infers the clinical status and computes the risk score for a newly hospitalized patient according to (14) and (15). The implementation of the algorithms $\mathcal{A}_{off}$ and $\mathcal{A}_{on}$ is confronted with the following challenges:

1) The number of patient subtypes G is unknown, and the subtype memberships of the reference patients is not declared in $\mathcal{D}$.
2) The relationship between the admission information Y and the latent subtype Z is unknown and needs to be learned from the data.
3) The physiological model for the clinically deteriorating patients is non-stationary, and hence, for newly admitted patients, we need to estimate the latent epoch index $\bar{k}$ in real-time in order to synchronize the patient's physiological signal with our model, and properly compute the patient's risk score described by (14) and (15).
4) The physiological model for the clinically deteriorating patients has many parameters (i.e. $K\left(\frac{D(D+1)}{2} + D + 1\right)$ parameters), but the number of clinically deteriorating patients in the dataset $\mathcal{D}$ is relatively small (ICU admission rate is usually less than 10%).

In the following, we provide an implementation for the offline algorithm $\mathcal{A}_{off}$ that addresses challenges (1-3), and then we present an implementation for the online algorithm $\mathcal{A}_{on}$ that addresses challenge (4).

The offline algorithm $\mathcal{A}_{off}$

The objective of the offline algorithm $\mathcal{A}_{off}$ is to learn from $\mathcal{D}$ the number of subtypes G, the parameter set $(\mathbf{\Theta}_o^1, \ldots, \mathbf{\Theta}_o^G, \mathbf{\Theta}_1^1, \ldots, \mathbf{\Theta}_1^G)$, and the probability of a patient's membership in each subtype given her admission information, i.e. $\mathbb{P}(Z = z|Y = y)$. In the rest of this Section, we use the following notations $\Gamma_v = (\mathbf{\Theta}_v^1, \ldots, \mathbf{\Theta}_v^G), v \in \{0,1\}$, and $\beta_z(y) = \mathbb{P}(Z = z|Y = y)$..

Recall from (14) that the risk score $R(t,y)$ can be written as

$$R(t,y) = \sum_{z \in \mathcal{Z}} R_z(t) \cdot \beta_z(y), \quad (18)$$

where

$$R_z(t) = \mathbb{P}\left(V = 1 \,|\{x_{ij}, t_{ij}\}_{i,j}, \mathbf{\Theta}_o^z, \mathbf{\Theta}_1^z\right). \quad (19)$$

The formulation of the risk score $R(t,y)$ in (18) explicates the impact of the patient's latent subtype on her risk assessment. The score $R(t,y)$ is a weighted average of the posterior probabilities $R_z(t) = \mathbb{P}(V = 1 \,|\{x_{ij}, t_{ij}\}_{i,j}, \mathbf{\Theta}^z_o, \mathbf{\Theta}^z_1)$, i.e. the probabilities of the alternative hypothesis $\mathcal{H}_1$ given the evidential physiological data and the latent subtype being $Z = z$, over all possible latent subtypes for the patient. The weight $\beta_z(y)$ associated with the term $R_z(t)$ corresponds to the probability that the patient with admission information $Y = y$ belongs to subtype $Z = z$. We denote $R_z(t)$ as the "*expert for subtype z*", whereas the weight $\beta_z(y)$ is denoted as the "*responsibility of expert z*". Therefore, computing the risk score $R(t,y)$ entails invoking a *mixture* of GP experts, and assigning the mixture weights in accordance to the experts' responsibilities determined by $\beta_z(y)$.

The algorithm $\mathcal{A}_{off}$ operates in 3 steps. In step 1, we *discover the experts*, i.e. we apply the expectation-maximization (EM) algorithm to the dataset $\mathcal{D}_o$ in order to estimate the latent patient subtypes and the physiological model parameters for the clinically stable patients. We apply the Bayesian Information Criterion (BIC) for model selection in order to select the number of subtypes G. This ensures statistical efficiency in learning the number of subtypes and the model parameters since the physiological model for the clinically stable patients in $\mathcal{D}_o$ has only $\frac{D(D+1)}{2} + D + 1$ parameters. In step 2, we use a (transductive) *transfer learning approach* to learn the *experts' responsibilities* $\beta_z(y)$ as a function of the admission information. Finally, in step 3, we use a transfer learning approach to learn the parameters of the physiological model for the clinically deteriorating patients through the dataset $\mathcal{D}_1$ using the model learned for the clinically stable patients from the dataset $\mathcal{D}_o$. In the following, we specify the detailed steps of the algorithm $\mathcal{A}_{off}$.

Step 0. Align the temporal physiological streams in the dataset $\mathcal{D}_1$: Before implementing the 3 steps of the algorithm $\mathcal{A}_{off}$, we need to ensure that the recorded (non-stationary) physiological streams in $\mathcal{D}_1$are aligned with respect to a common reference time in order to properly estimate the GP parameters for every epoch $k \in \{1, 2, \ldots, K\}$. This is achieved by considering the ICU admission time as a surrogate marker for the latent epoch index k. That is, we consider that the samples in the last T_1 period of time in every physiological streams to be designated as epoch K (i.e. the last epoch), and then we go backwards in time and label the preceding epochs as $K-1, K-2$, etc. This procedure is applied to all the physiological streams of the reference patients in $\mathcal{D}_1$, and hence all the training physiological streams become aligned in time which allows for a straight-forward epoch-specific parameter estimation. The epoch length T_1, and the number of epochs K are hyper-parameters that are optimized via cross-validation. The distribution of the initial epoch index $f(\bar{k})$ is a *truncated negative binomial* distribution with support $\{1, \ldots, K\}$, and can be straightforwardly estimated given the patients' length of stay information.

Step 1. Discover the Experts through Clinically Stable Patients: In this step, we learn both the number of subtypes G (which is also the number of experts), as well as the parameter sets Γ_o. This is accomplished through an iterative approach in which we use the expectation-maximization (EM) algorithm for estimating the parameters in Γ_o for given values of G, and then use the Bayesian information criterion (BIC) to select the number of experts.

The detailed implementation of the EM algorithm is given in lines 4-18 in Algorithm 1. The algorithm is executed on the dataset $\mathcal{D}_o$ by iterating over the values of G, with an initial number of experts $G = 1$. For every M, we implement the usual E-step and M-step of the EM-algorithm: starting from an initial parametrization Γ_o, in the p^{th} iteration of the EM-algorithm, the auxiliary function $Q(\Gamma_o; \Gamma_o^{p-1})$ is computed as

$$Q(\Gamma_o; \Gamma_o^{p-1}) = \mathbb{E}[\log(\mathbb{P}(\mathcal{D}_o, \{Z^{(n)}\}_{n=1}^{N_o} \,|\, \Gamma_o)) \,|\, \mathcal{D}_o, \Gamma_o^{p-1}],$$

where $Z^{(n)}$ is the latent subtype of the n^{th} entry of the dataset $\mathcal{D}_o$. The parametrization is updated in the M-step by maximizing $Q(\Gamma_o; \Gamma_o^{p-1})$ with respect to Γ_o (closed-form expressions are available for the jointly Gaussian data in $\mathcal{D}_o$ as per the GP model). The p^{th} iteration is concluded by updating expert z's responsibility towards the n^{th} patient in the dataset $\mathcal{D}_o$ as follows

$$\begin{aligned}\beta^{(n)}_{z,p} &= \mathbb{P}(Z^{(n)} = z \,|\, \{x^{(n)}_{ij}, t^{(n)}_{ij}\}_{i,j}, \Gamma^p_o) \\ &= \frac{\pi^p_z \, f(\{x^{(n)}_{ij}, t^{(n)}_{ij}\}_{i,j} \,|\, \mathbf{\Theta}^{p,z}_o)}{\sum_{z'=1}^{G} \pi^p_{z'} \, f(\{x^{(n)}_{ij}, t^{(n)}_{ij}\}_{i,j} \,|\, \mathbf{\Theta}^{p,z'}_o)}, \end{aligned} \tag{20}$$

where π^p_z is the estimate for $\mathbb{P}(Z = z)$ in the p^{th} iteration, and $f(.)$ is the Gaussian distribution function. The term $\beta^{(n)}_{z,p}$ represents the posterior probability of patient n's membership in subtype z given the realization of her physiological data $\{x_{ij}, t_{ij}\}_{i,j}$. The iterations of the EM-algorithm stop when the claimed responsibilities of the G experts towards the N_o reference patients in $\mathcal{D}_o$ converges to within a precision parameter ϵ (line 14).

After each instantiation of the EM-algorithm, we compare the model with G experts to the previous model with $G-1$ experts found in the previous iteration. Comparison is done through the Bayes factor $B_{G,G-1}$ (computed in line 16 via the BIC approximation), which is simply a ratio between Bayesian criteria that trade-off the likelihood of the model being correct with the model complexity (penalty for a model with G experts is given by Ψ_G in line 15, such a penalty corresponds to the total number of hyper-parameters in the model with G experts). We stop adding new experts when the Bayes factor $B_{G,G-1}$ drops below a predefined threshold $\bar{B}$.

Step 2. Recruit the Experts via (transductive) Transfer Learning[2]: Having discovered the experts by learning the parameter set $\Gamma_o = (\mathbf{\Theta}^1_o, \ldots, \mathbf{\Theta}^G_o)$, we need to learn how to associate different experts to the patients based on the initial information we have about them, i.e. the admission

[2] Our terminologies with respect to transfer learning paradigms follow those in [39].

Algorithm 1 The Offline Algorithm $\mathcal{A}_{off}$

```
1: Input: Dataset $\mathcal{D}$, precision, $\epsilon$, threshold $\bar{B}$.
2: Implement step 1 (Discover the experts):
3: Extract dataset $\mathcal{D}_o$ of clinically stable patients with label $v^{(n)} = 0$.
4: Initialize $G = 1$
5: repeat
6:     $p \leftarrow 1$
7:     Initialize $\Gamma_o^p = \{\Theta_o^{p,z}\}_{z=1}^{G}$.
8:     repeat
9:         E-step: Compute $Q(\Gamma_o; \Gamma_o^{p-1})$.
10:        M-step: $(\boldsymbol{\Theta}_o^p, \{\pi_z^p\}_{z=1}^G) = \arg\max_{\Gamma_o} Q(\Gamma_o; \Gamma_o^{p-1})$.
11:        $Q_G^* \leftarrow \max_{\Gamma_o} Q(\Gamma_o; \Gamma_o^{p-1})$.
12:        Update responsibilities using Bayes rule $\beta_{z,p}^{(n)} = \frac{\pi_z^p\, f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} | \boldsymbol{\Theta}_o^{p,z})}{\sum_{z'=1}^{G} \pi_{z'}^p\, f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} | \boldsymbol{\Theta}_o^{p,z'})}$
13:        $p \leftarrow p + 1$.
14:    until $\frac{1}{N_o G} \sum_{i=1}^{N_o} \sum_{z=1}^{G} \left| \beta_{z,p}^{(n)} - \beta_{z,p-1}^{(n)} \right| < \epsilon$
15:    $\Psi_G = G\left(\frac{D(D+1)}{2} + D + 1\right)$
16:    $B_{G,G-1} \approx \frac{\exp\left(Q_G^* - \frac{1}{2}\Psi_G \log(N_o)\right)}{\exp\left(Q_{G-1}^* - \frac{1}{2}\Psi_{G-1} \log(N_o)\right)}$
17:    $G \leftarrow G + 1$.
18: until $B_{G,G-1} < \bar{B}$
19: Implement step 2 (Recruit the experts):
20: Construct the dataset $\left\{y^{(n)}, (\beta_1^{(n)}, \ldots, \beta_G^{(n)})\right\}_{n=1}^{N_o}$.
21: Find linear regression coefficients for $\beta_z(y) = [w_1^z, \ldots, w_S^z]^T y$.
22: Implement step 3 (Transfer learning):
23: For every $n \in \mathcal{D}_1$ and $z \in \{1, \ldots, G\}$, sample a random variable $c_{n,z} \sim \text{Bernoulli}(\beta_z(y^{(n)}))$.
24: For every expert $z$, construct a dataset $\mathcal{D}_{1,z} = \{n \in \mathcal{D}_1 : c_{n,z} = 1\}$.
25: Find the MLE estimates of $\Gamma_1$ using the samples in the corresponding datasets $\{\mathcal{D}_{1,1}, \ldots, \mathcal{D}_{1,G}\}$.
```

Vital signs	Lab tests	Admission
Diastolic blood pressure	Glucose	Transfer Status
Eye opening	Urea Nitrogen	Gender
Glasgow coma scale score	White blood cell count	Age
Heart rate	Creatinine	Transplant
Respiratory rate	Hemoglobin	Floor ID
Temperature	Platelet Count	ICD-9 codes
O_2 Device Assistance	Potassium	Race
O_2 Saturation	Sodium	Ethnicity
Best motor response	Total CO_2	
Best verbal response	Chloride	
Systolic blood pressure		

TABLE I: Physiological data and admission information associated with the patient cohort under study.

features (e.g. transfer status, age, gender, ethnicity, etc). In other words, we aim to learn a mapping rule $\beta_z(y) : \mathcal{Y} \rightarrow \mathcal{Z}$. The function $\beta_z(y)$ reflects the extent to which we rely on the different experts when scoring the risk of a patient with admission information $Y = y$.

A transfer learning approach is used to learn the function $\beta_z(y)$. That is, we use the estimates for the posterior $\beta_z^{(n)}$ obtained from step 1 (see line 12 in Algorithm 1) for every patient n in $\mathcal{D}_o$, and then we label the dataset $\mathcal{D}_o$ with these posteriors, and transfer these labels to the domain of admission features, thereby constructing a dataset of the form $\left\{y^{(n)}, (\beta_1^{(n)}, \ldots, \beta_G^{(n)})\right\}_{n=1}^{N_o}$. We use a linear regression analysis to fit the function $\beta_z^{(n)}$ (see lines 20-21 in Algorithm 1).

Step 3. Discover the Experts of Clinically Deteriorating Patients: The knowledge of the parameter set $\Gamma_1 = (\boldsymbol{\Theta}_1^1, \ldots, \boldsymbol{\Theta}_1^G)$ needs to be gained from the dataset $\mathcal{D}_1$. We use a self-taught transfer learning approach to transfer the knowledge obtained using unsupervised learning from the dataset $\mathcal{D}_o$, i.e. the domain of stable patients, to "label" the dataset $\mathcal{D}_1$ and learn the set of experts associated with the clinically acute patients [39], [40].

Self-taught learning is implemented by exporting the number of experts G that we estimated from $\mathcal{D}_o$ directly to the population of patients in $\mathcal{D}_1$, picking a subset of patients in $\mathcal{D}_1$ to estimate the parameter set $\boldsymbol{\Theta}_1^z$ of expert z by sampling patients from $\mathcal{D}_1$ using their responsibility vectors (line 23 in Algorithm 1).

The online algorithm $\mathcal{A}_{on}$

An aggregate risk score for every patient with admission information $Y = y$ is obtained by weighting the opinions of the G experts with their responsibilities $\{\beta_z(y)\}_{z=1}^G$. The risk score for a newly hospitalized patient i with admission information $Y = y$ at time t is then given by

$$R(t, y) = \sum_{z=1}^{G} \beta_z(y)\, R_z(t).$$

Note that computing $R_z(t)$ is not possible unless we know the latent epoch index $\bar{k}$ for the monitored patient. Since $\bar{k}$ is a hidden variable, we estimate $\bar{k}$ and evaluate $R_z(t)$ by averaging over its posterior distribution, i.e.

$$\begin{aligned} R_z(t) &= \mathbb{E}_{\bar{k}}\left[\mathbb{P}(V = 1 | \{x_{ij}, t_{ij} \leq t\}_{ij}, \bar{k}, \Gamma_o, \Gamma_1)\right] \\ &= \sum_{1 \leq \bar{k} \leq K} \mathbb{P}(V = 1 | \{x_{ij}, t_{ij} \leq t\}_{ij}, \bar{k}, \Gamma_o, \Gamma_1) \times \\ &\quad \mathbb{P}(\bar{k} | \{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1), \end{aligned} \tag{21}$$

where $\mathbb{P}(V = 1 | \{x_{ij}, t_{ij} \leq t\}_{ij}, \bar{k}, \Gamma_o, \Gamma_1)$ is evaluated via Bayes rule as clarified in (15). Hence, the online algorithm $\mathcal{A}_{on}$ continuously estimates the latent epoch index $\bar{k}$ as more physiological data is gathered, and synchronized the monitored physiological stream with the learned (non-stationary) GP model. Algorithm 2 shows the a pseudo-code for the operations implemented in the real-time stage. We note that if the current patient's length of stay exceeded $K\, T_1$, we use the deterioration model for the K^{th} epoch throughout her remaining time in the ward.

Algorithm 2 The Online Algorithm $\mathcal{A}_{on}$

```
Input: Physiological measurements $\{x_{ij}, t_{ij}\}_{i,j}$, admission features $Y = y$, a set of experts' parameters $\Gamma_o$ and $\Gamma_1$.
Estimate the experts' responsibilities $\beta_z(y)$.
Compute the posterior epoch index distribution $\mathbb{P}(\bar{k}|\{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1)$.
For every expert $z$, compute the risk score
    $R_z(t) = \sum_{1 \leq \bar{k} \leq K} \mathbb{P}(V = 1|\{x_{ij}, t_{ij} \leq t\}_{ij}, \bar{k}, \Gamma_o, \Gamma_1) \times \mathbb{P}(\bar{k}|\{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1).$
Compute the final risk score as a mixture of the individual experts' risk assessments weighted by their individual responsibilities toward the monitored patient
    $R(t, y) = \sum_{z=1}^{G} \beta_z(y)\, R_z(t).$
```

IV. Experiments and Results

A. Data Description

Experiments were conducted on a cohort of 6,321 patients who were hospitalized in a general medicine floor in the Ronald Reagan UCLA medical center during the period between March 3^{rd} 2013, to February 4^{rd} 2016 (excluding patients who were initially admitted to the ICU and then transferred to the ward after stabilization since for those patients the data were not recorded in the EHR). The patients' population is heterogeneous with a wide variety of diagnoses and ICD-9 codes: the patient's cohort included an overall number of 1,643 ICD-9 codes; the most frequent of which corresponded to conditions such as shortness of breath, hypertension, septicemia, sepsis, fever, pneumonia and renal failure. The cohort included patients who were not on immunosuppression and others who were on immunosuppression, including patients that have received solid organ transplantation. In addition, there were some patients that had diagnoses of leukemia or lymphoma. Some of these patients received stem cell transplantation as part of their treatment. Because these patients receive chemotherapy to significantly ablate their immune system prior to stem cell transplantation, they are at an increased risk of clinical deterioration. The vast heterogeneity of the patients' cohort motivates the need for a "personalized" risk model, and suggests the general applicability of the experimental results presented in this Section.

Patients in the dataset $\mathcal{D}$ were monitored for 11 vital signs (e.g. O_2 saturation, heart rate, systolic blood pressure, etc) and 10 lab tests (e.g. Glucose, white blood cell count, etc). Hence, the dimension of the physiological stream for every patient is $D = 21$. Table I lists all the vital signs and lab tests included in the experiment, in addition to the set of admission information Y that are used for personalizing the computed risk scores. The ICD-9 code ranges were converted to a set of 18 categorical values, where each value bundles a set of ICD-9 codes for "related diseases"; such a "categorization" allows the algorithm to handle newly hospitalized patients with rare ICD-9 codes that were not present in the dataset. The sampling rate for the physiological streams $\{x_{ij}, t_{ij}\}_{i,j}$ ranges from 1 hour to 4 hours, and the length of hospital stay for the patients ranged from 2 to 2,762 hours. Correlated feature selection (CFS) was used to select the physiological streams that are relevant to predicting the endpoint outcomes (i.e. ICU admission) [41]; the CFS algorithm selected 7 vital signs (Diastolic blood pressure, eye opening, Glasgow coma scale score, heart rate, temperature, O_2 device assistance and O_2 saturation), and 3 lab tests (Glucose, Urea Nitrogen and white blood cell count).

Throughout the experiments conducted in this Section, the training and testing datasets are constructed as follows. The training set comprises 5,130 patients who were admitted to the ward in the period between March 2013 and July 2015. Among those patients, the ICU admission rate was 8.34%. The algorithms are trained via this dataset, and then tested on a separate dataset that comprises the remaining 1,191 patients who were admitted to the ward in the period between July 2015 and April 2016 (ICU admission rate is 8.13%). The training set is split into a set of 4,130 patients for training, and 1,000 patients for validation. The validation set is used for feature selection and tuning T_1 and K.

B. Subtype Discovery

When running the risk scoring algorithm on the 5,130 patients in the testing set, the algorithm was able to discover 6 patient subtypes ($G = 6$), and train the corresponding GP experts. Setting the number of subtypes as $G = 6$ experts is optimal given the size of the dataset $\mathcal{D}$; the offline algorithm $\mathcal{A}_{off}$ stops after computing the Bayes factor $B_{6,5}$.

Having discovered the latent patient subtypes, we investigate how the hospital admission features Y are associated to the patients' subtypes, i.e. we are interested in understanding *which* of the admission features are most representative of the latent patient subtypes. Table II lists the admission features ranked by their "importance" in deciding the responsibilities of the 6 experts corresponding to the 6 subtypes. Since we normalize all feature to the range [0,1], the importance, or relevance, of an admission feature can be quantified by the weight of that feature $(w_1, \ldots, w_S)$ in the learned linear regression function $\beta_z(y)$ averaged over all subtypes (see line 21 in Algorithm 1). As shown in Table II, stem cell transplant turned out to be the feature that is most relevant to the assignment of responsibilities among experts. This is consistent with domain knowledge: patients receiving stem cell transplantation are at a higher risk of deterioration due to their severely compromised immune systems, thus it is extremely important to understand their physiological state [42].

Surprisingly, gender turned out to be the third most relevant feature for expert assignments. This means that vital signs and lab tests for males and females should not be interpreted in the same way when scoring the risk of clinical deterioration, i.e. different GP experts needs to handle different genders (recall the demonstration in Fig. 1). The fact that the transfer status of a patient is an important admission factor (ranked fourth in

Rank	Admission feature	Regression coefficient
1	Stem cell transplant	0.1091
2	Floor ID	0.0962
3	Gender	0.0828
4	Transfer status	0.0827
5	ICD-9 code	0.0358
6	Age	0.0109

TABLE II: Relevance of the patients' admission features to the latent subtype memberships.

the list) is consistent with prior studies that demonstrate that patients transferred from outside facilities have a higher acuity with increased mortality [43].

C. Prognosis and Early Warning Performance

We validated the utility of the proposed risk scoring model by constructing an EWS that issues alarms for ICU admission based on the real-time risk score (i.e. ICU alarms are issued whenever the risk score $R(t, y)$ crosses a threshold η), and evaluating the performance of the EWS in terms of the PPV and the TPR as defined in (16) and (17). The accuracy of the proposed risk model is compared with that of the state-of-the-art risk scores (Rothman, MEWS, APACHE and SOFA) by evaluating the Receiver Operating Characteristics (ROC) curves in Fig 2a. The implementation of the MEWS and Rothman indexes followed their standard methodologies in [44] and [5], whereas the implementations of SOFA and APACHE followed [45].

As shown in Fig. 2a, the proposed risk model with $G = 6$ subtypes consistently outperforms all the other risk scores for any setting of the TPR and PPV. The proposed score offers gains of 12% with respect to the (most competitive) Rothman score (p-value $<$ 0.01). This promising result shows the prognostic value of replacing the currently deployed scores in wards with scores that captures the patients' heterogeneity, considers the temporal aspects of the physiological data, and accounts for the correlations among different physiological streams. The same comparison is carried out in Fig. 2b, but in terms of the TPR and the false positive rate (FPR) performances, and it can be seen that the AUC of the proposed score (0.806) outperforms that of the Rothman index (0.72) and all other risk scoring methods. Moreover, as shown in Fig. 2c, the proposed risk score also outperforms state-of-the-art machine learning techniques (logistic regression, linear regression, random forest, and LASSO); it provides an AUC gain of around 10% with respect to these techniques (p-value $<$ 0.01).

It is important to note that the proposed risk score significantly reduces the false alarm rates as compared to the state-of-the-art risk scores. This can be seen for the numerical values in Table III and is also reflected in the TPR/PPV performance comparison in Fig. 2a, where we can see that for any fixed TPR, the proposed risk score achieves a much higher PPV than the Rothman index, e.g. at a TPR of 60%, the proposed score achieves a PPV of 30%, which is double of that achieved by the Rothman index (15%). This significant reduction in the false alarm rate can be attributed to the fact that the proposed algorithm computes a risk score based on a trajectory of measurements rather than instantaneous ones. Fig. 3c illustrates this effect by depicting a realization for the risk scores' trajectory of a clinically stable patient in the testing dataset. We can see that the MEWS and Rothman indexes exhibit drastic fluctuations over time as they only consider the most recent vital signs and lab tests, which makes them easily triggered by instantaneous measurements or transient phenomena. Our score offers a smoother trajectory that is more resilient to false alarms since it computes a posterior probability that is conditioned on the entire physiological history.

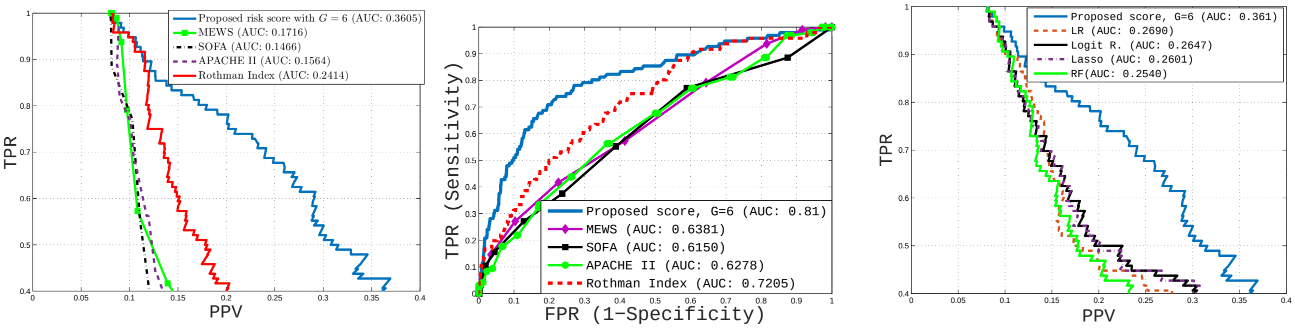
Reductions in the false alarm rates are further demonstrated in Table III, where we specify the number of false alarms per one true alarm for both the proposed risk score and the state-of-art scores at different settings of the TPR. At a TPR of 50%, our risk score leads to only 2.16 false alarms for every 1 true alarm, whereas the Rothman index lead to 4.56 false alarms per true alarm, i.e. the rate of the false alarms caused by the Rothman index is more than double of that caused by the proposed algorithm. Thus, our risk score can ensure more confidence in its issued ICU alarms, which would mitigate alarm fatigue and enhance a hospital's resource utilization [26], [46]. Table III shows that our risk score offers a lower false alarm rate compared to all other scores and benchmark algorithms for all settings of the TPR.

Fig. 3a illustrates the trade-off between the timeliness of the ICU alarm and its accuracy for a fixed TPR of 50% (the achieved gains hold for any setting of the TPR). In Fig. 3a, we select an alarm threshold η that corresponds to a fixed TPR of 50%, and then compute the PPV for the alarms issued at different time horizons prior to ICU admission. We can see that the proposed risk score consistently outperforms all the other scores in terms of the timeliness of its ICU alarms for all the PPV settings. For instance, for a PPV greater than 25%, our score offers a 12-hour earlier predictions with respect to the actual physician-determined ICU admission event. This level of timeliness is not feasible for any of the other risk scores. Combining the results shown in Fig. 3a and Table III, one can see that the proposed risk score is able to both warn the clinician earlier and provide a more confident signal as compared to the state-of-the-art risk scores, thus providing the ward staff with a safety net for patient care by giving them sufficient time to intervene in order to prevent clinical deterioration.

The value of personalization is depicted in Fig. 3b and Fig. 3c, where we plot the ROC and timeliness curves for our algorithm once with one subtype (i.e. $G = 1$ and no personalization is taken into account), and once with $G = 6$ subtypes. If we were to take $G = 1$, our model would prompt ICU alarms that warns the clinicians 5 hours earlier than the physicians' determination. When we take $G = 6$, our model prompts ICU

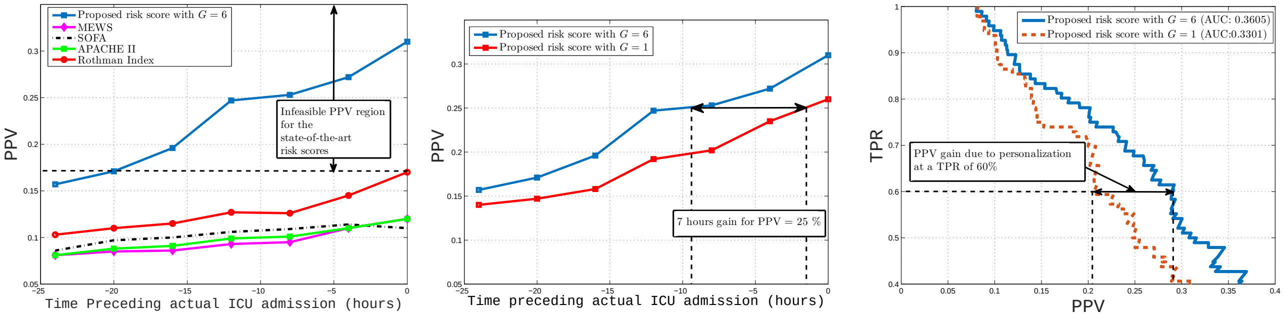
TPR	Proposed score ($G = 6$)	LR*	Logit. R.*	LASSO	RF*	MEWS	SOFA	APACHE	Rothman
40%	1.76	2.58	2.3	2.3	3.31	5.9	7.26	6.41	3.98
50%	2.16	4.46	3.95	3.44	4.62	7.13	7.77	7.13	4.56
60%	2.44	5.13	4.99	4.95	5.45	7.06	7.06	7.77	5.62
70%	3.15	6.09	6.25	6.09	6.41	8.8	8.52	8.62	6.35
80%	4.81	6.63	7.2	7.2	6.94	9.31	9.31	9.75	7.33

TABLE III: Number of false alarms per one true alarm (* LR = Linear regression, Logit. R. = Logistic regression, and RF = Random forest).



(a) (b) (c)

Fig. 2: (a) TPR and PPV performance comparisons (ROC curve) with respect to state-of-the-art risk scores. (b) TPR and FPR performance comparisons (ROC curve) with respect to state-of-the-art risk scores. (c) TPR and PPV performance comparisons (ROC curve) with respect to state-of-the-art machine learning techniques.



(a) (b) (c)

Fig. 3: (a) Timeliness of the proposed risk score. (b) Impact of personalization on the timeliness of the ICU alarms. (c) Impact of personalization on the ROC curve.

alarms 12 hours earlier than physician determination. Thus, even the unpersonalized version of our model is significantly quicker than the physician determination, but is sluggish in comparison to the personalized one. A similar gain is attained due to personalization in terms of the PPV. As shown in Fig. 3c, personalization leads to a 10% higher PPV at a TPR of 60% as compared to a non-personalized version of our model.

V. Conclusion

In this paper, we have developed a personalized risk scoring algorithm for critically ill patients in wards that allows transferring deteriorating patients to the ICU in a timely manner. The algorithm learns a granular risk scoring model that is tailored to the individual patient's traits by modeling the patient's physiological processes via a mixture of multitask Gaussian Processes, the weights of which are determined by the patient's baseline admission information and the latent subtypes discovered from the training data. We have demonstrated the utility of the proposed risk scoring algorithm through a set of experiments conducted on a heterogeneous cohort of 6,321 critically ill patients who were recently admitted to Ronald Reagan UCLA medical center. The experiments have shown that the proposed risk score significantly outperforms the currently deployed risk scores, such as the Rothman index, MEWS, APACHE and SOFA scores, in terms of timeliness, true positive rate, and positive predictive value. The results suggest the possibility of reducing the annual sub-acute care mortality rates through precision medicine.

Appendix A: Literature Review

Previous Works in the Medical Literature

Hospitals have been investigating and investing in prognostic risk scoring systems that quantify and anticipate the acuity of critically ill inpatients in real-time based on their (temporally evolving) physiological signals in order to ensure timely ICU transfer [3]–[9]. Prognosis in hospital wards is feasible since unanticipated adverse events are often preceded by disorders in a patient's physiological parameters [11], [12]. However, the subtlety of evidence for clinical deterioration in the physiological parameters makes the problem of constructing an "informative" risk score quite challenging: overestimating a patient's risk can lead to alarm fatigue and inefficient utilization of clinical resources [44], whereas underestimating her risk can undermine the effectiveness of consequent therapeutic interventions [13], [47].

Recent systematic reviews have shown that currently deployed expert-based risk scores, such as the MEWS score [24], provide only modest contributions to clinical outcomes [25]–[27]. Alternatives for expert-based risk scores can be constructed by training a risk scoring model using the data available in the electronic health records (EHR) [4]. Recently, a data-driven risk score, named the Rothman index, has been developed using regression analysis [5], and was shown to outperform the MEWS score and its variants [9]. However, this score lacks a principled model for the hospitalized patient's physiological parameters, and is mainly constructed using a "one-size-fits-all" approach that leaves no room for personalized risk assessment that is tailored to the individual patient. Personalized models that account for the patient's individual traits are anticipated to provide significant accuracy and granularity in risk assessments [48].

Two broad categories of risk models and scores that quantify a patient's risk for an adverse event have been developed in the medical literature. The first category comprises *early-warning scores* (EWS), which hinge on expert-based models for triggering transfer to ICU [24]. Notable examples of such scores are MEWS and its variant VitalPAC [7]. These scores rely mainly on experts to specify the risk factors and the risk scores associated with these factors [44]. A major drawback of this class of scores is that since the model construction is largely relying on experts, the implied risk functions that map physiological parameters to risk scores do not have any rigorous validation. Recent systematic reviews have shown that EWS-based alarm systems only marginally improve patient outcomes while substantially increasing clinician and nursing workloads [25]–[27]. Other expert-based prognostication scores that were constructed to predict mortality in the ICU, such as SOFA and APACHE scores, has been shown to provide a reasonable predictive power when applied to predict deterioration for patients in wards [45].

The second category of risk scores relies on more rigorous, data-intensive regression models to derive and validate risk scoring functions using the electronic medical record. Examples for such risk scores include the regression-based risk models developed by Kirkland et al. [4], and by Escobar et al. [49]. Rothman et al. build a more comprehensive model for computing risk scores on a continuous basis in order to detect a declining trend in time [5], [9]. The risk score computed therein, which is termed as the "Rothman index", quantifies the individual patient condition using 26 clinical variables (vital signs, lab results, cardiac rhythms and nursing assessments). Table I summarizes the state-of-the-art risk scores used for critical care prognostication.

The Rothman index is the state-of-the-art risk scoring technology for patients in wards: about 70 hospitals and health-care facilities, including Houston Methodist hospital in Texas, and Yale-New Haven hospital in Connecticut, are currently deploying this technology [50]. While validation of the Rothman index have shown its superiority to MEWS-based models in terms of false alarm rates [9], the risk scoring scheme used for computing the Rothman index adopts various simplifying assumptions. For instance, the risk score computed for the patient at every point of time relies on instantaneous measurements, and ignores the history of previous vital sign measurements (see Equation (1) in [5]). Moreover, correlations among vital signs are ignored, which leads to double counting of risk factors. Finally, the Rothman scoring model is fitted to provide a reasonable "average" predictive power for the whole population of patients, but does not offer "personalized" risk assessments for individual patients, i.e. it ignores baseline and demographic information available about the patient at admission time. Our risk scoring model addresses all these limitations, and hence provides a significant gain in the predictive power as compared to the Rothman index as we show in Section IV.

Previous Works in the Machine Learning Literature

The problem of modeling multivariate physiological time series has been recently investigated by the machine learning community [2], [6], [10], [21], [22], [28]–[30], [59], [60]; some of the previous works have also adopted multitask GP models [6], [10], [21], [22], [28]. However, most of these works have focused on a *forecasting* problem in which the goal is to predict the future values of an observable bio-marker. For instance, [28] focuses on predicting the PFVC clinical marker (a measure of lung severity) for scleroderma patients, [6], [10], [21], [22] focus on predicting the future values of SOFA, APACHE and SAPS scores for ICU patients, and [30] focuses on predicting the GFR bio-marker values for patients with chronic kidney disease. Unfortunately, a major challenge encountered in our setting is that patients in regular wards have no such strongly indicative bio-markers; we face this challenge by resorting to a *latent class* modeling approach, in which different classes correspond to different severity states. Our model adopts two latent classes, which allows the risk scoring problem to be formulated as a *sequential hypothesis test* [19]. Consequently, our multitask GP model serves as a tool for computing the optimal test statistic, and not for performing GP regression

Reference	Risk scores	Details	Limitations
[2], [7], [24], [44], [51]–[53]	MEWS, ViEWS and TREWS	Expert-based risk assessment methodologies (also known as "track and trigger" systems)	• Neither personalized nor data-driven, does not take advantage of the EHR. • Modest performance reported by recent systematic reviews in [25]–[27].
[45], [54]–[56]	SOFA	A combination of organ dysfunction scores for respiratory, coagulation, liver, cardiovascular and renal systems. Originally developed for predicting mortality in ICU patients, but was shown in [45] to function as a prognostication tool for non-ICU ward patients.	• Not personalized, i.e. uses the same scoring scheme for all patients (see Table 3. in [54]). • Does not consider correlations between organ dysfunction scores and endpoint outcomes. • Predictions can corporate the mean statistics of the computed score over time but does not consider the full temporal trajectory.
[45], [57], [58]	APACHE II and III	A disease severity score used for ICU patients (usually applied within 24 hours of admission of a patient to the ICU [57]). It has been shown in [45] that it can be used for prognostication in regular wards.	• Does not consider the temporal trajectory of score evaluations during the patients stay in ICU (or in the ward).
[5], [9]	Rothman index	A regression-based data-driven model that utilizes physiological data to predict mortality, 30-days readmission, and ICU admissions.	• Not personalized. Uses vital signs and lab tests to construct a "one-size-fits" all population-level model. • Ignores correlations between vital signs, and hence may double-count risk factors (see Eq. (1) in [5]). • Uses the instantaneous vital signs and lab tests measurements, and ignores the physiological stream trajectory.

TABLE IV: Summary of the state-of-the-art critical care risk scores.

as it is the case in the forecasting problems in [6], [10], [21], [22], [28]. To the best of our knowledge, our model is the first to conceptualize real-time risk scoring as a sequential testing procedure.

Our risk scoring model handles the heterogeneity of the patients' population via *subtyping*. Unlike previous works on subtyping in longitudinal disease progression models [28], [30], in which one set of subtypes is learned for the entire population of "sick" patients, the nature of the critical care setting (manifesting in our sequential testing framework) entails the need for learning different sets of subtypes for both clinical stability and deterioration. This imposes the challenge of learning a separate set of subtypes for the clinically deteriorating patients under class imbalance (ICU admission rate is less than 10%); we face this challenge via a novel learning algorithm that uses ideas from transfer learning to transfer the knowledge learned from the clinical stable population to the deteriorating population.

Most of the previous works on clinical risk prognosis used clinical endpoints (ICU admission or discharge) as "*surrogate labels*" for a patient's clinical deterioration, and hence used those labels to train a supervised (regression) model using the physiological data in a fixed-size time window before censoring. The supervised models used in the literature included logistic regression [31], [32] and SVMs [33]. We compare the performance of our model with these methods in Section IV. A detailed, tabulated comparisons with other risk scoring methodologies is provided in Appendix A in the supporting document.

Various other important tools for risk prognosis that do not rely on GP models have been recently developed. In [2] and [29], a Cox regression-based model was used to develop a sepsis shock severity score that can handle data streams that are censored due to interventions. However, this approach does not account for personalization in its severity assessments, and relies heavily on the existence of ordered pairs of comparisons for the extent of disease severity at different times, which may not always be available and cannot be practically obtained from experts. Our model does not suffer from such limitations: it does not rely on proportional hazard estimates, and hence does not require ordered pairs of disease severity temporal comparisons, and can be trained using the raw physiological stream records that are normally fed into the EHR during the patients' stay in the ward.

In [61] and [62], personalized risk factors are computed for a new patient by constructing a dataset of K "similar patients" in the training data, and train a predictive model for that patient. This approach would be computationally very expensive when applied in real-time for patients in a ward since it requires re-training a model for every new patient, and more importantly, it does not recognize the extent of heterogeneity of the patients, i.e. the constructed dataset has a fixed size of K irrespective of the underlying patients' physiological heterogeneity. Hence, such methods may incur efficiency loss if K is underestimated, and may perform unnecessary computations if the underlying population is already homogeneous. Our model overcomes this problem by learning the number of latent subtypes from the data, and hence it can adapt to both homogeneous and heterogeneous patient populations.

Table V presents a detailed comparisons with state-of-the-art risk scoring methodologies, highlighting the limitations of these methods that were addressed by out model.

Appendix B: Data Description

The Patient Cohort and ICD-9 Codes

Experiments were conducted on a cohort of 6,321 patients who were hospitalized in a general medicine floor in the Ronald Reagan UCLA medical center during the period between March 3^{rd} 2013, to February 4^{rd} 2016 (excluding patients who were initially admitted to the ICU and then transferred to the ward after stabilization since for those patients the data were not recorded in the EHR). The patients' population is heterogeneous with a wide variety of diagnoses and ICD-9 codes: the patient's cohort included an overall number of 1,643 ICD-9 codes; the most frequent of which corresponded to conditions such as shortness of breath, hypertension, septicemia, sepsis, fever, pneumonia and renal failure. The distribution of the ICD-9 codes associated with the patients in the cohort is illustrated in Fig. 5 and Table VI. The cohort included patients who were not on immunosuppression and others who were on immunosuppression, including patients that have received solid organ transplantation. In addition, there were some patients that had diagnoses of leukemia or lymphoma. Some of these patients received stem cell transplantation as part of their treatment. Because these patients receive chemotherapy to significantly ablate their immune system prior to stem cell transplantation, they are at an increased risk of clinical deterioration. Of the 6,321 patients (the dataset $\mathcal{D}$), 524 patients experienced clinical deterioration and were admitted to the ICU (the dataset $\mathcal{D}_1$), and 5,788 patients were discharged home (the dataset $\mathcal{D}_o$). Thus, the ICU admission rate is 8.30%. The vast heterogeneity of the patients' cohort motivates the need for a "personalized" risk model, and suggests the general applicability of the experimental results presented in the paper.

Patients in the dataset $\mathcal{D}$ were monitored for 11 vital signs (e.g. O_2 saturation, heart rate, systolic blood pressure, etc) and 10 lab tests (e.g. Glucose, white blood cell count, etc). Hence, the dimension of the physiological stream for every patient is $D = 21$. Each physiological stream is a temporal, irregularly sampled time series, which resemble the data structure depicted in Fig. 4. The ICD-9 code ranges were converted to a set of 18 categorical values, where each value bundles a set of ICD-9 codes for "related diseases"; such a "categorization" allows the algorithm to handle newly hospitalized patients with rare ICD-9 codes that were not present in the dataset. The ICD-9 ranges used for categorization are shown in Table VII. The sampling rate for the physiological streams $\{x_{ij}, t_{ij}\}_{i,j}$ ranges from 1 hour to 4 hours, and the length of hospital stay for the patients ranged from 2 to 2,762 hours. Correlated feature selection (CFS) was used to select the physiological streams that are relevant to predicting the endpoint outcomes (i.e. ICU admission); the CFS algorithm selected 7 vital signs (Diastolic blood pressure, eye opening, Glasgow coma scale score, heart rate, temperature, O_2 device assistance and O_2 saturation), and 3 lab tests (Glucose, Urea Nitrogen and white blood cell count).

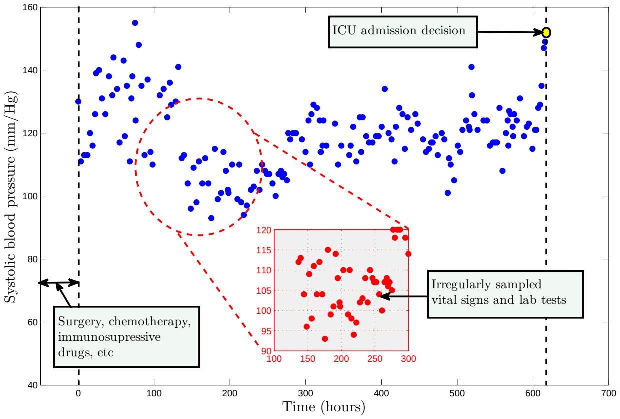


Fig. 4: An exemplary physiological stream for a patient hospitalized in a regular ward.

For all the experiments conducted in the paper, the training and testing datasets are constructed as follows. The training set comprises 5,130 patients who were admitted to the ward in the period between March 2013 and July 2015. Among those patients, the ICU admission rate was 8.34%. The algorithms are trained via this dataset, and then tested on a separate dataset that comprises the remaining 1,191 patients who were admitted to the ward in the period between July 2015 and April 2016.

In Figure 6 we display a snapshot for the temporal risk score trajectories computed by various risk scoring methods for one clinically stable patient, and one clinically deteriorating patient. All risk scores are normalized such that their optimal alarm threshold is fixed at 0.7. For the clinically stable patient, the proposed score as a function of time displays a higher level of smoothness as opposed to the MEWS and Rothman scores which falsely alarm for an ICU admission for that patient because of their drastic fluctuations. For the clinically deteriorating patient, the proposed score is able to track the trend of the patient's clinical deterioration, and hit the alarm threshold quicker than the Rothman index, whereas the MEWS score even fails to identify the patients clinical deterioration. In this case, the patient's clinical status starts to progressively worsen approximately 250 hours prior to the emergent ICU transfer. Our risk model demonstrates a steady increase in risk of clinical deterioration until it crosses the threshold where a warning would be sent to the clinician taking care of the patient. Even after that point, the risk model continues to

Reference	Method	Details	Limitations
[6], [10], [21], [22]	Multitask GPs	Model physiological time series data with a Multitask GP likelihood	• Does not capture non-stationarity. • Does not account for latent patient subtypes. • Estimate observable severity score (which is not available for patients in wards).
[28], [30]	GPs for disease progression models	Model long-term longitudinal disease progression (via severity scores) using subtypes and GP regression for the severity scores	• Does not capture non-stationarity. • Uses the same set of sub-types for the entire population. • Estimate observable severity score (which is not available for patients in wards). • Does not fit for distinguishing between patient latent classes of patients; models only the physiological trajectory of a sick patient.
[31]–[33]	Sliding-window regression	Use the clinical endpoints (ICU admission or discharge) as surrogate labels for a patient's clinical deterioration, and hence used those labels to train a supervised (regression) model using the physiological data in a fixed-size time window before censoring	• Does not capture non-stationarity. • No time-series model: does not exploit the information conveyed in different adjacent sliding window.
[2], [29]	Proportional Hazard Models	Cox regression-based model used to develop a sepsis shock severity score that can handle data streams that are censored due to interventions	• Does not capture non-stationarity. • Relies on the existence of ordered pairs of comparisons for the extent of disease severity at different times (not available for ward patients. • Does not incorporate static information or patient subtypes.

TABLE V: Summary of the state-of-the-art risk scoring methodologies.

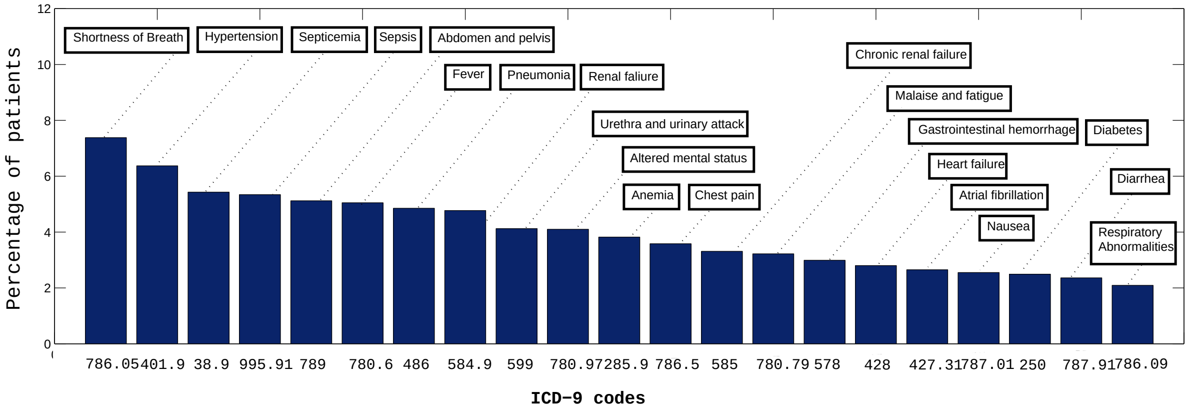


Fig. 5: Distribution of the ICD-9 codes in the patient cohort.

cross the threshold until the patient finally decompensates to the point that the clinician makes the decision to transfer to the ICU. It is worth mentioning that many patients in the cohort under study were receiving chemotherapy or stem cell transplantation, and hence their immune systems often do not recover for several days during which time they are at increased risk of infection. The fact that our risk model can predict several days prior to the actual clinical deterioration event provides hope that an earlier intervention can be provided to reverse the course of decompensation.

Appendix C: Algorithmic Details

The EM Algorithm

We show the E and M steps for the EM algorithm (Algorithm 1) for the clinically stable patients. The same steps are conducted for the deteriorating patients but separately for every epoch.

We start by writing the proximal likelihood function as follows:

$$Q(\Gamma_o;\Gamma_o^{p-1}) = \mathbb{E}[\log(\mathbb{P}(\mathcal{D}_o,\{Z^{(n)}\}_{n=1}^{N_o}\,|\,\Gamma_o))\,|\,\mathcal{D}_o,\Gamma_o^{p-1}],$$

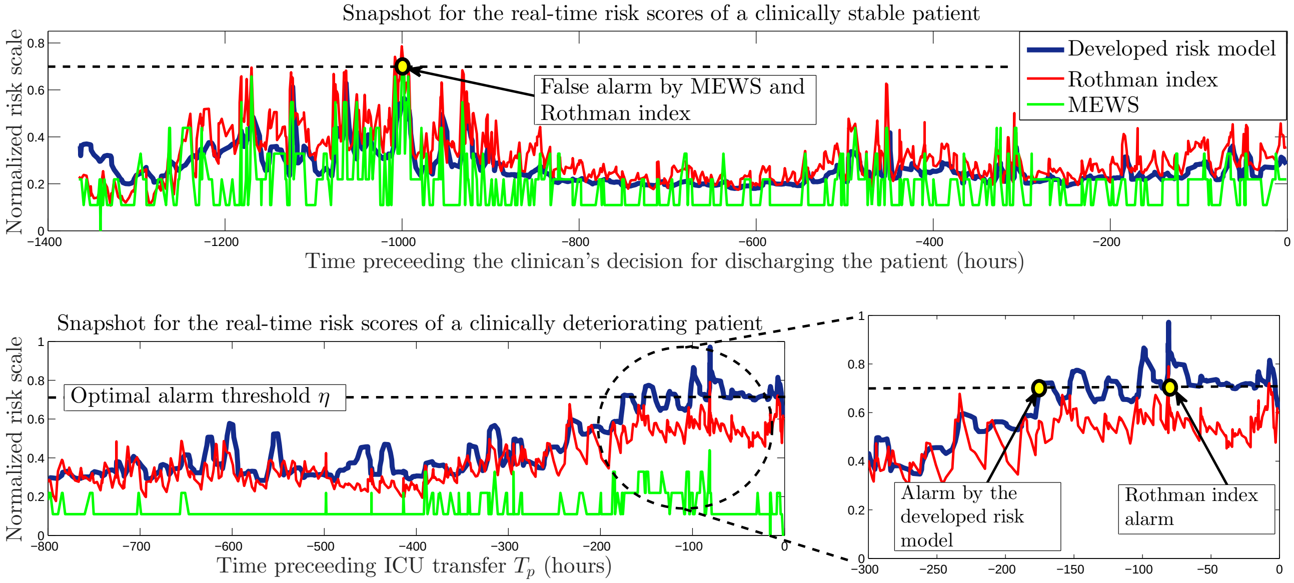


Fig. 6: Snapshots for real-time risk scores computed for two typical patients.

TABLE VI: ICD-9 codes in the patient cohort under study.

ICD-9 code	Diagnosis	% Freq.
(786.05)	Shortness of Breath	7%
(401.9)	Hypertension	6%
(38.9)	Septicemia	5%
(995.91)	Sepsis	5%
(780.6)	Fever	5%
(486)	Pneumonia	5%
(584.9)	Renal failure	5%
(599)	Urethra and urinary attack	5%
(780.97)	Altered mental status	4%
(285.9)	Anemia	4%
(786.5)	Chest pain	4%
(585)	Chronic renal failure	4%
(780.79)	Malaise and fatigue	3%
(578)	Gastrointestinal hemorrhage	3%
(428)	Heart failure	3%
(427.31)	Atrial fibrillation	3%
(787.01)	Nausea	3%
—	Other	22.5%

where $Z^{(n)}$ is the latent subtype of the n^{th} entry of the dataset $\mathcal{D}_o$. The parametrization is updated in the M-step by maximizing $Q(\Gamma_o;\Gamma_o^{p-1})$ with respect to Γ_o (closed-form expressions are available for the jointly Gaussian data in $\mathcal{D}_o$ as per the GP model). The p^{th} iteration is concluded by updating expert z's responsibility towards the n^{th} patient in the dataset $\mathcal{D}_o$ as follows

$$\begin{aligned}\beta_{z,p}^{(n)} &= \mathbb{P}(Z^{(n)} = z \,|\, \{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}, \Gamma_o^p) \\ &= \frac{\pi_z^p\, f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} \,|\, \mathbf{\Theta}_o^{p,z})}{\sum_{z'=1}^{G} \pi_{z'}^p\, f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} \,|\, \mathbf{\Theta}_o^{p,z'})},\end{aligned}$$

where π_z^p is the estimate for $\mathbb{P}(Z = z)$ in the p^{th} iteration, and $f(.)$ is the Gaussian distribution function. The term $\beta_{z,p}^{(n)}$ represents the posterior probability of patient n's membership in subtype z given the realization of her physiological data $\{x_{ij}, t_{ij}\}_{i,j}$.

Given the above, we can rewrite the proximal likelihood function as follows

$$\begin{aligned}Q(\Gamma_o;\Gamma_o^{p-1}) &= \mathbb{E}[\log(\mathbb{P}(\mathcal{D}_o, \{Z^{(n)}\}_{n=1}^{N_o} \,|\, \Gamma_o)) \,|\, \mathcal{D}_o, \Gamma_o^{p-1}] \\ &= \mathbb{E}[\log(\prod_{n=1}^{N_o} \mathbb{P}(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}, Z^{(n)} \,|\, \Gamma_o)) \,|\, \Gamma_o^{p-1}] \\ &= \sum_{n=1}^{N_o} \mathbb{E}[\log(\mathbb{P}(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j}, Z^{(n)} \,|\, \Gamma_o)) \,|\, \Gamma_o^{p-1}] \\ &= \sum_{n=1}^{N_o} \sum_{z} \beta_{z,p}^{(n)} \log(f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} \,|\, \Theta_o^z)),\end{aligned}$$

where the expression for the Gaussian distribution $f(\{x_{ij}^{(n)}, t_{ij}^{(n)}\}_{i,j} \,|\, \Theta_o^z)$ can be easily formulated by constructing the corresponding covariance matrix.

Similar to conventional Gaussian mixture models, the M-step proceeds as follows:

$$\begin{aligned}\pi_z^{p+1} &= \frac{1}{N_o} \sum_{n=1}^{N_o} \beta_{z,p}^{(n)} \\ m_o^{z,p+1}(t,i) &= \frac{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)} \sum_i x_{ij}^{(n)}}{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)}},\end{aligned}$$

where $m_o^{z,p+1}(t,j)$ is the constant mean function for the j^{th} physiological stream in subtype z. Our adoption of a constant mean function allows us to use the direct weighted sample mean as the updated mean function in each EM iteration. The covariance parameters $\mathbf{\Sigma}$ and ℓ are estimated separately conditioned on every subtype using the gradient method in [36] (an online MATLAB package for hyper-parameter tuning is provided by the authors), this yields a set of subtype-specific estimates $\hat{\mathbf{\Sigma}}_z^n$ and $\hat{\ell}_z^n$ for every patient's time series, which are

TABLE VII: ICD-9 codes in the patient cohort under study.

ICD-9 code	Category	Categorical value
001-139	Infectious and parasitic diseases	1
140-239	Neoplasms	2
240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders	3
280-289	Blood diseases and blood-forming organs	4
90-319	Mental disorders	5
320-359	Nervous system diseases	6
360-389	Sense organs diseases	7
390-459	Circulatory system diseases	8
460-519	Respiratory system diseases	9
520-579	Digestive system diseases	10
580-629	Genitourinary system diseases	11
630-679	Pregnancy, childbirth, and the puerperium complications	12
680-709	Skin and subcutaneous tissue diseases	13
710-739	Musculoskeletal system and connective tissue diseases	14
740-759	Congenital anomalies	15
760-779	Conditions originating in perinatal period	16
780-799	Symptoms, signs, and ill-defined conditions	17
800-999	Injury and poisoning	18

used to update the covariance hyper-parameters as follows:

$$\hat{\ell}_o^{z,p+1} = \frac{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)} \hat{\ell}_z^n}{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)}}$$

$$\boldsymbol{\Sigma}_o^{z,p+1} = \frac{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)} \hat{\boldsymbol{\Sigma}}_z^n}{\sum_{n=1}^{N_o} \beta_{z,p}^{(n)}}.$$

Depending on the problem and the size of the dataset, this process can be computationally expensive, in which case the EM algorithm can be terminated after a predefined number of iterations.

Computation of $\mathbb{P}(\bar{k}|\{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1)$

We compute $\mathbb{P}(\bar{k}|\{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1)$ recursively every T_1 hours in a similar manner to the forward filtering algorithm used for inference in Hidden Markov Models; forward messages are fixed over segments of length T_1. We define the *forward message* $\alpha_t(h_t)$ as follows

$$\alpha_k(h_k) = \mathbb{P}(h_k, \{x_{ij}, (k-1)\,T_1 \leq t_{ij} \leq k\,T_1\}_{ij}, \Gamma_1),$$

where h_k is the latent epoch index, and $\alpha_o(h_k = \bar{k}) = f(h_k = \bar{k})$. The forward messages can be computed using the following dynamic programming recursion

$$\alpha_k(h_k) = \mathbb{P}(\{x_{ij}, (k-1)\,T_1 \leq t_{ij} \leq k\,T_1\}_{ij} \,|\, h_k, \Gamma_1) \times$$
$$\mathbb{P}(h_k|h_{k-1} = h_k - 1)\, \alpha_{k-1}(h_{k-1}).$$

Note that the valid values of h_k are restricted such that $h_k \in \{K - k + 1, \ldots, K\}$. The posterior distribution of the latent epoch index is easily evaluated using Bayes rule as follows

$$\mathbb{P}(\bar{k}\,|\,\{x_{ij}, t_{ij} \leq t\}_{ij}, \Gamma_1) = \frac{\alpha_k(h_k)}{\sum_{h=K-k+1}^{K} \alpha_k(h)}.$$

Appendix D: Details of the Benchmark Algorithms

- **Feature Selection:** The correlated feature selection (CFS) algorithm was used to select the relevant features for all the algorithms [41]. The same relevant features were used for all the benchmarks. All excluded features were realized to be highly irrelevant by virtue of their CFS relevance scores. The CFS selected 7 vital signs (Diastolic blood pressure, eye opening, Glasgow coma scale score, heart rate, temperature, O_2 device assistance and O_2 saturation), and 3 lab tests (Glucose, Urea Nitrogen and white blood cell count). These features, augmented with all the static admission information were used to train the benchmarks.
- **Validation:** We divided the data into a training set of 4,130 patients and a validation set of 1,000 patients. The same splits were used for all the benchmarks. The validation set was used to tune the hyper-parameters of each algorithm by optimizing its AUC.
- **Feature Extraction:** In order to ensure that the information in the clinical endpoints are utilized properly by all the sliding-window predictors, we trained every predictor by constructing a training dataset that comprises: (1) the physiological data gathered within a temporal window before the terminating event (ICU admission or patient discharge), and using the clinical endpoints as the labels, (2) summary statistics of the entire time series episode (means, standard deviations, skewness, kurtosis, maximum and minimum values), and (3) the static features. This creates a fixed length training set to train the model. In real-time, a sliding-window is used to extract sequential data from the running time series, augments it with the summary statistics up to the current time and the static features, and a risk score is used as a sliding-window regression outcome. The size of this window is a hyper-parameter that is tuned separately for every predictor.

Appendix E: Modeling Rationale, Assumptions and Some Comments

Connection to Latent Variable Models

A latent variable model with state transitions (such as the Markov model depicted in Figure 7) is indeed a very natural approach to model the patients' clinical states. We note that our model is a latent variable model; one can think of our model as

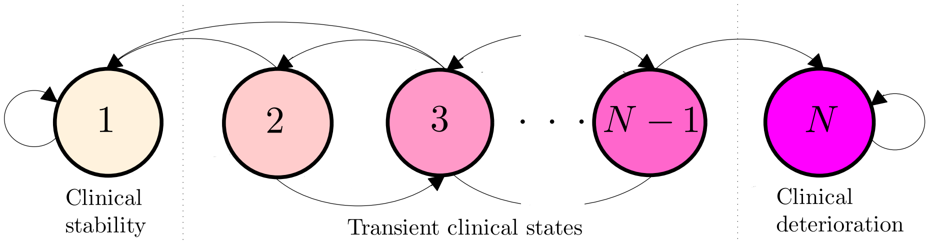


Fig. 7: A latent variable model for the clinical state.

a state model with 2 latent absorbing states (this corresponds to the model in Figure 7 but with only states 1 and N), in which risk scoring boils down to testing the true hypothesis about the identity of the hidden absorbing state that generates a patient's physiological trajectory. Thus, our theoretical formulation of the problem as a sequential hypothesis test with an uncertain time horizon is equivalent to a limiting case of real-time filtering of a state-space model in which we have only two states. We note that both types of models capture non-stationarity: in our model, the "fixed" latent states has non-stationary "emission distribution", whereas in state-space models, the states have a stationary emission distribution and non-stationarity is captured by "state-switching" over time. We have tried both types of models as conceptual apparatuses for risk scoring, and we decided to go with the sequential hypothesis testing framework for the following reason. A latent variable model with more than 2 states will entail the need for inferring the **hidden** state trajectories for every patient' physiological stream. Since the ICU data is not labeled by clinical state at any point other than the endpoint of ICU admission or deterioration, one would need an unsupervised algorithm to learn these hidden clinical state representations. This means that in addition to the patient subtype variables which are hidden, we will also have a hidden state trajectory for every patient. This significantly complicates the learning problem, and the usage of the EM algorithm for learning such a model may converge to a considerably bad local optimum. We believe that it is much more reasonable to reduce the number of hidden variables in the model in order to ensure robustness and consistency of different versions of the model that would be learned whenever the EHR data is updated.

The Conception of Subtyping

Figure 8 depicts what we believe to be the most accurate and expressive conception of patient subtypes; such a conception has been developed under the guidance of our clinical collaborator. To illustrate our conception of subtyping, let us assume that we only have one static feature, say the patient's ICD-9 code, and there are two possible types of patients: type A and type B. If the ICD-9 code corresponds to a blood cancer (e.g. Leukemia), then the patient is allocated to type A, whereas if the ICD-9 code corresponds to Pneumonia, then the patient is allocated to type B. Both types of patients have very different stability patterns since their different illnesses (or even different gender and ethnicity) dictate different nominal values for their stable physiological data. Both patients also have different "possible" patterns of deterioration, depending on the nature of the adverse event they may encounter. Type A patients are more likely to experience Leukemia-related adverse events (e.g. adverse cytogenetics outcomes), whereas type B patients are more likely to experience respiratory-related adverse events (e.g. respiratory arrests); and hence, conditioned on the patients' diagnoses, they have very different deterioration patterns.

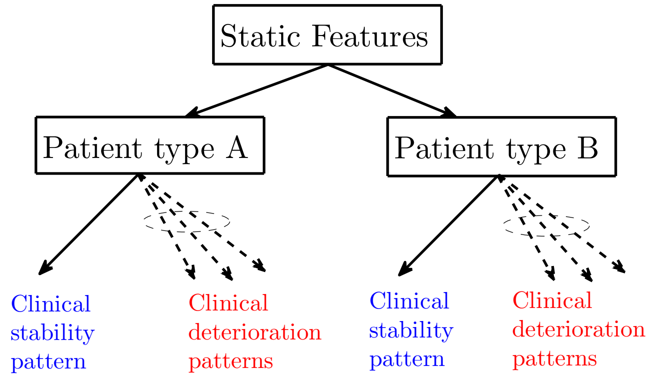

Static Features
Patient type A
Patient type B
Clinical stability pattern
Clinical deterioration patterns
Clinical stability pattern
Clinical deterioration patterns

Fig. 8: Configuration of clinical stability and deterioration patterns.

The model described above assumes that the patient's subtype is fully determined by her static admission features, and then conditioned on her subtype, there is one nominal stability pattern and multiple possible deterioration patters. Following this model, we can use the same z-classifier network to allocate patients to subtypes, and then apply an$M$*-ary sequential hypothesis test* to test whether the patient is stable, or experiencing 1 out of $M-1$ possible deterioration patterns. Since we had no labels that designate the different adverse events in the dataset, our model is an approximate version of the one in Figure 8, in which we treat all deterioration pattern as coming from one, more dispersed distribution (i.e. this reflects in the form of a larger variance in the GP parameters), but this "average" model sufficiently differs from the stability model and hence a simple binary sequential test is sufficient for risk scoring. Our usage of the same z-classifier network for stable and deteriorating groups is motivated by the fact that even if the patients' deterioration patterns are more diverse and could be clustered into more "deterioration subtypes", those finer "deterioration subtypes" are not logically independent of the "stability subtypes", but they are rather subsets of them (as conceptualized in Figure 8). The only approximation we do here is that we collapse all deterioration patterns within a subtype into one representative model, and the motive behind such an approximation is the lack of data on what adverse event is associated with every patient, and unsupervised model for further clustering the deteriorating cohort seems infeasible due to the scarcity of data in that cohort. We also note that grouping models of stability and deterioration together into logically related subtypes has also an advantage in terms of medical interpretability, which would be lost if we have two different groups of clusters that are conditioned on the patient's (unseen) clinical state.

We also note that our sub-typing model is not only found to be more plausible from the clinical perspective, but it is more statistically efficient as well. That is, since the ICU admission rate is only (less than) 10%, if we attempt to learn a disjoint group of sub-types (a different z-classifier network) for the deteriorating group, we will have much less data and we will be required to learn more sub-types than the 6 discovered from the stable patients (since as we argued earlier, there are more patterns of deterioration than for stability. Selecting a separate deterioration sub-type model using Bayes factors yields only 4 clusters!). By "transferring" the knowledge gained from the clinically stable cohort to the clinically deteriorating one (in terms of subtype definitions), we are able to re-sample a reasonably large dataset from the deteriorating cohort for every sub-type and hence accurately learn the deterioration model parameters (step 23 in Algorithm 1), without needing to bear the burden of jointly discover the sub-typing configuration already learned from the stable patient. Our ability to transfer the sub-typing knowledge from stable to deteriorating patients hinges on the logical association between the two groups; such a logical association assumption is believed by our medical collaborator to be clinically sound.

Multi-task Gaussian Processes

It is important to note that multi-task Gaussian processes with an intrinsic correlation model for the co-variance structure entails the assumption of a common temporal length-scale for all the physiological stream. This is does not reflect the differences in the rate of fluctuations of the different streams; for instance, heart rate changes much faster than a signal like creatinine level. We have initially tried to construct a kernel function that captures heterogeneous length-scales by using a linear corregionalization model that adds multiple kernels with diverse length scale, but this led to an unnecessarily much more complicated model with many more parameters and a less tractable likelihood function. Our choice for a multitask Gaussian process is justified by the fact that we are not interested in finding a good fit for the physiological data, but we are rather interested in capturing the aspects of the physiological streams that distinguish stable and deteriorating patients. The correlation structure significantly differ between stable and deterioration patients (for instance, respiratory rate and heart rate are much less correlated for deteriorating patients at different epochs as compared to stable patients.), whereas the length-scale parameter does not differ much for the two models. To demonstrate the difference between the correlation structures of the stable and deteriorating patients, see below the correlation matrices for the stable patient's model, and the deteriorating patients' model for the K^{th} epoch for the physiological streams (diastolic blood pressure, heart rate, respiratory rate, SpO2, Glucose, urea nitrogen):

$$\mathbf{\Sigma}_o = \begin{bmatrix} 138 & 20 & 2 & 1 & 7 & 0 \\ 20 & 237 & 4 & -1 & 12 & -12 \\ 2 & 4 & 6 & 0 & 1 & 0 \\ 1 & -1 & 0 & 4 & -1 & -1 \\ 7 & 12 & 1 & -1 & 315 & -1 \\ 0 & -12 & 0 & -1 & -1 & 20 \end{bmatrix},$$

$$\mathbf{\Sigma}_{1,K} = \begin{bmatrix} 259 & 43 & -3 & 9 & -58 & -42 \\ 43 & 185 & 3 & -2 & -28 & -18 \\ -3 & 3 & 12 & -1 & 7 & 7 \\ 9 & -2 & -1 & 61 & -11 & 39 \\ -58 & -28 & 7 & -11 & 958 & 175 \\ -42 & -18 & 7 & 39 & 175 & 885 \end{bmatrix},$$

where the correlation coefficients are rounded to the nearest integer. As can be seen in $\mathbf{\Sigma}_o$ and $\mathbf{\Sigma}_{1,K}$, not only that the extent of correlation between the different physiological variables differ under the two hypothesis, but the nature of correlations differ as well (i.e. some physiological measurements are positively correlated for stable patients and negatively correlated for deteriorating ones). Hence, in terms of the accuracy of the sequential hypothesis test, we much better off by considering the distinguishing inter-stream correlation structures than when ignoring correlations and consider the non-distinguishing stream specific length-scales.

The Graphical Model

The conditional independence assumptions in eq. (13) can be interpreted as follows: conditioned on the patient's static features, the clinical state is independent of the subtype, and conditioned on the subtype, the clinical state is independent on the static features. This means that one can generate samples from our model by drawing a clinical state from the prior distribution, and then drawing a static feature instance (independent of the clinical state), and then drawing a sub-type indicator variable conditioned on that instance. The reason that we assumed that the clinical state is independent on the patient's subtype (and static feature) is that the ICU admission rate is very balanced across all the patient groups in Table VII (and consequently the ICU admission rate is balanced across all the 6 discovered subtypes). This encourages adopting the simplifying assumption of the clinical state being independent of the sub-type, which further simplifies the real-time computation of the Bayesian posterior probability.

Length of Stay Exceeding $K\,T_1$

Most patients in the data set have hospitalization times that do not exceed the length $K\,T_1$. If the patient's length of stay exceeds $K\,T_1$, the model corresponding to hypothesis $\mathcal{H}_1$ is assumed to be trapped in the last epoch, i.e. the physiological streams become stationary after this point. Patients who have very long stays in the ward and never admitted to the ICU are overwhelmingly more likely to be stable and undergoing a routine hospitalization procedure; for these patients one can safely assume a stationary model for deterioration without losing predictive power.

Impact of Temporal Alignment and Length of Stay on Training Data

The temporal alignment via the clinical endpoints is indeed a source of imbalance in the number of data points available for training every epoch. Fortunately, as shown in Figure 9 the consequences of this imbalance affects the earlier epochs but does not affect the latest epochs, which are much more crucial since they are closer to the clinical deterioration onset. The impact of the availability of few data points for earlier epochs is that it leads to higher false alarm rates, but it does not affect the detection probability in any way. We also note that even for the earlier epochs for which less data points are available, there is enough temporal data within the same patient's temporal stream to obtain a decent estimate for the GP hyper-parameters. We truncated the physiological stream lengths to exclude epoch numbers that would have fewer than 5 patients (every epoch for a single patient still have hundreds of temporal data points, which allows for a decent estimate for the length scale and mean parameters).

The Usage of Fixed Epoch Lengths

One can think of our model as a semi-Markov model with restricted left-to-right transitions among two groups of disconnected states as shown in Figure 10, and with the epoch intervals being the states' sojourn times. In this case, T_1 (and T_o) are random and drawn from a pre-defined distribution. We have initially modeled T_1 as a random variable drawn from an epoch-specific Gamma distribution, and we used the non-parametric E-divisive change-point detection algorithm to estimate the epoch length distributions. This turned out not be useful for the following reasons:

- This distributions for the different epochs' lengths were quite similar.
- The estimated Gamma distributions had a significantly large *shape* parameter, which implies a small value for the variance.

Updating the posterior probabilities while considering random epoch lengths did not provide us with statistically significant AUC gains. For this reason, we adopted a simpler model in which the epoch lengths are modeled as a degenerate random variable that only differs between stable and deteriorating patients.

References

[1] N. Broutet, F. Krauer, M. Riesen, A. Khalakdina, M. Almiron, S. Aldighieri, M. Espinal, N. Low, and C. Dye, "Zika virus as a cause of neurologic disorders," *New England Journal of Medicine*, vol. 374, no. 16, pp. 1506–1509, 2016.

[2] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[3] M. M. Churpek, T. C. Yuen, S. Y. Park, R. Gibbons, and D. P. Edelson, "Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards," *Critical care medicine*, vol. 42, no. 4, p. 841, 2014.

[4] L. L. Kirkland, M. Malinchoc, M. OByrne, J. T. Benson, D. T. Kashiwagi, M. C. Burton, P. Varkey, and T. I. Morgenthaler, "A clinical deterioration prediction tool for internal medicine patients," *American Journal of Medical Quality*, vol. 28, no. 2, pp. 135–142, 2013.

[5] M. J. Rothman, S. I. Rothman, and J. Beals, "Development and validation of a continuous measure of patient condition using the electronic medical record," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 837–848, 2013.

[6] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian process regression in vital-sign early warning systems," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 6161–6164.

[7] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, "Viewstowards a national early warning score for detecting adult inpatient deterioration," *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.

[8] M. P. Young, V. J. Gooder, K. Bride, B. James, and E. S. Fisher, "Inpatient transfers to the intensive care unit," *Journal of general internal medicine*, vol. 18, no. 2, pp. 77–83, 2003.

[9] G. D. Finlay, M. J. Rothman, and R. A. Smith, "Measuring the modified early warning score and the rothman index: advantages of utilizing the electronic medical record in an early warning system," *Journal of hospital medicine*, vol. 9, no. 2, pp. 116–119, 2014.

[10] M. A. Pimentel, D. A. Clifton, L. Clifton, P. J. Watkinson, and L. Tarassenko, "Modelling physiological deterioration in post-operative patient vital-sign data," *Medical & biological engineering & computing*, vol. 51, no. 8, pp. 869–877, 2013.

[11] J. Kause, G. Smith, D. Prytherch, M. Parr, A. Flabouris, K. Hillman *et al.*, "A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in australia and new zealand, and the united kingdomthe academia study," *Resuscitation*, vol. 62, no. 3, pp. 275–282, 2004.

[12] H. Hogan, F. Healey, G. Neale, R. Thomson, C. Vincent, and N. Black, "Preventable deaths due to problems in care in english acute hospitals: a retrospective case record review study," *BMJ quality & safety*, pp. bmjqs–2012, 2012.

[13] D. Mokart, J. Lambert, D. Schnell, L. Fouché, A. Rabbat, A. Kouatchet, V. Lemiale, F. Vincent, E. Lengliné, F. Bruneel *et al.*, "Delayed intensive care unit admission is associated with increased mortality in patients with cancer with acute respiratory failure," *Leukemia & lymphoma*, vol. 54, no. 8, pp. 1724–1729, 2013.

[14] P. S. A. Group, "Safe use of opioids in hospitals," *Sentinel Event Alert*, pp. bmjqs–2012, 2014.

[15] J. D. Mardini L, Lipes J, "Adverse outcomes associated with delayed intensive care consultation in medical and surgical inpatients," *Journal of critical care*, vol. 27, no. 6, pp. 688–693, 2012.

[16] R. M. Merchant, L. Yang, L. B. Becker, R. A. Berg, V. Nadkarni, G. Nichol, B. G. Carr, N. Mitra, S. M. Bradley, B. S. Abella *et al.*, "Incidence of treated cardiac arrest in hospitalized patients in the united states," *Critical care medicine*, vol. 39, no. 11, p. 2401, 2011.

[17] G. Kumar, N. Kumar, A. Taneja, T. Kaleekal, S. Tarima, E. McGinley, E. Jimenez, A. Mohan, R. A. Khan, J. Whittle *et al.*, "Nationwide trends of severe sepsis in the 21st century (2000-2007)," *Chest Journal*, vol. 140, no. 5, pp. 1223–1231, 2011.

[18] C. Hershey and L. Fisher, "Why outcome of cardiopulmonary resuscitation in general wards is poor," *The Lancet*, vol. 319, no. 8262, pp. 31–34, 1982.

[19] A. Wald, *Sequential analysis*. Courier Corporation, 1973.

[20] G. Peskir and A. Shiryaev, *Optimal stopping and free-boundary problems*. Springer, 2006.

[21] R. Durichen, M. A. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask gaussian processes for multivariate physiological time-series analysis," *Biomedical Engineering, IEEE Transactions on*, vol. 62, no. 1, pp. 314–322, 2015.

[22] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data." in *AAAI*, 2015, pp. 446–453.

[23] S. Saria and A. Goldenberg, "Subtyping: What it is and its role in precision medicine," *Intelligent Systems, IEEE*, vol. 30, no. 4, pp. 70–75, 2015.

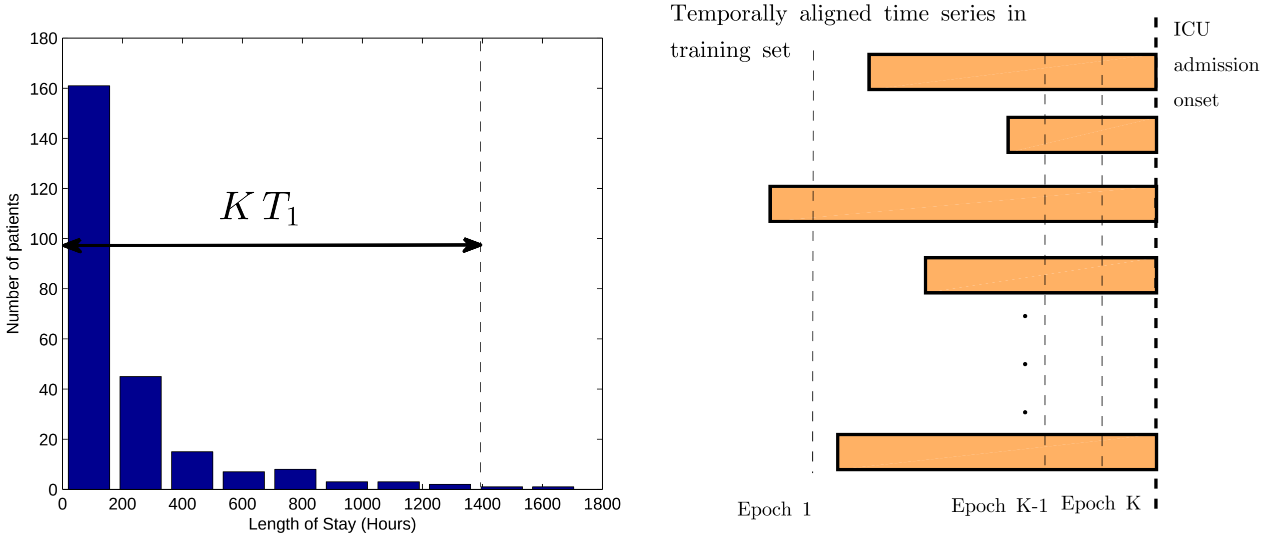


Fig. 9: Interplay between length of stay and training data available for every epoch.

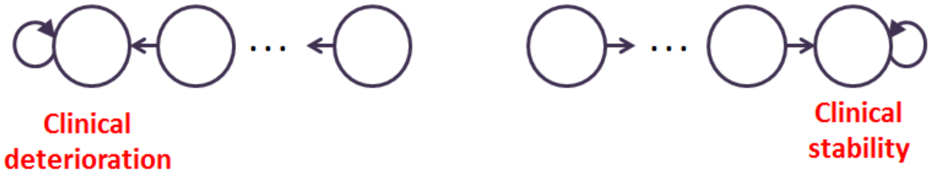


Fig. 10: Deterministic left-to-right transitions.

[24] R. Morgan, F. Williams, and M. Wright, "An early warning scoring system for detecting developing critical illness," *Clin Intensive Care*, vol. 8, no. 2, p. 100, 1997.

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

[26] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomedical Instrumentation & Technology*, vol. 46, no. 4, pp. 268–277, 2012.

[27] J. P. Bliss and M. C. Dunn, "Behavioural implications of alarm mistrust as a function of task workload," *Ergonomics*, vol. 43, no. 9, pp. 1283–1300, 2000.

[28] P. Schulam and S. Saria, "A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure," in *Advances in Neural Information Processing Systems*, 2015, pp. 748–756.

[29] K. Dyagilev and S. Saria, "Learning (predictive) risk scores in the presence of censoring due to interventions," *Machine Learning*, pp. 1–26, 2015.

[30] J. Futoma, M. Sendak, C. B. Cameron, and K. Heller, "Predicting disease progression with a model for multivariate longitudinal clinical data," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, 2016, pp. 42–54.

[31] J. C. Ho, C. H. Lee, and J. Ghosh, "Imputation-enhanced prediction of septic shock in icu patients," in *Proceedings of the ACM SIGKDD Workshop on Health Informatics*, 2012, pp. 21–27.

[32] S. Saria, A. K. Rajani, J. Gould, D. Koller, and A. A. Penn, "Integration of early physiological responses predicts later illness severity in preterm infants," *Science translational medicine*, vol. 2, no. 48, pp. 48ra65–48ra65, 2010.

[33] J. Wiens, E. Horvitz, and J. V. Guttag, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," in *Advances in Neural Information Processing Systems*, 2012, pp. 467–475.

[34] J. Yoon, A. Alaa, S. Hu, and M. van der Schaar, "Forecasticu: A prognostic decision support system for timely prediction of intensive care unit admission," pp. 1680–1689, 2016.

[35] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized risk scoring for critical care patients using mixtures of gaussian process experts," *ICML workshop on Computational Frameworks in Personalization*, 2016.

[36] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[37] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[38] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 1, pp. 193–197, 2013.

[39] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[40] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[41] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[42] O. Hayani, A. Al-Beihany, R. Zarychanski, A. Chou, A. Kharaba, A. Baxter, R. Patel, and D. Allan, "Impact of critical care outreach on hematopoietic stem cell transplant recipients: a cohort study," *Bone marrow transplantation*, vol. 46, no. 8, pp. 1138–1144, 2011.

[43] F. Rincon, T. Morino, D. Behrens, U. Akbar, C. Schorr, E. Lee, D. Gerber, J. Parrillo, and T. Mirsen, "Association between out-of-hospital emergency department transfer and poor hospital outcome in critically ill stroke patients," *Journal of critical care*, vol. 26, no. 6, pp. 620–625, 2011.

[44] C. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *Qjm*, vol. 94, no. 10, pp. 521–526, 2001.

[45] S. Yu, S. Leung, M. Heo, G. J. Soto, R. T. Shah, S. Gunda, and M. N. Gong, "Comparison of risk prediction scoring systems for ward patients:

a retrospective nested case-control study," *Critical Care*, vol. 18, no. 3, p. 1, 2014.

[46] D. M. Bliss JP, "Behavioural implications of alarm mistrust as a function of task workload," *Ergonomics*, vol. 43, no. 9, pp. 1283–1300, 2010.

[47] V. Liu, P. Kipnis, N. W. Rizk, and G. J. Escobar, "Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system," *Journal of hospital medicine*, vol. 7, no. 3, pp. 224–230, 2012.

[48] R. Snyderman, "Personalized health care: From theory to practice," *Biotechnology journal*, vol. 7, no. 8, pp. 973–979, 2012.

[49] G. J. Escobar, J. C. LaGuardia, B. J. Turk, A. Ragins, P. Kipnis, and D. Draper, "Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record," *Journal of hospital medicine*, vol. 7, no. 5, pp. 388–395, 2012.

[50] L. Landro, "Hospitals find new ways to monitor patients 24/7 (link: http://www.wsj.com/articles/hospitals-find-new-ways-to-monitor-patients-24-7-1432560825)," *The Wall Street Journal*, 2015.

[51] N. Alam, E. Hobbelink, A. van Tienhoven, P. van de Ven, E. Jansma, and P. Nanayakkara, "The impact of the use of the early warning score (ews) on patient outcomes: a systematic review," *Resuscitation*, vol. 85, no. 5, pp. 587–594, 2014.

[52] D. Goldhill, A. McNarry, G. Mandersloot, and A. McGinley, "A physiologically-based early warning score for ward patients: the association between score and outcome*," *Anaesthesia*, vol. 60, no. 6, pp. 547–553, 2005.

[53] C. S. Parshuram, J. Hutchison, and K. Middaugh, "Development and initial validation of the bedside paediatric early warning system score," *Crit Care*, vol. 13, no. 4, p. R135, 2009.

[54] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive care medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[55] A. E. Jones, S. Trzeciak, and J. A. Kline, "The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation," *Critical care medicine*, vol. 37, no. 5, p. 1649, 2009.

[56] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent, "Serial evaluation of the sofa score to predict outcome in critically ill patients," *Jama*, vol. 286, no. 14, pp. 1754–1758, 2001.

[57] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "Apache ii: a severity of disease classification system." *Critical care medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[58] A. Goel, R. G. Pinckney, and B. Littenberg, "Apache ii predicts long-term survival in copd patients admitted to a general medical ward," *Journal of general internal medicine*, vol. 18, no. 10, pp. 824–830, 2003.

[59] C. Proust-Lima, J.-F. Dartigues, and H. Jacqmin-Gadda, "Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach," *Statistics in medicine*, vol. 35, no. 3, pp. 382–398, 2016.

[60] Z. Liu and M. Hauskrecht, "A regularized linear dynamical system framework for multivariate time series analysis," in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2015. NIH Public Access, 2015, p. 1798.

[61] K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized predictive modeling and risk factor identification using patient similarity," *AMIA Summits on Translational Science Proceedings*, vol. 2015, p. 132, 2015.

[62] X. Wang, F. Wang, J. Hu, and R. Sorrentino, "Towards actionable risk stratification: a bilinear approach," *Journal of biomedical informatics*, vol. 53, pp. 147–155, 2015.