A Personalized Predictive Framework for Multivariate Clinical Time Series via Adaptive Model Selection

Zitao Liu*
Pinterest
651 Brannan St
San Francisco, California 94107
zitaoliu@pinterest.com

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
Pittsburgh, Pennsylvania 15260
milos@cs.pitt.edu

ABSTRACT

Building of an accurate predictive model of clinical time series for a patient is critical for understanding of the patient condition, its dynamics, and optimal patient management. Unfortunately, this process is not straightforward. First, patient-specific variations are typically large and population-based models derived or learned from many different patients are often unable to support accurate predictions for each individual patient. Moreover, time series observed for one patient at any point in time may be too short and insufficient to learn a high-quality patient-specific model just from the patient's own data. To address these problems we propose, develop and experiment with a new adaptive forecasting framework for building multivariate clinical time series models for a patient and for supporting patient-specific predictions. The framework relies on the adaptive model switching approach that at any point in time selects the most promising time series model out of the pool of many possible models, and consequently, combines advantages of the population, patient-specific and short-term individualized predictive models. We demonstrate that the adaptive model switching framework is very promising approach to support personalized time series prediction, and that it is able to outperform predictions based on pure population and patient-specific models, as well as, other patient-specific model adaptation strategies.

KEYWORDS

Personalized medicine; Multivariate time series; Forecasting

1 INTRODUCTION

Personalized medicine is a new and promising approach to healthcare that focuses on an individual patient and patient-specific care. While historically the personalized medicine has emphasized genetics and pharmacology, it is now expanding to include other important aspects of healthcare such as personal preferences, nutrition, lifestyle, and disease, recapturing the importance of personalized health [18, 19, 30]. Along the lines of personalized medicine efforts, our objective is to build personalized clinical time series prediction models that better mimic patient-specific temporal behaviors and variations. More specifically we are interested in developing models and strategies that can predict future values of real-valued multivariate time series (MTS) for a patient given his or her past observations, as well as, time series data for any past patient.

Clinical multivariate time series represent fresh physiological signals collected from patients, which reflect the patients' most recent health condition. Building temporal models of clinical time series is important for modeling the risk of adverse events, or for understanding of the dynamics of a disease and effects of various patient management interventions, especially in clinical settings with a high degree of uncertainty such as intensive care unit where decisions are made frequently [10, 34]. For example, [12] assess patient acuity (or severity of illness) by learning forecasting models from real ICU patients' time series data, which poses immediate practical use for clinicians.

However, making personalized and accurate MTS forecasting is rather challenging due to the characteristics of clinical MTS and the computational and modeling trade-offs arising from them. Briefly, when the time series of past observations for the patient are short, it may be hard to learn a patient-specific model from the patient's own data, and the population-based model may be a better option. On the other hand, learning a population-based model by using all available patients' time series data may bias the forecasting model towards the population mean. Hence, the learned population-based model may not able to reflect the patient-specific future behavior. The most common way to alleviate this problem is to identify a subpopulation of patients most similar to the target patient and learn a model using only examples from this subset. However, such approaches usually rely on some predefined similarity measures to evaluate similarity between the the patient that needs to be predicted and all available training patients. The similarity measures become problematic in time series setting where the atemporal features such as demographic information are very limited (See details in Section 2.1). Moreover, short-term clinical variability and deviations from typical behaviors may prefer the ideal forecasting models that can adapt quickly to just a few recent observations. Overall, the prediction model should provide flexible and customized predictions for each new patient given his or her current health condition, and should benefit from what is known about other patients when the patient-specific model is not available. The majority of existing approaches proposed for clinical MTS prediction in the literature are not able to cover all necessary model behaviors.

We propose and develop an adaptive clinical time series prediction framework that reflects the fact that predictions at different times may be driven by the different types of prediction models. In general, this type of problem is tackled in the machine learning literature by adaptive model selection methods. Briefly, these methods assume a pool of candidate prediction models and each of them is associated with an optimized weight that reflects how much they

*This work was done when the author was at the University of Pittsburgh.

CIKM'17 , Singapore, Singapore
2017. 978-x-xxxx-xxxx-x/YY/MM.
DOI: 10.1145/3132847.3132859

contribute to the prediction solution. The adaptive model selection framework we propose uses the online switching approach [11, 22] that uses a mix of population-based and patient-specific prediction models. The switching is driven by the weighted sum of prediction errors (or deviations) of each model on past patient's data. The weights are set so that more recent errors are more important. The method which makes fewer errors recently is more likely to be selected.

The quality of the adaptive model switching framework ultimately depends on the quality of prediction models included in the pool of time series models and their variety assuring the coverage of many different modes and behaviors. In general one can choose and put any arbitrary model into the pool. However, in this work we narrow our focus to study the trade-offs related to population-based and patient-specific models. This is reflected on the choices of our models. Briefly in addition to simple population and patient-specific baselines we also include and consider more advanced population-based dynamic linear model (DLM), patient-specific DLM, as well as, population and patient-specific versions of two Gaussian process models: one that relies on a set of independent univariate Gaussian process (GP) models (a time series of each clinical variable is modeled by a GP) [29] and a multi-task Gaussian process (MTGP) where entire MTS and interactions among variables are modeled together [3].

We test our framework on clinical laboratory time series data extracted from electronic health records (EHRs). We first show that the population-based models tend to make better predictions when little is known about the current patient instance, and that patient-specific models trained on the patient's own data tend to dominate when observation sequences are sufficiently long. Second, we show that our model switching strategy penalizing more recent prediction errors combines the advantages of all the models and leads to the best personalization strategy for time series forecasting tasks.

2 BACKGROUND

Various methodologies built by machine learning community can be adopted with less or more effort to achieve model personalization. Time series models are among the most challenging ones. In general personalized time series prediction models can be divided into three categories:

- Subpopulation models that build instance-specific models for each instance (Section 2.1).
- Model adaptation methods that adjust the population-based model to fit better the specific instance (Section 2.2).
- Adaptive model selection approaches that instance-dependently combine a pool of predictive models which are built either from the entire population or a subpopulation of instances (Section 2.3).

2.1 Subpopulation Models

The most common way to build a patient-specific model is to identify a subpopulation of patients most similar to the target patient and learn a model using only examples from this subset. We call these models *Subpopulation Models*. The subpopulation approaches usually rely on some pre-defined similarity measures to evaluate similarity between the target example (the patient that needs to be predicted) and all training examples (all available past patients), that is, a past patient is used to build a model for the target patient only if it is highly similar to the target patient.

The main challenge when adopting the subpopulation approach is to define proper similarity among patients and their respective time series. The majority of approaches in the literature assume the similarity among patients relies on some atemporal patient-specific information (such as demographics of the patient) to guide the personalized strategies. Deriving the similarity of two time series or mixed atemporal and temporal information is more complex. To measure the similarity of time series sequences of equal length, Euclidean distance, Pearson correlation, cosine distance and their variants are typically used [15]. For time series of different lengths, the similarity can be computed either explicitly by using dynamic time warping [1, 28] or implicitly by using the likelihood of generative probabilistic models defining the time series [16, 21].

However, there are several drawbacks that prevent subpopulation based approaches from providing accurate personalized forecasting in real-world clinical domain. First, it is difficult to find a patient-specific sequence's similar "neighbors" in the training set initially when the sequence is very short. Second, seeking appropriate neighbors of each instance becomes computationally expensive when the instance changes over time. In time series domain, this intensive neighbor searching process has to be redone once new observations arrive. Third, a subpopulation from which we start and learn a subpopulation model from may still be very large and exhibit a lots of patient-specific variations. So it may be necessary to further explore methods that can adapt the prediction model closer to the current patient.

2.2 Model Adaptation

Model adaptation methods try to bridge a possible gap in between population (or subpopulation models) and the target patient by adjusting the population model to fit better the specific patient. This usually includes two steps: first learn a population-based model from all available data and then calibrate the population-based model according on the unique characteristics of each instance. Broadly speaking, there are two types of mechanisms and strategies to modify the population-based model to reflect the instance-specific characteristics, *model parameter adaptation* and *instance-specific residual modeling*.

Model parameter adaptation approaches achieve the personalized prediction results by modifying the model parameters of population-based models based on instance-specific features. For example, Berzuini et al. [2] propose a general Bayesian network model for individualized therapeutic monitoring. Different from model parameter adaptation approaches, instance-specific residual based techniques add additional models to support the personalized predictive outcomes. In such approaches, residuals are defined as the difference between the true outcomes of the specific instance and the predictive results of the population-based models. For example, Liu and Hauskrecht [25] achieve better adaptive forecasting performance by learning extra prediction models learned

from the patient-specific residual time series, which is the difference between the patient observations and the predictions from the population-based model.

In spite of the successful applications of the model adaptation techniques, they have some limitations. For model parameter adaptation approaches, designing and deriving adaptation is very difficult and varies from model to model. Even under Bayesian adaptation framework, issues of how to design the model and parameter priors to achieve fast adaptation remain open. Furthermore, both model parameter adaptation and instance-specific residual modeling usually require more instance-specific features or observations to perform sufficient contributions. However, time series observed for one patient are often too short to support adequate adaptation or residual modeling.

2.3 Adaptive Model Selection

The methodology we pursue in this work solves the personalization problem via adaptive model selection approach. The adaptive model selection approach assumes a pool of candidate predictive models that may contribute to the prediction. A different model or combination of models may support the predictions at the different time. Briefly, each of the candidate models is associated with weight that reflects how much it contributes to the final solution. Two different strategies: ensemble (Section 2.3.1) and online (Section 2.3.2) methods are commonly be used to choose (optimize) the weights in the machine learning literature.

2.3.1 Ensemble Methods. Ensemble methods are general techniques in machine learning for combining several models to create a more accurate prediction [5]. Related research work focuses on either creating more candidate models, such as bagging [4], boosting [11] or by wisely optimizing their combination weights, such as exponential weighting, stacking [33], etc. In medical and clinical practice, the ensemble methods can often significantly boost the performance of individual models. For examples, Jiang et al. [17] develop a data-driven approach to utilize individualized confidence intervals to select the most "appropriate" model from a pool of candidates to predict patient's specific clinical condition. Visweswaran and Cooper [35] perform a selective Bayesian model averaging for each individual patient where the prediction is made by first searching for models having the greatest impact on the target prediction and then averaging the predictions from selected models.

2.3.2 Online Algorithms. In online prediction problems, various techniques, such as the weighted majority algorithm [22], hedge algorithm [11] are proposed to select the best model from the candidate pool based on the knowledge of the past. The models with poorer performance receive larger penalties and become less likely to be picked in the future. There have been many papers that aim to apply online learning to solving real-world problems, for example, classifying handwritten digits [8], detects malicious Web sites [27], but as far as we know no work has been applied to time series forecasting in clinical settings.

Similarly to approaches in Section 2.1 and Section 2.2, Adaptive model selection techniques require a large number of observations to optimize the weights to combine/select models. This becomes become unrealistic and inapplicable when a new patient comes in and very few observations are known for that patient in the real clinical setting. Moreover, methods in adaptive model selection tend to treat all the past errors equally and optimize the weights to combine/select models to achieve best performance on average. However, due to rapid changes in clinical time series and standard ensemble and online algorithms fail to penalize the recent errors more in clinical time series forecasting.

Please note that models from the above three categories are complementary and they can be combined in the prediction process. For example, the model adaptation techniques can be applied to both population-based models and subpopulation models. Moreover, both subpopulation models and adaptive models can be candidate models in the pool of the adaptive model selection approaches. In the following, we briefly review the three approaches to build the personalized model.

3 METHODOLOGY

In this work, we develop a model switching framework that adaptively switches among many different time series models that may represent population, patient-specific trends or even short-term variabilities in the clinical time series. We build a pool of time series models with different characteristics that can be used in different stages of the time series prediction problem and when the sequences of patient's own past observations vary in length. In the following, we first introduce and formalize the prediction problem we want to solve. After that we describe a pool of widely used time series models and how we apply them to the collection of clinical MTS to build population and patient-specific forecasting models. Then we describe the online switching framework that combines these models with the aim of improving the overall prediction performance. Finally, we discuss the practical advantages of our model switching strategy.

3.1 The Prediction Problem

In this work, our objective is to predict the values of future clinical observations for the *target patient* given his/her past clinical data. More specifically, we assume that we have observed a sequence of q past observation-time pairs $(\mathbf{y}_i, t_i)_{i=1}^{q}$, such that, q is the number of past observations, $0 < t_i < t_{i+1}$, and $\mathbf{y}_i$ is a n-dimensional observation vector made at time (t_i). Time t^*, $t^* > t_q$, is the time at which we would like to predict the future observation $\hat{\mathbf{y}}_{t^*}$. Furthermore, in order to obtain a fine-grained experimental analysis, we denote a prediction task Γ^p_{j,t^*} as the smallest prediction unit in our problem, which is the estimation of patient p's jth clinical variable at time t^*.

3.2 Time Series Models

Our framework works by combining multiple different time series models and their strength to improve the prediction. Various time series models with the different assumptions exist [13] and may be considered. In this work we power our model switching framework with two widely used time series models - dynamic linear model and Gaussian process models, and develop robust population-based and patient-specific versions of these models and algorithms for learning them from data. The robustness assures the models can

applied to cases when the number of time series examples is small or the length of the individual time series is short-span.

3.2.1 Dynamic Linear Model. The dynamic linear model (DLM) model is a time series model used frequently for time series prediction. The DLM models real-valued MTS $\{\mathbf{y}_t \in \mathbb{R}^n\}_{t=1}^T$ using hidden states $\{\mathbf{z}_t \in \mathbb{R}^d\}_{t=1}^T$:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t \tag{1}$$

$$\mathbf{y}_t = C\mathbf{z}_t + \boldsymbol{\zeta}_t \tag{2}$$

Briefly, $\{\mathbf{z}_t\}$ is generated via the transition matrix $A \in \mathbb{R}^{d\times d}$. Observations $\{\mathbf{y}_t\}$ are generated from $\mathbf{z}_t$ via the emission matrix $C \in R^{n\times d}$ (see eq.(1) and eq.(2)). $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$ and $\{\boldsymbol{\zeta}_t\}_{t=1}^T$ are i.i.d. multivariate normal distributions with mean $\mathbf{0}$ and covariance matrices Q and R respectively. T is the length of the time series. In addition to A, C, Q, R, the DLM is defined by the initial state distribution for $\mathbf{z}_1$ with mean $\boldsymbol{\xi}$ and covariance matrix Ψ, i.e., $\mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{\xi}, \Psi)$. The complete set of the DLM parameters is $\Lambda = \{A, C, Q, R, \boldsymbol{\xi}, \Psi\}$. DLM represents the dynamics indirectly using hidden states which gives one additional flexibility to better capture the different modes the system may exhibit and is more robust when observations are noisy.

3.2.2 Regularized Dynamic Linear Model. One of the limitations of the DLM model is that we do not know a priori the dimensionality of its hidden state space. From eq.(1), we can see that the number of parameters representing transitions among hidden state components (a.k.a, transition matrix) is quadratic in the dimensionality of the hidden space. An inappropriate choice of hidden state dimension (d) may easily lead to either overfitting or underfitting problems. Overfitting may occur if the dimension of the hidden state space picked is large and when the training set used to learn the model is small. On the other hand, underfitting occurs when the hidden state space size is unnecessarily small, which limits model's expressibility and accuracy.

To avoid these issues, Liu and Hauskrecht [24] proposed an extension of the DLM, called regularized dynamic linear model (rDLM), which starts from a higher dimensional hidden state space, but is able to automatically adjust its dimensionality to prevent overfitting. The rDLM aims to find the optimal dimensionality of the hidden state space by regularizing the rank of the transition matrix (A) and hence, shuts down spurious and unnecessary dimensions of the DLM. In order to encourage a low-rank transition matrix of rDLM, a nuclear norm prior is assigned to the entire transition matrix, i.e., $p(A) \propto \exp(-\lambda\|A\|_*)$, and the EM algorithm is performed on the augmented Q function.

3.2.3 Population-based and Patient-specific DLM. In general, an DLM model can be learned either from a collection of many MTS sequences or from an individual sequence, which leads to either population-based models or patient-specific models. The population-based DLM model is learned from all available data sequences of patients and hence it summarizes the dynamics of all patients in the population. We expect a population-based model to be especially useful in the early stages of clinical predictions because at the beginning, observations of clinical variables for an individual patient are too short and insufficient to learn a high quality patient-specific model solely based on patient's own data.

However, because of the averaging effects of many patient data sequences the model is trained on, the population-based model usually fails to capture patient-specific variability. Since the prediction task is performed patient by patient, an ideal forecasting model should reflect and take into account the variations specific to the current patient. Furthermore, a patient may exhibit short-term variability reflecting the different events affecting the care and patient state [31]. Since the individual-specific model is trained on each sequence, the model is better at capturing the patient-specific variability and providing better customized predictions than population-based models.

We note that DLM based models belong to discrete time models which require that the time intervals between any two consecutive observations are same. When dealing with irregularly sampled time series, time series discretization techniques can be used as a data preprocessing step before learning the models [23].

3.2.4 Gaussian Process Models. The Gaussian process (GP) is a popular nonparametric nonlinear Bayesian model and is widely used in time series regression and forecasting tasks, where time stamps are modeled as the input of GP and observations are modeled through the predicted mean function of the time series [29]. In time series modeling, each GP is used to model an individual time series, which is represented by the mean function $m(t) = \mathbb{E}[f(t)]$ and the covariance function $K^G(t, t') = \mathbb{E}[(f(t) - m(t))(f(t') - m(t'))]$, where $f(t)$ is a real-valued process and t and t' are two time stamps. The GP can be used to calculate the posterior distribution $p(f(t^*)|(\mathbf{y}_i, t_i))$ of f values for an arbitrary time stamp t^*, given a set of observation-time pairs $(\mathbf{y}_i, t_i)$.

The advantages of GP based models is that (1) with the reasonable choice of the covariance function, GP based models are capable of capturing the short-term rapid changes in clinical time series [7, 12]; and (2) GP based models can be applied to time series modeling problem by representing observations as a function of time. As a result, there is no restriction on when the observations are made and whether they are regularly or irregularly spaced in time.

3.2.5 Multi-task Gaussian Process. One limitation of the basic GP model is that each clinical time series in the MTS must be modeled independently and the interactions among multiple clinical variables are ignored. The multivariate behaviors within the clinical MTS can be however captured by the multi-task Gaussian process (MTGP) [3]. The MTGP is an extension of GP to model multiple tasks (e.g., multivariate time series) simultaneously by utilizing the learned covariance between related tasks. MTGP uses K^C to model the similarities between tasks and uses K^G to capture the temporal dependence with respect to time stamps. The covariance function of the MTGP is shown as follows:

$$K^M = K^C \otimes K^G + D \otimes I_T \tag{3}$$

where K^C is a positive semi-definite matrix and $K^C_{j,k}$ measures the similarity between time series j and time series k. D is an $n \times n$ diagonal matrix in which $D_{j,j}$ is the noise variance δ_j^2 for the jth time series. $\otimes$ is the Kronecker product. Usually the MTGP model has the computation limitation that it has $O(n^3T^3)$ compared with $n \times O(T^3)$ for standard GP models. However, this limitation is not as relevant in our application setting, given that the number of

clinical observations is very limited and clinical time series are usually short span.

3.2.6 Population-based and Patient-specific GP and MTGP. Similarly to learning of the DLM based models, GP (or MTGP) models can be either learned from the collection of many time series or from an individual patient-specific sequence. Learning of the patient-specific model from each sequence is straightforward. For each target patient, a patient-specific model is learned solely from the patient's past q observation-time pairs $(\mathbf{y}_i, t_i)_{i=1}^{q}$. More specifically, we treat each clinical time series in $(\mathbf{y}_i, t_i)_{i=1}^{q}$ independently and learn a patient-specific GP model for each clinical variable. If we take into account of the correlations and interactions among clinical variables and learn a patient-specific MTGP model from $(\mathbf{y}_i, t_i)_{i=1}^{q}$. Both the GP and MTGP models has zero mean and a squared exponential covariance function (eq.(4)), which is the most frequently-used example in literature [29].

$$K^G(t_i, t_{i'}) = \alpha \exp\Big(- \frac{(t_i - t_{i'})^2}{2\beta^2} \Big) \tag{4}$$

In our work, we adopt the Cholesky decomposition and the "free-form" parameterization techniques ($K^C = LL^\top$) to learn the parameter set Λ by minimizing the negative log marginal likelihood via gradient descent [12, 29].

To learn the population-based models from a collection of MTS sequences, we learn the GP based models from each sequence in the training collection first and use the average of all the learned parameters as our estimates of the population-based models. We note that another possible approach to learn the population GPs (or MTGP) is to concatenate sequences for multiple patients into one long sequence and by separating the two consecutive patient sequences with a long time gap.

Both GP and MTGP are used in clinical time series domain to capture the short-term and long-term variability [7, 9, 12, 20, 31]. In the work by Clifton et al. [7], Lasko, Denny, and Levy [20], and Schulam, Wigley, and Saria [31], each clinical time series is modeled by a single GP separately which does not allow one to represent dependences among the different time series. Ghassemi et al. [12] and Durichen et al. [9] try to capture MTS and dependences among its time series by applying MTGP to clinical MTS modeling and forecasting. Since all above applications focus on individual-specific sequence, they tend to support more accurate and personalized time series predictions for each patient compared to population-based models. However, those models usually require long enough sequences to optimize the models' parameters. This becomes unrealistic and inapplicable when a new patient comes in and very few observations are known for that patient.

3.3 Online Model Switching

Due to the rapid changes in the clinical time series, it is difficult to develop a single model that consistently performs well over the time for each individual. Therefore, in this work, we make the prediction for patient p at time t^* from a pool of candidate models, which contains both the population-based model (DLM based and GP based) and patient-specific models (DLM based and GP based). Our objective is to develop a framework that is able to pick the best model from the pool to timely support accurate and personalized clinical predictions for each patient at every time stamp.

Although numerous ensemble and online methods exist, the majority of the methods require error feedback over longer periods of time to achieve any statistical guarantee of total errors made by the algorithms. However, in the real-world clinical setting, patients' time series are usually too short to obtain effective weights for both the ensemble and online algorithms. Furthermore, weight updating rules are often based on the overall performance of each model on all previously observed data and hence the recent errors are smoothed out by the errors made in the early stage of the process. Since clinical MTS may contain short-term variability (caused, for example, by acute infections, bleeding, surgeries, etc) standard weight updating rules are not able to respond to these changes quickly enough.

In this work, we propose and develop a novel online model switching strategy, i.e., "weighted Follow-the-Leader" (wFTL), to address the above problem. Different from traditional online learning algorithms that treat each past errors equally, we put more penalties on recent errors. The intuition lies in that the predictive models do not perform well initially can catch up the performance soon and they need to be selected as soon as possible given the short-span characteristics in clinical data. More precisely, for each model $\mathcal{M}_m$, all its past errors can be computed (up to current time stamp t_q) as $\mathbf{e}_m = [e_1^m, \cdots e_i^m, \cdots, e_q^m]$. The model being pick at time t^* is selected by

$$\mathcal{M}_* = \arg\min_m \sum_{i=1}^{q} w_i * e_i^m \tag{5}$$

where w_i is the error weight at time t_i.

In order to capture the recency effect, we compute the error weight by using the kernel functions which takes time stamps as inputs. The idea is that the errors made far aways should be less penalized compared to the most recent errors. In this work, we experiment with two standard kernel functions, i.e., the square exponential kernel (eq.(6)) and the mean reverting kernel (eq.(7)).

$$K_{se}(t_i, t^*) = \exp\Big(- \frac{(t_i - t^*)^2}{\gamma} \Big) \tag{6}$$

$$K_{mr}(t_i, t^*) = \exp\Big(- \frac{|t_i - t^*|}{\gamma} \Big) \tag{7}$$

where t^* is the time stamp of the target prediction. t_i is the all the past time stamps, $i = 1, 2, \cdots q$ and γ is the bandwidth parameter.

As we can see from eq.(5), the proposed approach downgrades to the "Follow-the-Leader" (FTL) strategy when all the weights (w_is) become 1 [32]. The FTL strategy simply selects the best prediction model by integrating the loss across past t steps and neglects the recency effect. While wFTL always selects the prediction model with the minimum weighed loss over time. As a result, it is more sensitive to the recent observations that reflect the most current trend and change of the state of the target patient. By evaluating the candidate models' predictions and focusing on the recent performance, the proposed strategy is able to discover sudden changes and quickly switch to the best model. Compared with eq.(6) and eq.(7), the square exponential kernel squares the time difference which vanishes the past errors much quicker than mean reverting kernel. The hyper parameter γ controls the magnitude of the recency effect. wFTL with either eq.(6) or eq.(7) becomes FTL when γ goes to infinity.

3.4 Model Switching versus Model Averaging

Although model switching approaches can be viewed as a special case of model averaging where only the selected model has the non-zero weight, it has several advantages compared to model averaging in practice. First, model switching based prediction systems usually have higher development velocity and lower maintenance cost. It decouples the entire prediction system and adding or deleting candidate from the pool is more easier compared to model averaging based prediction system. Any failure of a candidate model won't influence the prediction pipeline in terms of system stability. Second, model switching based prediction system is more robust to poor-quality candidates in the pool. Even though the combining weights can be designed to penalize more to the worse candidates, it is difficult to completely vanish the prediction errors.

4 EXPERIMENTAL EVALUATION

In this section, we evaluate our approach on a real-world clinical time series dataset derived from the Complete Blood Count panel. We conduct a series of experiments to explore and demonstrate the benefits of our adaptive model switching framework. First, we study the quality of population-based versus patient-specific models for observations histories of the different length. Second, we focus on the MTS forecasting and the evaluation of the proposed model switching approach to other models.

4.1 Clinical Data

We test our adaptive model switching framework on a clinical MTS data obtained from EHRs of post-surgical cardiac patients [14, 26]. We take 500 patients from the database who had their *Complete Blood Count* (CBC) tests [1] done during their hospitalization. The MTS data consists of six individual CBC lab time series: mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), mean platelet volume (MPV), red blood cell (RBC) and red cell distribution width (RDW). In the following experiments, we have randomly selected 100 patients out of 500 as a test set and used the remaining 400 patients for training the models.

4.2 Evaluation Metrics

We evaluate and compare the performance of the different methods by calculating the average Mean Absolute Percentage Error (Average-MAPE) of models' predictions. Average-MAPE measures the prediction deviation proportion in terms of true values:

$$\text{Average-MAPE} = \frac{\sum_{l=1}^{N}\sum_{j=1}^{n}\sum_{i=1}^{T_l}|1-\hat{y}_{j,i}^{l}/y_{j,i}^{l}|}{n\sum_{l=1}^{N}T_l}\times 100\%$$

where $|\cdot|$ denotes the absolute value; $y_{j,i}^{l}$ and $\hat{y}_{j,i}^{l}$ are true and predicted values from time series j at time t_i for patient l.

Usually in MTS data, different time series are in different scales and simply averaging the error values themselves is not appropriate. Average-MAPE measures the prediction deviation proportion in terms of the true values, which is more sensible than computing the mean of root mean square errors (RMSE), mean square errors (MSE) or mean absolute errors (MAE) of each time series' predictions.

[1] CBC panel is used as a broad screening test to check for such disorders as anemia, infection, and other diseases.

4.3 Time Series Models

During our evaluations we consider a variety of time series prediction models used commonly in both clinical pharmacology and machine learning and their population-based and patient-specific versions. All these can be put into the pool of candidate models into the pool of methods our framework uses. For the population-based models, we choose (1) *P_Mean*: mean of the entire population; (2) *P_rDLM*: a regularized DLM learned from other patient data $\mathcal{D}$; (3) *P_GP*: learning a population GP model from $\mathcal{D}$; and (4) *P_MTGP*: learning a population MTGP model from $\mathcal{D}$. For patient-specific models, we choose (1) *I_Mean*: Mean value for the individual patient up to the current time stamp; (2) *I_rDLM*: learning a rDLM model from the MTS sequence of the target patient; (3) *I_GP*: Gaussian process regression model for each individual time series of the target patient; and (4) *I_MTGP*: multi-task Gaussian process model for the MTS sequence of the target patient.

4.4 Baselines

In the following experiments, we denote the wFTL with the square exponential kernel (eq.(6)) as wFTL_se and denote the wFTL with the mean reverting kernel (eq.(7)) as wFTL_mr. We set γ in eq.(6) and eq.(7) by the internal cross validation approach while optimizing modelsfi predictive performances. We compare our wFTL_se and wFTL_mr model switching strategies to other approaches one can use for personalized predictive modeling.

- *Sub*: represents a subpopulation approach. For each patient at each time stamp, top k similar patients are selected and are used to train the rDLM model. The similarity is defined by the Euclidean distance between the sample means of clinical variables of the target patients and the sample means of available training patients. In this experiments, we vary k to 50, 100 and ALL where ALL means all the training examples.
- *rDLM+reGP*: is a model adaptation approach [25]. In rDLM + reGP, a population rDLM model is trained first and the time series of past observations for the target patient is expressed in terms of residuals (or differences in between predictions made by the population-based model and actually observed values). Then each of the residual time series is modeled by a GP.
- *rDLM+reMTGP*: is another model adaptation method that is similar to rDLM+reGP but the all residual time series are modeled by an MTGP [25].
- *En_Avg*: is a simple averaging method in which the prediction is made by uniformly averaging the results from all the models in the pool.
- *En_Err*: is the inverse-error weighted average method. Assuming M be the number of models in the pool. Let e_m be the sum of prediction errors of the model m over the past t time steps (rounds) and w_m be the mixture weight corresponding to model m. In En_Err, w_m is computed as $w_m = \frac{1}{e_m S}$ where $S = \sum_{m=1}^{M}\frac{1}{e_m}$.

- *OL_FTL*: Follow-the-Leader method that selects the best model based on the loss integrated over past t time stamps.
- *OL_MW*: Multiplicative weights algorithm [6] that at each round t, makes the selection is based on the probability distribution $\mathbf{p} = \{w_1/\Phi, \cdots, w_M/\Phi\}$, where $\Phi = \sum_{m=1}^{M} w_m$. w_m is updated by penalizing the costly predictions, i.e., $w_m^+ = w_m(1 - \eta e_m)$ where $\eta, \eta \leq 0.5$ is the discounting factor.
- *OL_Hedge*: Hedge algorithm [11] that is similar to OL_MW but uses an exponential factor instead of a linear cost $(1 - \eta e_m)$. The weight update is $w_m^+ = w_m \exp(-\eta e_m)$.

4.5 Population-based versus Patient-specific Models

We first explore the prediction performance of each model in the prediction model pool individually. Instead of averaging all the prediction results, we compute the Average-MAPE results of population-based, patient-specific methods and our proposed wFTL model switching approaches (wFTL_se and wFTL_mr) when they start to predict with a delay corresponding to the different number of initial observations (initial observation sequence length). For example, when the initial observation sequence length is set to 4 the Average-MAPE reflects the errors of all one-step-ahead predictions the method makes when starting from four or more initial observations for the target patient (that is, the model starts to make predictions from the 5th time stamp). The Average-MAPE results with different initial observation lengths are shown in Figure 1. Due to the poor performance of the P_Mean and I_Mean methods, we don't visualize them in Figure 1.

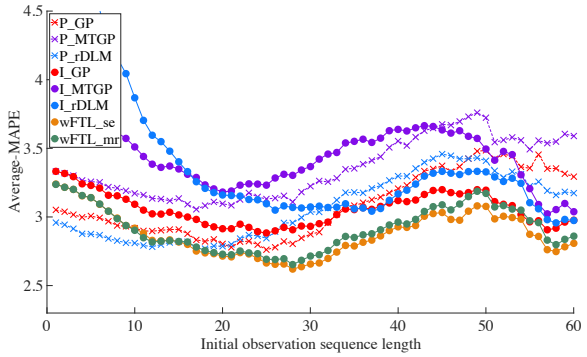

Figure 1: Average-MAPE results of all models in the pool and two wFTL methods for the different initial observation lengths.

First, Figure 1 shows the trade-off between the population-based (P_rDLM, P_GP and P_MTGP) and the patient-specific (I_rDLM, I_GP and I_MTGP) models. Briefly, the performance of patient-specific models built from patient's own past observations tends to gradually improve and eventually outperforms the population-based models that are the best initially when little is known about the target patient. More specifically we observe that, P_rDLM model built on the population of past patients starts strong but deteriorates when more values are observed. We explain this deterioration by the fact that longer the patients stay in the hospital the more likely they deviate from the population-based models. This is also reflected by the deterioration of the population-based GP models (P_GP and P_MTGP) for longer observation sequences. On the other hand, we observe that patient-specific models can adapt to the specifics of the patient but they also take a longer time (number of observations) to learn, especially when the model is more complex. While I_GP is relatively fast to adapt to the specifics and short-variability of the target patient, I_MTGP is slower because of increased model complexity and more parameters it needs to learn. In addition, from Figure 1, we can see that different models have various prediction performance when the number of observations change, which confirms the motivation of dynamically switching to the most appropriate model during the prediction. By using the different kernel functions (eq.(6) and eq.(7)), our wFTL strategies penalize the most recent errors made by each candidate model. As shown in Figure 1, the proposed wFTL approaches are slightly worse compared to P_rDLM initially. But they catch up the performance of P_rDLM rapidly and consistently have the best performance among all the population-based and patient-specific models when enough initial observations are obtained.

4.6 Prediction Accuracy

In this experiment, we compute and compare one-step-ahead prediction accuracy of wFTL to various state-of-the-art personalization approaches. We present the prediction results against baselines in different categories separately to make the differences clear. The results are shown in Figures 2 - 4. To evaluate the statistical significance of performance difference, we apply paired t-tests at 0.05 significance level.

4.6.1 Comparison of results for ensemble and online methods. As we can see from Figure 2, when initial observation sequence length is short, our wFTL strategies perform slightly worse than the inverse-error weighted average method (En_Err). But in the long run, our wFTL strategies have the best performance among all the other adaptive model selection based baselines. Clinical time series contain lots of short-term variability due to different causes [31]. For example, the blood tests may be affected by events like infection, bleeding, transfusion, or a particular medication treatment. patient-specific models can adapt better to this variability while population-based models tend to average the variability out (treat them as a noise) so they likely do not perform well when these "exceptions" occur. Since wFTL strategies not only consider the past errors but also focus on the most recent performance of each predictor, they are able to quickly adapt to the short-term variability and rapid changes. On the contrary, the standard adaptive model selection approaches (ensemble methods and online algorithms) are all based on weighting schema extracted from the entire history. These historical observations are too long and may prevent us from adapting to these short-term variability. Furthermore, in order to change the prediction behavior of these methods, the weights must be changed. Since there are many different weights it may take a long time for them to be adapted. This is also reflected by the improvement of the online learning approaches (OL_MW and OL_Hedge) for longer observation sequences. Also we find from the statistical significance test results (due to the space limit,

we don't list the full statistical significance test results here), both OL_MW and OL_Hedge have the comparable performance to our wFTL strategies only when the initial observation length is large than 49.

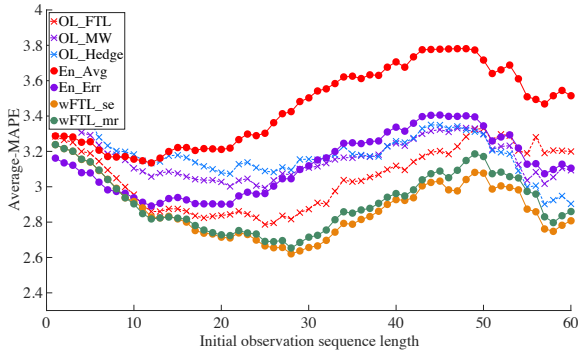

Figure 2: Average-MAPE results of the proposed wFTL approaches compared to the ensemble and online methods.

4.6.2 Comparison of results for subpopulation methods. Figure 3 compares the prediction performance of our wFTL strategies and subpopulation methods. Similarly to the prediction results in Figure 2, the subpopulation methods achieve better performance when initial observation sequence length is less than 12. This is because patients start to differentiate and exhibit their unique symptoms as their hospitalizations go by. For subpopulation methods, it is difficult to accurately find and represent the target patient's short-term changes by solely using the static examples from the training set. The top k similar subset might not be able to reflect the most recent temporal behavior of the target patient. Moreover, from Figure 3 and the statistical significance test results (due to the space limit, we don't list the full statistical significance test results here), we can see that the performance of subpopulation methods vary with different values of k. Choosing the optimal value of k is an challenging issue. In subpopulation methods, the top k subset is specific to each patient and it is re-constructed when every new observation is obtained for that specific patient, repeatedly searching for the best subset. As a result, the training of the subpopulation model becomes very time consuming. It is not practical to apply such methods in a large scale EHR data set.

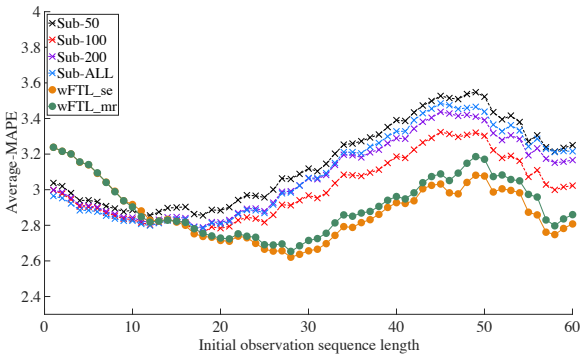

Figure 3: Average-MAPE results of the proposed wFTL approaches compared to the subpopulation methods.

4.6.3 Comparison of results for model adaptation approaches. We also compare our online model switching strategies (wFTL_mr and wFTL_se) with the residual based model adaptation techniques (rDLM+reGP and rDLM+reMTGP) and the results are shown in Figure 4. As we can see, our wFTL_mr and wFTL_se switching strategies have comparable performance to model adaptation techniques although they are slightly worse numerically than rDLM+reGP and rDLM+reMTGP. We run the pairwise t-test for each possible pair from these two categories of methods and none of them are statistically significantly different at 0.05 level (shown in Table 1). *Please note that even though the two approaches have similar performance, they are different by nature*: the wFTL strategies keep selecting the best predictor from a pool of candidate models based on the weighted average of past errors while the residual based model adaptation techniques rely on learning from patient specific residuals to capture the short-term variability in patient dynamics. Both rDLM+reGP and rDLM+reMTGP models have worse performance at the beginning is because they require enough residuals to fit the parameters of the GP or MTGP models.

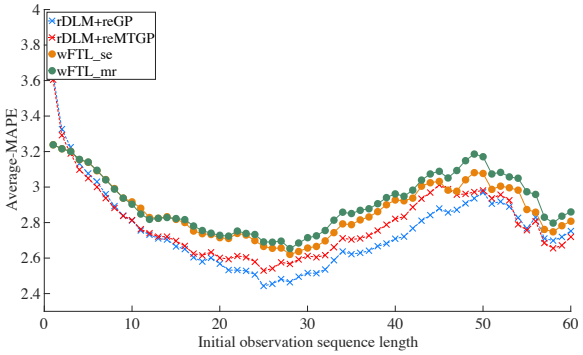

Figure 4: Average-MAPE results of the proposed wFTL approaches compared to the model adaptation based methods.

We also note that models from the subpopulation methods and model adaptation approaches are complementary and they can be combined in the prediction process. For example, the model adaptation techniques can be applied to both population-based models and subpopulation models. Moreover, both subpopulation models and adaptive models can be candidate models in the pool, which can be used by our online model switching strategies.

5 CONCLUSION

In this paper, we presented a new adaptive model switching framework for clinical time series forecasting. Compared to the traditional prediction models which rely on either population-based models or individual-specific models, our strategy benefits from both and at the same time it is able to quickly adapt to patient-specific variability. Experimental results on a real-world clinical data demonstrated that our approach outperforms other state-of-the-art prediction approaches in terms of Average-MAPE. In the future, we plan to explore (1) event-specific models of short-term variability; and (2) online strategies that would allow us to add or delete models dynamically from the candidate pool.

Table 1: Average-MAPE results (means and standard errors) of the proposed wFTL approaches compared to the model adaptation based personalization approaches for the different initial observation sequence lengths. The best performing method is shown in bold. Also in bold are the methods that are not statistically significantly different from the best method at 0.05 significance level.

Initial Length	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
reGP	3.64±0.06	3.33±0.05	3.22±0.05	3.13±0.05	**3.08±0.06**	**3.03±0.06**	**2.96±0.06**	**2.89±0.06**	**2.84±0.06**	**2.81±0.06**	**2.76±0.06**	**2.73±0.06**	**2.71±0.06**	**2.70±0.07**	**2.67±0.07**
reMTGP	3.60±0.06	3.29±0.05	**3.19±0.05**	**3.10±0.05**	**3.05±0.05**	**3.00±0.06**	**2.94±0.06**	**2.88±0.06**	**2.84±0.06**	**2.81±0.06**	**2.76±0.06**	**2.74±0.06**	**2.72±0.06**	**2.72±0.07**	**2.70±0.07**
wFTL_se	**3.24±0.05**	**3.22±0.05**	**3.20±0.05**	**3.15±0.06**	3.14±0.06	3.09±0.06	3.04±0.05	2.99±0.06	2.94±0.05	2.92±0.06	2.88±0.06	2.83±0.06	2.82±0.06	2.83±0.07	2.82±0.07
wFTL_mr	**3.24±0.05**	**3.22±0.05**	**3.20±0.05**	**3.16±0.06**	3.14±0.06	3.09±0.05	3.04±0.05	2.99±0.05	2.94±0.05	2.90±0.05	2.85±0.05	2.82±0.06	2.82±0.06	2.83±0.06	2.82±0.07
Initial Length	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
reGP	**2.65±0.07**	**2.60±0.08**	**2.58±0.08**	**2.60±0.08**	**2.57±0.09**	**2.53±0.09**	**2.53±0.09**	**2.53±0.10**	**2.51±0.09**	**2.44±0.09**	**2.46±0.09**	**2.48±0.10**	**2.46±0.10**	**2.49±0.11**	**2.52±0.11**
reMTGP	**2.67±0.07**	**2.63±0.07**	**2.62±0.08**	**2.63±0.08**	**2.60±0.08**	2.59±0.09	2.61±0.09	2.60±0.10	2.58±0.10	2.53±0.10	2.54±0.10	2.58±0.11	2.57±0.11	2.59±0.12	2.61±0.12
wFTL_se	2.80±0.07	2.75±0.08	2.74±0.08	2.73±0.08	2.71±0.09	2.71±0.09	2.74±0.10	2.73±0.10	2.70±0.10	2.66±0.09	2.65±0.10	2.66±0.10	2.62±0.11	**2.64±0.11**	**2.66±0.12**
wFTL_mr	2.82±0.07	2.78±0.07	2.75±0.08	2.74±0.08	2.73±0.08	2.72±0.09	2.75±0.09	2.74±0.09	2.73±0.09	2.69±0.09	2.69±0.10	2.70±0.10	2.65±0.10	2.68±0.11	2.71±0.12
Initial Length	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
reGP	**2.51±0.12**	**2.54±0.12**	**2.59±0.13**	**2.64±0.14**	**2.62±0.14**	**2.63±0.14**	**2.64±0.14**	**2.67±0.15**	**2.68±0.15**	**2.71±0.16**	**2.72±0.16**	**2.77±0.17**	**2.81±0.17**	**2.84±0.17**	**2.88±0.18**
reMTGP	**2.60±0.13**	**2.62±0.13**	**2.66±0.14**	**2.71±0.15**	**2.70±0.15**	**2.71±0.16**	**2.73±0.16**	**2.75±0.17**	**2.79±0.17**	**2.82±0.18**	**2.83±0.18**	**2.89±0.19**	**2.93±0.19**	**2.97±0.20**	**3.01±0.20**
wFTL_se	**2.67±0.12**	**2.70±0.13**	**2.74±0.14**	**2.79±0.14**	**2.79±0.15**	**2.81±0.15**	**2.83±0.15**	**2.86±0.16**	**2.90±0.16**	**2.93±0.17**	**2.92±0.17**	**2.94±0.18**	**3.01±0.18**	**3.03±0.19**	**3.03±0.19**
wFTL_mr	2.73±0.12	2.76±0.13	2.81±0.14	2.86±0.14	2.85±0.15	2.87±0.15	2.88±0.15	2.91±0.16	2.94±0.16	2.96±0.17	2.95±0.17	**2.98±0.18**	**3.04±0.18**	**3.07±0.19**	**3.09±0.19**
Initial Length	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
reGP	**2.86±0.18**	**2.87±0.19**	**2.91±0.19**	**2.94±0.20**	**2.97±0.21**	**2.91±0.20**	**2.92±0.20**	**2.89±0.20**	**2.83±0.21**	**2.77±0.20**	**2.83±0.21**	**2.71±0.16**	**2.70±0.17**	**2.72±0.17**	**2.75±0.18**
reMTGP	**2.99±0.21**	**2.96±0.20**	**2.96±0.21**	**2.97±0.22**	**2.98±0.23**	**2.94±0.22**	**2.96±0.23**	**2.93±0.23**	**2.79±0.20**	**2.76±0.21**	**2.81±0.22**	**2.68±0.17**	**2.65±0.18**	**2.67±0.18**	**2.72±0.18**
wFTL_se	**2.98±0.19**	**2.98±0.19**	**3.04±0.20**	**3.08±0.20**	**3.08±0.21**	**2.99±0.19**	**3.01±0.20**	**3.00±0.21**	**2.98±0.21**	**2.87±0.19**	**2.86±0.18**	**2.76±0.16**	**2.75±0.16**	**2.78±0.17**	**2.81±0.17**
wFTL_mr	**3.05±0.20**	**3.09±0.20**	**3.15±0.20**	**3.19±0.21**	**3.17±0.21**	**3.07±0.20**	**3.08±0.21**	**3.06±0.21**	**3.05±0.22**	**2.97±0.21**	**2.96±0.20**	**2.83±0.16**	**2.80±0.16**	**2.84±0.16**	**2.86±0.16**

6 ACKNOWLEDGMENT

The work in this paper was supported by grant R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1] Donald J Berndt and James Clifford. 1994. Using Dynamic Time Warping to Find Patterns in Time Series.. In *KDD workshop*, Vol. 10. 359–370.
[2] Carlo Berzuini, Riccardo Bellazzi, Silvana Quaglini, and David J Spiegelhalter. 1992. Bayesian networks for patient monitoring. *Artificial intelligence in medicine* 4, 3 (1992), 243–260.
[3] Edwin V Bonilla, Kian M Chai, and Christopher Williams. 2007. Multi-task Gaussian process prediction. In *NIPS*. 153–160.
[4] Leo Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140.
[5] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *ICML*. ACM, 18–25.
[6] Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. 2007. Improved second-order bounds for prediction with expert advice. *Machine Learning* 66, 2-3 (2007), 321–352.
[7] Lei Clifton, David A Clifton, Marco Pimentel, Peter J Watkinson, Lionel Tarassenko, and others. 2013. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering* 60, 1 (2013), 193–197.
[8] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *JMLR* 7 (2006), 551–585.
[9] Robert Durichen, Marco Pimentel, Lei Clifton, Achim Schweikard, David A Clifton, and others. 2015. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering* 62, 1 (2015), 314–322.
[10] Lloyd D Fisher and Danyu Y Lin. 1999. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health* 20, 1 (1999), 145–157.
[11] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139.
[12] Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *AAAI*. 446–453.
[13] James Douglas Hamilton. 1994. *Time Series Analysis*. Vol. 2. Princeton University Press.
[14] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F Cooper, and Gilles Clermont. 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics* 46, 1 (2013), 47–55.
[15] Zhengxing Huang, Wei Dong, Huilong Duan, and Haomin Li. 2014. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. *IEEE Journal of Biomedical and Health Informatics* 18, 1 (2014), 4–14.
[16] Zhengxing Huang, Wei Dong, Fei Wang, and Huilong Duan. 2015. Medical Inpatient Journey Modeling and Clustering: A Bayesian Hidden Markov Model Based Approach. In *AMIA*, Vol. 2015. 649.
[17] Xiaoqian Jiang, Aziz A Boxwala, Robert El-Kareh, Jihoon Kim, and Lucila Ohno-Machado. 2012. A patient-driven adaptive prediction technique to improve personalized risk estimation for clinical decision support. *JAMIA* 19, e1 (2012), e137–e144.
[18] Jan Trøst Jørgensen. 2009. New era of personalized medicine: a 10-year anniversary. *The Oncologist* 14, 5 (2009), 557–558.
[19] Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R Mishra, James Fogarty, Sean A Munson, and Julie A Kientz. 2015. A framework for self-experimentation in personalized health. *JAMIA* (2015).
[20] Thomas A Lasko, Joshua C Denny, and Mia A Levy. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS One* 8, 6 (2013), e66341.
[21] T Warren Liao. 2005. Clustering of time series datafi!?a survey. *Pattern recognition* 38, 11 (2005), 1857–1874.
[22] Nick Littlestone and Manfred K Warmuth. 1994. The weighted majority algorithm. *Information and Computation* 108, 2 (1994), 212–261.
[23] Zitao Liu and Milos Hauskrecht. 2015. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial Intelligence in Medicine* 65, 1 (2015), 5–18.
[24] Zitao Liu and Milos Hauskrecht. 2015. A Regularized Linear Dynamical System Framework for Multivariate Time Series Analysis. In *AAAI*. 1798–1804.
[25] Zitao Liu and Milos Hauskrecht. 2016. Learning Adaptive Forecasting Models from Irregularly Sampled Multivariate Clinical Data. In *AAAI*. 1273–1279.
[26] Zitao Liu and Milos Hauskrecht. 2016. Learning Linear Dynamical Systems from Multivariate Time Series: A Matrix Factorization Based Framework. In *SIAM International Conference on Data Mining*. 810–818.
[27] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Identifying suspicious URLs: an application of large-scale online learning. In *ICML*. 681–688.
[28] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.
[29] Carl Edward Rasmussen. 2006. *Gaussian Processes for Machine Learning*.
[30] Sebastian Schleidgen, Corinna Klingler, Teresa Bertram, Wolf H Rogowski, and Georg Marckmann. 2013. What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Medical Ethics* 14, 55 (2013).
[31] Peter Schulam, Fredrick Wigley, and Suchi Saria. 2015. Clustering longitudinal clinical marker trajectories from electronic health data: applications to phenotyping and endotype discovery. In *AAAI*. 2956–2964.
[32] Shai Shalev-Shwartz. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* 4, 2 (2011), 107–194.
[33] Padhraic Smyth and David Wolpert. 1999. Linearly combining density estimators via stacking. *Machine Learning* 36, 1-2 (1999), 59–83.
[34] Anne Thiébaut and Jacques Bénichou. 2004. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in medicine* 23, 24 (2004), 3803–3820.
[35] Shyam Visweswaran and Gregory F Cooper. 2004. Instance-specific Bayesian model averaging for classification. In *NIPS*. 1449–1456.