Supporting Document for: "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*

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 [1]–[7]. Prognosis in hospital wards is feasible since unanticipated adverse events are often preceded by disorders in a patient's physiological parameters [8], [9]. 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 [10], whereas underestimating her risk can undermine the effectiveness of consequent therapeutic interventions [11], [12].

Recent systematic reviews have shown that currently deployed expert-based risk scores, such as the MEWS score [13], provide only modest contributions to clinical outcomes [14]–[16]. 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) [2]. Recently, a data-driven risk score, named the Rothman index, has been developed using regression analysis [3], and was shown to outperform the MEWS score and its variants [7]. 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 [17].

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 [13]. Notable examples of such scores are MEWS and its variant VitalPAC [5]. These scores rely mainly on experts to specify the risk factors and the risk scores associated with these factors [10]. 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 [14]–[16]. 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 [18].

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. [2], and by Escobar et al. [19]. 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 [3], [7]. 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 [20]. While validation of the Rothman index have shown its superiority to MEWS-based models in terms of false alarm rates [7], 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 [3]). 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

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).

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 [4], [22], [30]–[37]; some of the previous works have also adopted multitask GP models [4], [30]–[33]. 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, [33] focuses on predicting the PFVC clinical marker (a measure of lung severity) for scleroderma patients, [4], [30]–[32] focus on predicting the future values of SOFA, APACHE and SAPS scores for ICU patients, and [37] 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* [38]. 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 [4], [30]–[33]. 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 [33], [37], 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 [39], [40] and SVMs [41]. 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 [22] and [34], 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 [42] and [43], 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 II 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. 2 and Table III. The cohort included patients who were not on immunosuppression and others who were on immunosuppression, including patients that have received

Reference	Risk scores	Details	Limitations
[5], [10], [13], [21]–[24]	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 [14]–[16].
[18], [25]–[27]	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 [18] 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 [25]). • 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.
[18], [28], [29]	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 [28]). It has been shown in [18] 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).
[3], [7]	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 [3]). • Uses the instantaneous vital signs and lab tests measurements, and ignores the physiological stream trajectory.

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

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. 1. 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 IV. The sampling rate for the physiological streams $\{x_{ij}, t_{ij}\}_{i,j}$

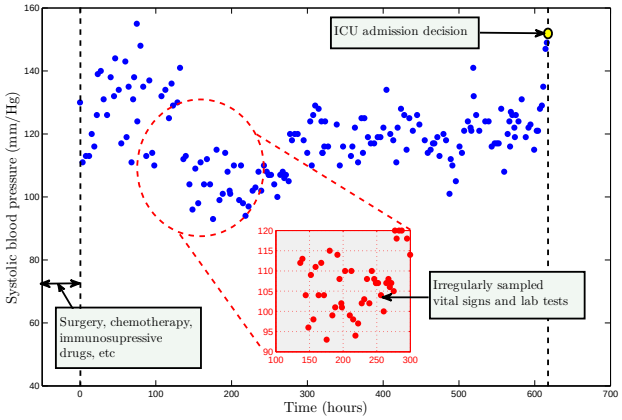

Fig. 1: An exemplary physiological stream for a patient hospitalized in a regular ward.

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

Reference	Method	Details	Limitations
[4], [30]–[32]	Multitask GPs	Model physiological time series data with a Multitask GP likelihood	• Does not capture non-stationarity. • Does not account for latent patient sub-types. • Estimate observable severity score (which is not available for patients in wards).
[33], [37]	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.
[39]–[41]	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.
[22], [34]	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 II: Summary of the state-of-the-art risk scoring methodologies.

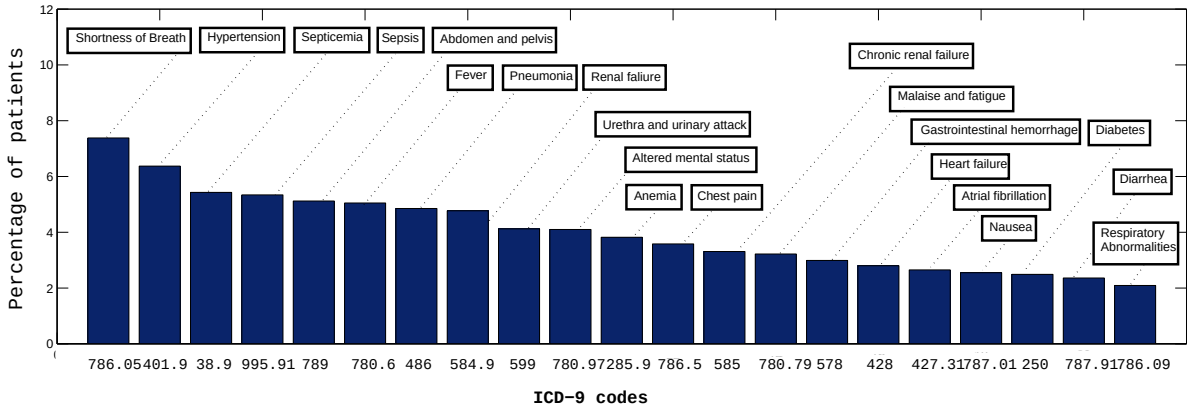

Fig. 2: Distribution of the ICD-9 codes in the patient cohort.

blood cell count).

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 3 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

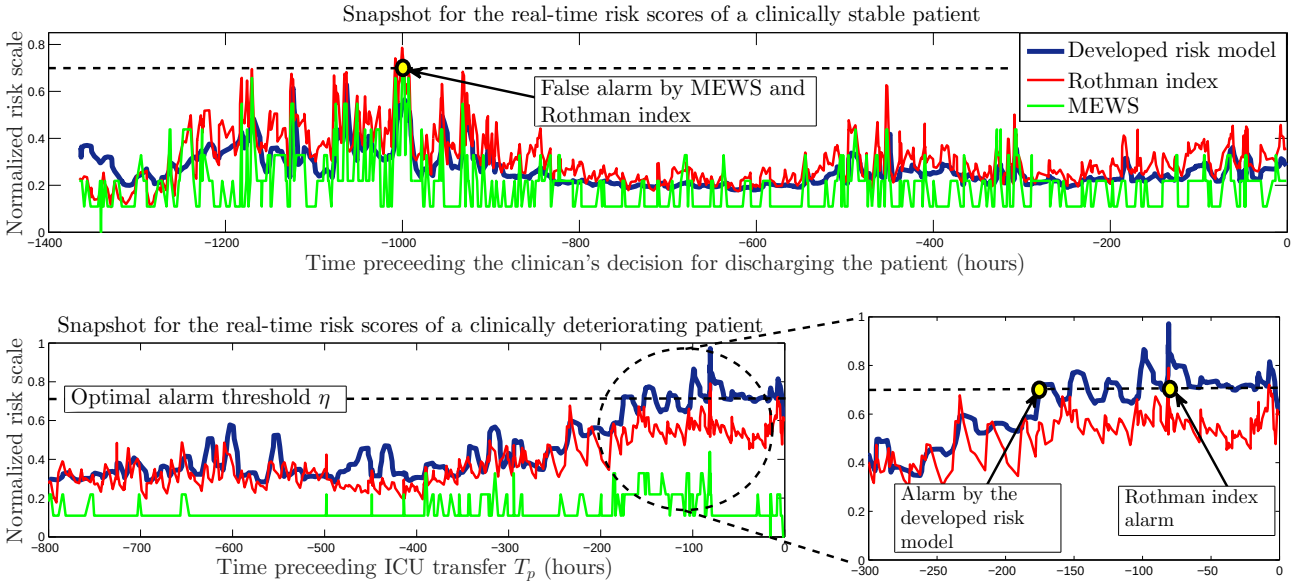

Fig. 3: Snapshots for real-time risk scores computed for two typical patients.

TABLE III: 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%

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 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:

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

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 $\{x_{ij},t_{ij}\}_{i,j}$.

TABLE IV: 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

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:

$$\pi_z^{p+1} = \frac{1}{N_o}\sum_{n=1}^{N_o}\beta_{z,p}^{(n)}$$

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

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 [44] (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 used to update the covariance hyper-parameters as follows:

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

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

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

We compute $\mathbb{P}(\bar{k}|\{x_{ij},t_{ij}\le t\}_{ij},\Gamma_1)$ 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

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

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

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

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

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

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 [45]. 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 4) 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 a state model with 2 latent absorbing states (this corresponds to the model in Figure 4 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 5 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.

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 5, 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 5). 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

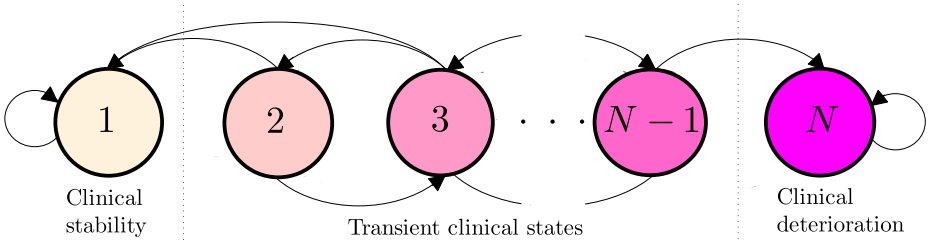

Fig. 4: A latent variable model for the clinical state.

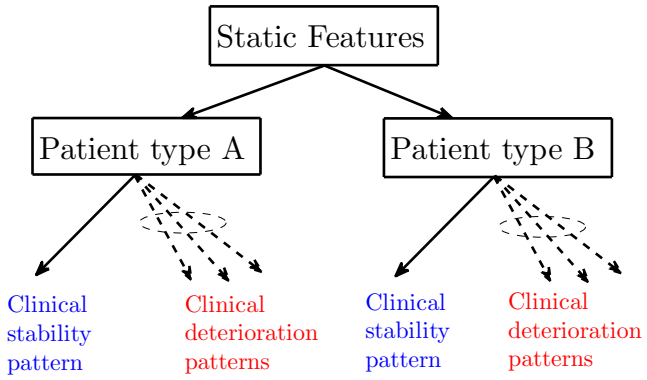

Fig. 5: Configuration of clinical stability and deterioration patterns.

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 IV (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 6 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 7, 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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

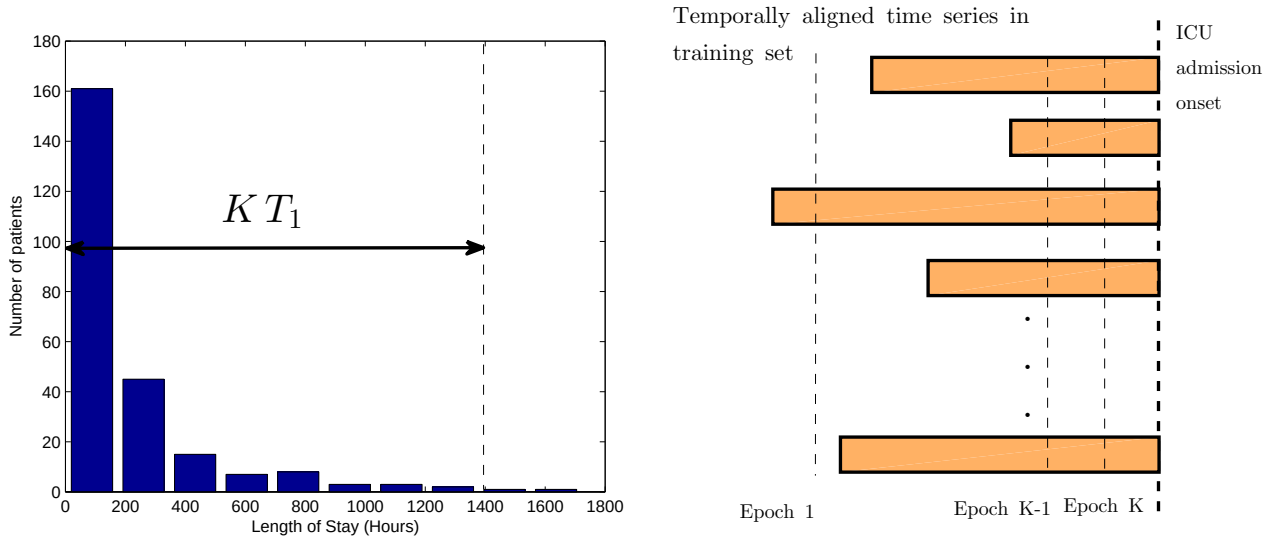

Fig. 6: Interplay between length of stay and training data available for every epoch.

Fig. 7: Deterministic left-to-right transitions.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.