



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Continuous-Time Probabilistic Models for Longitudinal Electrical Health Records

A. D. Kaplan, U. Tipnis, J. C. Beckham, N. A.
Kimbrel, D. W. Oslin, B. H. McMahon

November 19, 2021

Journal of Biomedical Informatics

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Continuous-Time Probabilistic Models for Longitudinal Electronic Health Records

Alan D. Kaplan^{a,*}, Uttara Tipnis^a, Jean C. Beckham^{b,c,d}, Nathan A. Kimbrel^{b,c,d,e}, David W. Oslin^{f,g}, the MVP Suicide Exemplar Workgroup, Benjamin H. McMahon^h

^a*Computational Engineering Division, Lawrence Livermore National Laboratory, 7000 East Ave., Livermore, CA 94550*

^b*Durham Veterans Affairs (VA) Health Care System, Durham, NC, USA*

^c*VA Mid-Atlantic Mental Illness Research, Education and Clinical Center, Durham, NC, USA*

^d*Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC, USA*

^e*VA Health Services Research and Development Center of Innovation to Accelerate Discovery and Practice Transformation, Durham, NC, USA*

^f*VISN 4 Mental Illness Research, Education, and Clinical Center, Center of Excellence, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA*

^g*Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, PA, USA*

^h*Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA*

Abstract

Analysis of longitudinal Electronic Health Record (EHR) data is an important goal for precision medicine. Difficulty in applying Machine Learning (ML) methods, either predictive or unsupervised, stems in part from the heterogeneity and irregular sampling of EHR data. We present an unsupervised probabilistic model that captures nonlinear relationships between variables over continuous-time. This method works with arbitrary sampling patterns and captures the joint probability distribution between variable measurements and the time intervals between them. Inference algorithms are derived that can be used to evaluate the likelihood of future using under a trained model. As an example, we consider data from the United States Veterans Health Administration (VHA) in the areas of diabetes and depression. Likelihood ratio maps are produced showing the likelihood of risk for moderate-severe vs minimal depression as measured by the Patient Health Questionnaire-9 (PHQ-9).

Keywords: electronic health records, probabilistic models, mixture models, time-dependent modeling

*Corresponding author

Email address: `kaplan7@llnl.gov` (Alan D. Kaplan)

1. Introduction

Improved individualized patient care is a central goal of precision medicine [1, 2]. Historical records containing clinical measurements, laboratory results, diagnoses, and outcomes offer an opportunity to learn about the individual characteristics that lead to increased risk of disease or disorder. Using such records may help to inform the development of more precise and individualized treatment. Large quantities of Electronic Health Records (EHRs) have been collected in part to help with the development of data-driven precision medicine methods and systems [3].

One limitation that has been studied in this context is the development of data-driven prediction of outcomes, such as mortality. This can be done in either acute (e.g. emergency department, inpatient, ICU), or outpatient settings. Many machine learning (ML) techniques have been applied toward such prediction problems, such as decision trees and random forests [4, 5, 6, 7, 8, 9], neural networks [10, 11], and regression techniques [12, 13, 14]. Rather than computing an outcome prediction, probabilistic unsupervised methods are geared toward a related, but different goal [15, 16]. The aim of these methods is to construct a probability distribution model of the data that quantifies the likelihood of a collection of variables. This model can then be used to perform inference and compute probabilities for uncertain events.

However, EHR data have irregularities that make applying ML methods difficult [17]. EHR data contain a large number of variables, each of which can be sampled irregularly. In addition, the data types are heterogeneous, containing discrete, ordinal, and continuous value variables. Many existing methods address these challenges by transforming the data into a form that is amenable to application of the ML methods, such as time-windowing and quantization. However, these processes result in a typically unknown loss of information. Using a constant time window across variables may also be problematic since different time scales may apply to different variables.

In this work, we address these challenges by constructing continuous-time unsupervised probabilistic models. The model is a pairwise joint probability distribution between two measured variables and the time interval between measurements. It can be trained on longitudinal data containing arbitrary sample patterns in a computationally scalable manner, and does not use any lossy transformations of the data or time-windowing. In this way, all available timepoints can be used in estimating the model parameters. Once trained, the model can be used to compute probabilities of events of interest, such as the odds ratio of an outcome as a function of time. In addition, multiple models can be composed to form a model of an entire EHR collection.

Predictive methods are geared towards accurate prediction of a target variables. Several approaches have been developed for EHR data that operate on longitudinal data streams. Neural network models for sequence data, such as Long-Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been used to incorporate the temporal effects of EHR data in predictive methods. These can be applied in discretized steps [18], or directly operate on time

intervals [19]. These recurrent neural networks track the state dynamics and generate representations of the data at each step that can be used to predict outcomes. Rather than a predictive approach, we focus on probabilistic representations of data that can be used to compute statistical likelihoods of various outcomes with a single model.

Probabilistic models trainable on sequential EHR data have also been developed. Markov-based modeling has been used to characterize state transitions over time [20]. Continuous-time versions of these probabilistic methods include the continuous-time Markov chain and continuous-time hidden Markov model. These methods and continuous-time Bayesian networks model the time intervals between measurements in a Markov state sequence [21, 22]. The present model is designed to capture the joint distribution between time intervals and measurements, rather than the transitions between timepoints.

A related set of techniques based on probabilistic topic models can be used to uncover underlying patterns in an unsupervised model. This joint probability modeling in the form of graphical models can be used to model heterogeneous data types [23, 24]. These approaches incorporate the use of a latent variable to control conditional dependence between data elements, and can be used to derive data-driven phenotypes. Versions of Latent Dirichlet Allocation (LDA) have been applied to EHR data [25]. Similar to these methods, our approach is a latent variable model, however it models continuous time intervals explicitly.

Survival modeling is another related set of approaches for predicting the time to an event. The Cox proportional hazards model is a regression-based approach that can be used to estimate risk over time given a fixed set of covariates (see e.g. [26]). Versions incorporating time-varying regression coefficients allow for covariates to have time-dependent effects. Adaptations for longitudinal EHR data with irregularly sampled response variables can be formulated as a marked point process [27]. Covariates can reflect internal dynamics such as the time since the last measurement. These methods generate partial likelihood functions, unlike the full joint probability distributions that the method in this paper describes.

In this work we used data from the Department of Veteran Affairs (VA) electronic medical record. The VA is the largest health care system in the US comprising of more than 150 medical centers and more than 1000 community based outpatient clinical sites¹. The VA EHR has been in existing from the late 1990's. For this project we used data from 2000 to 2019.

2. Methods

2.1. Continuous-time Model

Model definition. Our method is based on estimating the joint probability distribution between two variables and the time that has elapsed between their

¹<https://www.va.gov/health/>

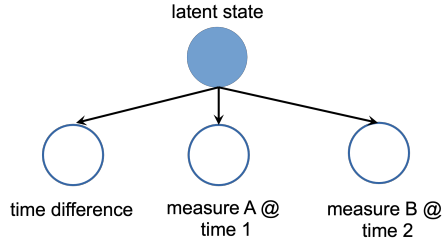


Figure 1: Graphical representation of the model, which is a mixture across two measurements and the time interval between them. Arrows indicate conditional dependence structure in the joint probability distribution.

measurement. The model can be viewed as a latent variable graphical model (Figure 1). The density function is a mixture of distributions,

$$f(\delta, x, y; N_Z) = \sum_{z=1}^{N_Z} \alpha_z f_{exp}(\delta; \lambda_z) f_G(x; \mu_z, \sigma_z^2) f_G(y; \nu_z, \xi_z^2), \quad (1)$$

where δ is the time difference, and x and y are the variables. Within the mixture, the time difference is characterized by exponential distributions $f_{exp}(\delta; \lambda)$, and the two variables are characterized by Gaussian distributions $f_G(x; \mu, \sigma^2)$. Mixing parameters are $\alpha_z \geq 0$ with $\sum_{z=1}^{N_Z} \alpha_z = 1$. The number of components is N_Z and the total parameter count for the model is $p = 6N_Z - 1$. See Appendix A for a description of the notation used in this work.

Estimation. Estimation of the parameters is performed using Expectation Maximization (EM) (see e.g. [28]). Given N samples $(\delta^{(1)}, x^{(1)}, y^{(1)}), \dots, (\delta^{(N)}, x^{(N)}, y^{(N)})$, we use the EM procedure to estimate the parameters: $\alpha, \lambda, \mu, \sigma^2, \nu, \xi^2$ that maximize $f(\boldsymbol{\delta}, \mathbf{x}, \mathbf{y}; N_Z) = \prod_{i=1}^N f(\delta^{(i)}, x^{(i)}, y^{(i)}; N_Z)$. The EM algorithm for this model is shown in Algorithm 1. See Appendix B for details on how this is derived.

Convergence of the estimation algorithm can be performed by either examining the parameters or the likelihood function. In our experiments we keep track of the log-likelihood per sample for every iteration. We claim convergence if the fractional increase in the fit is less than a fixed value (0.01) for all of the last 10 iterations.

Model selection. The number of components, N_Z , is a parameter that controls the model complexity. A greater N_Z increases the complexity and expressiveness of the model, but reduces its generalizability. Selecting an appropriate value for N_Z balances model complexity while reducing overfitting of the model to training data. We use the Bayesian Information Criterion (BIC) as a target to optimize N_Z (see e.g. [29]). The BIC for our model is $BIC(N_Z) = (6N_Z - 1) \log N - 2 \log f(\boldsymbol{\delta}, \mathbf{x}, \mathbf{y}; Z)$. The goal is to find N_Z that minimizes the BIC score. Although it is possible to perform a linear search over N_Z , in some

Algorithm 1 Parameter Estimation

Initialize $\alpha \leftarrow \alpha^{(0)}, \lambda \leftarrow \lambda^{(0)}, \mu \leftarrow \mu^{(0)}, \sigma^2 \leftarrow \sigma^{2(0)}, \nu \leftarrow \nu^{(0)}, \xi^2 \leftarrow \xi^{2(0)}$
 $j \leftarrow 0$
while not converged do
 $\gamma_i(z) \leftarrow \frac{\alpha_z^{(j)} f_{exp}(\delta_i; \lambda_z^{(j)}) f_G(x_i; \mu_z^{(j)}, \sigma_z^{2(j)}) f_G(y_i; \nu_z^{(j)}, \xi_z^{2(j)})}{f(\delta_i, x_i, y_i)}$
 $\gamma(z) = \sum_{i=1}^n \gamma_i(z)$
 $\alpha_z^{(j+1)} \leftarrow \frac{\gamma(z)}{n}$
 $\lambda_z^{(j+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(z) \delta_i}{\sum_{i=1}^n \gamma_i(z)}$
 $\mu_z^{(j+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(z) x}{\gamma(z)}$
 $\sigma_z^{2(j+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(z) (x - \mu_z)^2}{\gamma(z)}$
 $\nu_z^{(j+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(z) x}{\gamma(z)}$
 $\xi_z^{2(j+1)} \leftarrow \frac{\sum_{i=1}^n \gamma_i(z) (x - \nu_z)^2}{\gamma(z)}$
 $j \leftarrow j + 1$
end while

cases this can be an overly expensive approach. In our experiments we utilize
 110 Bayesian Optimization (BO) methods to optimize the BIC [30]. This established
 approach estimates a Gaussian Process to characterize the uncertainty of the
 $BIC(N_Z)$ function and selects points to sample that satisfy a defined criterion.
 For model selection, the BO operates on an outer loops, with EM estimation
 occurring in the inner loop. Every outer iteration estimates a new model with a
 115 value of N_Z chosen by the BO. The model with the lowest BIC is chosen as the
 final model. See Appendix C for details on parameters used for this approach.

Inference algorithms. Once a model is trained, there are a number of inference
 procedures that can be applied. The distribution of one variable given another
 over time can be computed by calculating

$$f(y|x, \delta) = \sum_{z=1}^{N_Z} \Pr(Z = z|x, \delta) f_G(y; \nu_z, \xi_z^2), \quad (2)$$

which is a Gaussian Mixture Model with mixing coefficients determined by the
 inputs x and δ . The distribution of the latent variable given x and δ can be
 found using

$$\Pr(Z = z|x, \delta) = \frac{\alpha_z f_{exp}(\delta; \lambda_z) f_G(x; \mu_z, \sigma_z^2)}{f(x, \delta)}.$$

The normalizing constant in the denominator does not depend on z and ensures
 that $\sum_z \Pr(Z = z|x, \delta) = 1$.

More than one model can be composed to perform inference on a common
 variable. In this setup, we would like to compute the likelihood of a target
 variable Y given input variables X_1, \dots, X_M and time differences between these

variables and Y , $\Delta_1, \dots, \Delta_M$. This can be performed using a collection of models f_1, \dots, f_M , where $f_i(\delta, x, y)$ is the joint distribution between (Δ_i, X_i, Y) . Composing these models together yields

$$f(y|x_1, \dots, x_M, \delta_1, \dots, \delta_M) = \prod_{i=1}^M f(y|x_i, \delta_i). \quad (3)$$

This framework can be used to infer future values of Y given past values of the input variables. If the current time is $t = 0$ and the input variables were collected at time t_1, \dots, t_m , then using the trained models the distribution of Y at time t is $f(y|x_1, \dots, x_m, t - t_1, \dots, t - t_m)$.

2.2. Baseline Models

We formulate two baseline models. These are used to evaluate and compare the fit of data to our model. The first model is a multivariate Gaussian:

$$f(\delta, x, y) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}([\delta, x, y] - [\mu_\delta, \mu_x, \mu_y])^T \Sigma^{-1} ([\delta, x, y] - [\mu_\delta, \mu_x, \mu_y])}, \quad (4)$$

where μ_δ , μ_x , and μ_y are the mean parameters, and Σ is the covariance matrix.

The second baseline model is a conditional bivariate Gaussian model, where a bivariate Gaussian is conditioned on the value of an ordinal variable y . The probability density function for this model is,

$$f(\delta, x, y) = p_y \frac{1}{\sqrt{2\pi|\Sigma|_y}} e^{-\frac{1}{2}([\delta, x] - [\mu_{\delta|y}, \mu_{x|y}])^T \Sigma_y^{-1} ([\delta, x] - [\mu_{\delta|y}, \mu_{x|y}])}, \quad (5)$$

where $\mu_{\delta|y}$, $\mu_{x|y}$ are the mean parameters dependent on y , and Σ_y is the covariance matrix dependent on y . This model allows for increased expressiveness compared to (4), however it has the requirement that y is ordinal.

2.3. Variables and Data Extraction

In the following case studies, we use the following variables:

- A1C is the hemoglobin A1C level as recorded in a standard laboratory test. The normal range is less than 5.7%, 5.7%–6.4% indicates prediabetes, and greater than 6.5% indicates diabetes².
- Cholesterol is the total cholesterol as recorded in a standard laboratory test. The normal range is less than 200 mg/dL, 200 mg/dL - 239 mg/dL is borderline high, and greater than 240 mg/dL is considered high.
- 250.XX is any ICD9 code that starts with 250, indicating a diabetes diagnosis,

²<https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html>

- CRP is the level of C-Reactive Protein as recorded in a standard laboratory test, which is a measure of inflammation,
- 140 • PHQ-2 is the 2-item Patient Health Questionnaire used as a screening tools for depressed mood,
- PHQ-9 is the 9-item Patient Health Questionnaire used to assess depression severity on a 0-27 point scale, with high values indicating greater severity.

145 Using the VHA patient record data, we organize the subjects into 100 cohorts based on age. Each cohort contains approximately 230,000 patients. For each case study, we randomly sample 100 patients from each cohort that have at least one recorded value from each variable. We then remove instances that have incomplete recordings, such as laboratory values missing.

150 3. Results

In this section, we describe results on synthetic examples, diabetes, and depression. Synthetic examples (Section 3.1) are designed to demonstrate the ability to model nonlinear and time-dependent behavior. Results on diabetes data (Section 3.2) are based on well-known relationships between hemoglobin
 155 A1C and diabetes, and is used to verify that the approach can capture previously understood relationships. Depression (Section 3.3) focuses on much weaker and less understood relationships between laboratory results and survey-based depression tools. Table 4 shows the number of subjects, samples, and components that resulted from model selection for all models. PHQ-9 models were trained
 160 by performing 20 runs of the model selection procedure with different random initializations for each one and choosing the model with the best BIC value. Over the 20 runs, we show the mean and 2 standard deviations of the model fit in Table 4.

3.1. Synthetic Examples

165 The models for these synthetic examples were designed to express time-dependent interaction between two variables X and Y . After defining the models, we then sample from them and re-estimate the parameters. Then the expected value of Y given X is computed over time to illustrate the association between variables. This is compared to the same computation for the estimated
 170 model to visualize estimation accuracy.

Diminishing temporal effect. In the first example, we construct a model to show a positive association between two variables X and Y that decreases in strength over time. This is designed to represent commonly occurring behavior. Table 1 shows the parameters chosen to illustrate this relationship between X and Y .
 175 The 6 components in this model all have equal mixing coefficients. The mean of the exponential distribution ($1/\lambda_z$) for each component controls its temporal reach. Relative to the other components, components 1 and 2 have greater

Table 1: Synthetic Model Showing Diminishing Temporal Effect

z	α_z	λ_z	μ_z	ν_z
1	0.17	1.00	0.0	0.0
2	0.17	1.00	1.0	1.0
3	0.17	0.50	0.0	0.0
4	0.17	0.50	1.0	0.5
5	0.17	0.33	0.0	0.0
6	0.17	0.33	1.0	0.2

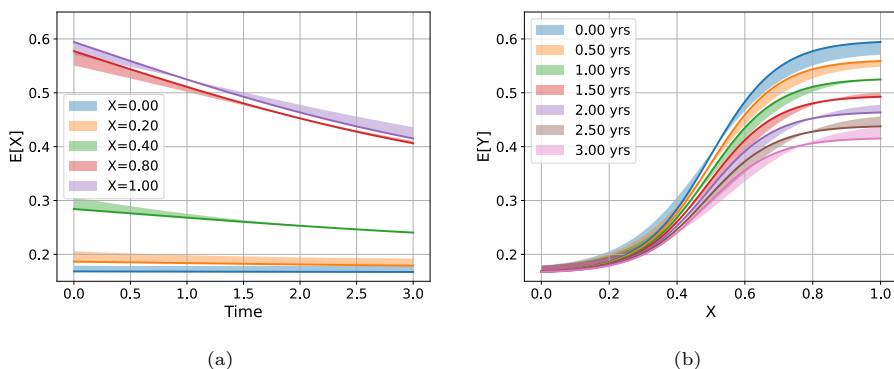


Figure 2: Inference results for synthetic example with diminishing positive association (Section 3.1). Figure 2a shows the expected value of Y over time for selected values of X . Figure 2b shows the same inference, but as a function of continuously valued X with selected time points. For both plots, solid lines are from the original model, and shaded regions are deviations to the re-estimated model.

influence in earlier times, while components 5 and 6 have weaker influence later in time. All variance parameters are set to 1.

180 The expected value of Y given X over time, $E[Y|X = x, \Delta = \delta]$ is shown in Figure 2a. This is computed by taking the expected value of the distribution in (2), $\sum_y y f(y|x, \delta)$. The solid lines are the expected values using the original model, and the shaded area is the region between the original model and a re-estimated model, trained on 10,000 samples drawn from the model. The same computation, but plotted as the expected value as a function of x is shown in 185 Figure 2b. As time increases, the effect that X has on Y decreases.

190 *Changing directionality over time.* In the second example, X has a positive association with Y initially, but over time inverts to a negative effect. Table 2 shows the parameters used in this 8 component model. As in the previous example, all components have equal mixing coefficients. The means of the time intervals in this model ($1/\lambda_z$) are 1, 1.5, 2, and 3. For each of these values, 2 components characterize the association between X and Y , initially with a positive directionality that weakens initially over time. Then the directionality

Table 2: Synthetic Model Showing changing directionality over time

z	α_z	λ_z	μ_z	ν_z
1	0.125	1.00	0.0	0.0
2	0.125	1.00	1.0	1.0
3	0.125	0.67	0.0	0.0
4	0.125	0.67	1.0	0.5
5	0.125	0.5	0.0	0.5
6	0.125	0.5	1.0	0.0
7	0.125	0.33	0.0	1.0
8	0.125	0.33	1.0	0.0

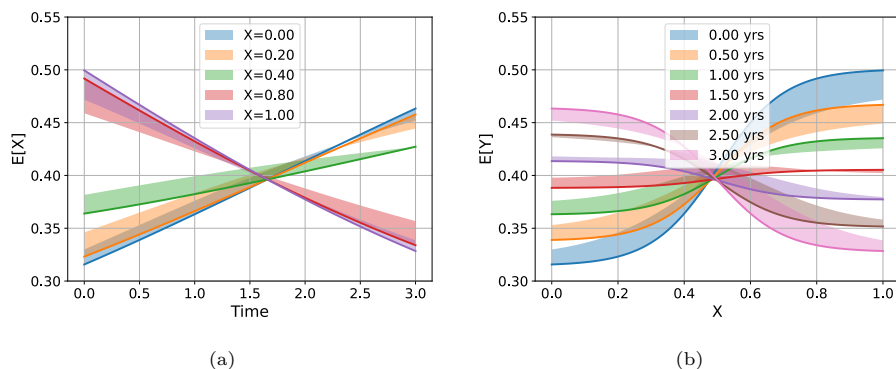


Figure 3: Inference results for synthetic example with inverting association over time (Section 3.1). Figure 3a shows the expected value of Y over time for selected values of X . Figure 3b shows the same inference, but as a function of continuously valued X with selected time points. For both plots, solid lines are from the original model, and shaded regions are deviations to the re-estimated model.

inverts and grows stronger in the other direction through components 5-8.

195 As in the previous example, we compute the expected value of Y given X over time. Figure 3 shows these computations, both for the original model and re-estimated model. The re-estimated model was trained on 10,000 samples drawn from the original model.

3.2. Diabetes

200 In these results, we model the well-understood relationship between hemoglobin A1C and diabetes. Diabetes diagnoses are typically given for A1C levels greater than 6.5%. In this study, each x is the A1C value and each δ is the time elapsed to the next occurring 250.XX diagnosis (y is not used). We extracted 5,661 samples from 2,539 subjects (see Table 4).

205 Running the Bayesian Optimization model selection resulted in $N_Z = 10$ components. Figure 4 shows a scatter plot of the data (A1C vs time interval) and equiprobable contours generated by the likelihood function of the model. As

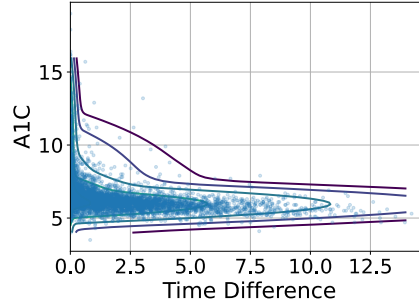


Figure 4: Scatter plot of data point used in training the A1C-250.XX model (Section 3.2) along with equiprobable contours of the trained likelihood function.

expected, high values of AIC result in a quick 250.XX diagnoses, while smaller values have a longer time interval to diagnoses. The model captures this L-shaped distribution.

3.3. Depression

In this section, we described results from three models trained to capture associations that Cholesterol, CRP, and PHQ-2 have with the PHQ-9. Using the trained models, likelihood ratios are computed for moderate-severe (PHQ-9 > 14) vs minimal depression (PHQ-9 < 5) for each variable. This is done in the two-dimensional space of measurement vs time. These models are then combined using (3). Combinations of measurements and time intervals are shown that contribute to varying levels of the likelihood ratios.

Data extraction. Table 4 shows the number of subjects and samples extracted for each model. One sample is a triplet that consists of a measured input (x), the PHQ-9 score (y) and the time interval in years between them (δ). The measured value x can be Cholesterol, CRP, or PHQ-2, depending on the model. Any two instances of x and y are included that occur in the subject's clinical record, as long as the PHQ-9 score occurs after x was recorded. Extracting data in this manner means that the same value x could be used in multiple samples, in cases where multiple PHQ-9 scores were recorded after the x value was measured.

Model training and comparison to baseline models. To evaluate and compare the model fit to the baseline models, we train on half of the samples and compute the likelihood on the held-out data. Table 3 shows the likelihood ratios between our model (1) and the two baseline models (4) and (5). These values indicate how many times more likely the held-out data is for our model than the baseline model. Thus, values greater than 1 indicate worse fit than our model. For example, the held-out data is approximately 3 times more likely for our model than the multivariate model for Cholesterol - PHQ-9.

Table 3: Model Fit Comparison to Baseline Models

Model	Cholesterol - PHQ-9	CRP - PHQ-9	PHQ-2 - PHQ-9
Multivariate (4)	3.01	11.72	9.38
Conditional Bivariate (5)	1.22	12.29	7.22

Moderate-severe and minimal depression outcome classes. Likelihood functions for two outcome classes based on the PHQ-9 are defined as $\text{PHQ-9} < 5$,

$$f_{min}(x, \delta) = \sum_{y < 5} f(x, y, \delta)$$

for minimal depression, and $\text{PHQ-9} > 15$,

$$f_{modsev}(x, \delta) = \sum_{y > 15} f(x, y, \delta)$$

for moderate-severe depression. For each model, we show (Figures 5a, 5b, 5c) the total likelihood, or prevalence score, relative to its mode,

$$p(x, \delta) = \frac{f_{min}(x, \delta) + f_{modsev}(x, \delta)}{\max[f_{min}(x, \delta) + f_{modsev}(x, \delta)]}, \quad (6)$$

235 which gives the likelihood of any x, δ combination relative to the most likely x, δ combination.

Likelihood ratio between outcome classes. The likelihood, or odds, ratio between the two severity classes is $f_{modsev}(x, \delta) / f_{min}(x, \delta)$. This ratio shows how much more likely moderate-severe depression is than minimal depression, as measured
240 by the PHQ-9. For example, a value of 2 means that the moderate-severe depression class is twice as likely as the minimal depression class, and a value of 0.5 indicates the opposite.

The model is capable of generating likelihood ratios for any continuous values of x, δ , even those far outside what was seen in the training data. In general, we
245 should only expect to make confident assessments of likelihood for values that have a reasonable prevalence. We use the prevalence score to help us determine appropriate regions to examine, and show (Figures 5d, 5e, 5f) the likelihood ratios for the region of x, δ that has a prevalence score $p(x, \delta)$ greater than 0.05. Figure 6 shows the prevalence score and likelihood ratios for PHQ-2 in a narrow
250 time range.

Composite likelihood. These three models can be combined to evaluate the likelihood of PHQ-9 over time with respect to Cholesterol, CRP, and PHQ-2. The likelihood function is composed using contributions from each model as described in (3). This results in a high-dimensional likelihood function that can
255 be difficult to visualize. Table 5 shows sample values of time intervals and measured values that contribute towards a given composite likelihood ratio. Each

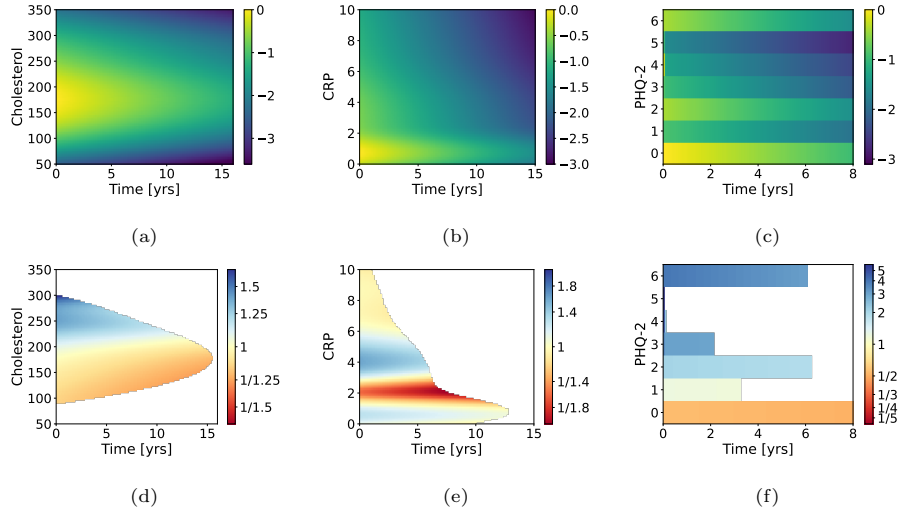


Figure 5: Normalized prevalence scores for the Cholesterol (5a), CRP (5b), and PHQ-2 (5c) models using a \log_{10} scale. These scores are the normalized likelihoods for both the moderate-severe and minimal depression outcomes, relative to the mode. Likelihood ratios for moderate-severe vs minimal depression for these models (5d, 5e, 5c). Yellow color indicates approximately equal likelihood between outcome classes, blue indicates higher risk for the moderate-severe class, and red corresponds to a lower risk. A version for PHQ-2 showing a narrow time range is shown in Figure 6.

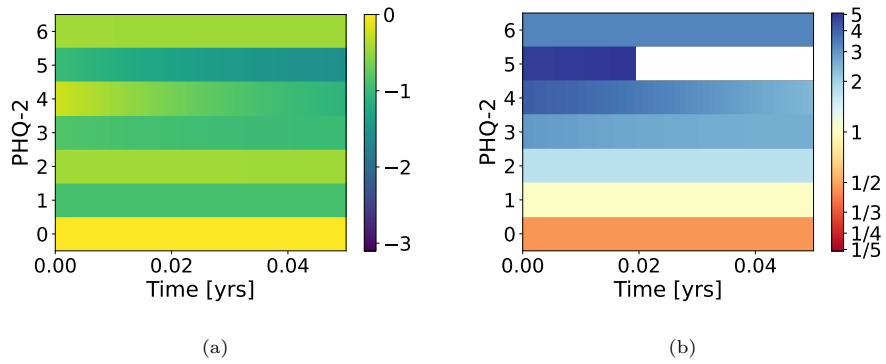


Figure 6: Prevalence scores and likelihood ratio for the PHQ-2 model. Zoomed version of Figures 5c and 5f showing earlier times.

Table 4: Number of subjects, samples, selected model order, and BIC variability due to initialization

Model	Subjects	Samples	Components	BIC
A1C-250.XX	2,539	5,661	10	-
Cholesterol-PHQ-9	9,177	402,841	292	19.93 ± 0.15
CRP-PHQ-9	5,932	53,542	44	16.10 ± 0.05
PHQ-2-PHQ-9	8,112	125,234	121	10.48 ± 0.19

Table 5: Sample values with fixed composite likelihood ratio of moderate-severe depression vs minimal depression.

likelihood ratio	$\delta_{\text{Cholesterol}}$	$x_{\text{Cholesterol}}$	δ_{CRP}	x_{CRP}	$\delta_{\text{PHQ-2}}$	$x_{\text{PHQ-2}}$
0.21 (min)	13.1	141	7.6	2.0	8.0	0
0.25	12.4	171	5.8	1.9	0.0	0
0.33	8.1	208	4.5	1.8	0.0	0
0.50	1.5	141	6.8	1.1	0.0	0
1.0	7.8	232	10.8	1.4	3.2	1
2.0	8.6	220	2.1	3.9	0.0	1
4.0	0.0	286	3.9	4.1	6.2	2
6.0	0.0	299	0.6	4.0	0.1	4
8.0	0.0	299	0.0	4.0	2.1	3
10.0	0.2	296	0.0	4.0	0.1	6
15.2 (max)	0.0	299	0.0	4.0	0.0	5

row corresponds to a likelihood ratio, and a single combination of variables that produces that odds ratio. The first row corresponds to the smallest likelihood ratio that the models output (0.21), and the last row has the largest likelihood ratio (15.2). Note that there are many different combinations of variables that could produce any single likelihood ratio.

4. Discussion

The synthetic experiments show two examples of the dynamics that can be expressed with the model. Although we chose parameters in the model in a specific way, there are many different parameter values that can lead to similar results. In general, we do not have identifiability properties for the model.

We used 10,000 samples in these examples to show that the re-estimated models have similar behavior to the original model. It would be possible to perform a more rigorous analysis of the estimate accuracy; however, there is also a strong dependency on the model structure.

The diabetes model was trained on extracted A1C measurements and the time until the next future 250.XX diagnosis. In this setup, several A1C measurements can be extracted with respect to the same 250.XX diagnosis. However, only the next 250.XX diagnoses is considered, not all future diagnoses.

275 In this problem, all of the subjects used to extract samples have 250.XX
diagnoses. This model can be used to infer how long it will take to receive a
A1C diagnosis, rather than if a subject will receive the diagnosis at all. Figure
4 shows the raw data along with equiprobable likelihood contours of the model.
Large values of A1C result in a quick 250.XX diagnosis, and smaller values result
280 in a longer time interval to the first 250.XX diagnosis. By visual inspection, we
confirm that the model is capturing the density of samples in the data.

For the depression examples, three models were trained that include the 0-27
point PHQ-9 tool. In contrast with from the diabetes model, we extracted all
pairwise instances where the PHQ-9 occurred after the measured Cholesterol,
285 CRP, or PHQ-2. All subjects had at least one PHQ-9 instance creating a bias
in the sample, as this implies that there is likely some indication of depression,
or reason to test for it.

Figures 5a and 5d show the prevalence scores and likelihood ratios as a function
of Cholesterol and the time interval. The normative Cholesterol range is less
290 than 200 mg/dL. These values of Cholesterol show a likelihood ratio of approx-
imately 1 or less than 1, indicating less risk of moderate-severe depression for
Cholesterol less than 200 mg/dL. On the other hand, values greater than 200
mg/dL start to show increased likelihood of moderate-severe depression relative
to minimal depression. This trend continues to approximately 300 mg/dL,
295 given a likelihood ratio around 2, indicating that moderate-severe depression is
approximately 2 times more likely.

Prevalence and likelihood ratios for CRP are shown in Figures 5b and 5e.
Based on the prevalence figures, a large majority of CRP values are in the range
0 mg/dL - 2 mg/dL. Values greater than 6 mg/dL have likelihood ratios around
300 1. The lowest likelihood ratios occur in a range around 2 mg/dL. The range 3
mg/dL - 6 mg/dL has the highest values. These patterns indicate that the CRP-
PHQ-9 relationship may be more complicated than that of Cholesterol-PHQ-9.

As Figures 5c and 6a show, prevalence for PHQ-2 is mostly in very early
times periods for $\text{PHQ-2} > 0$, and more elongated through time for $\text{PHQ-2} = 0$.
305 This is reflective of the usage of the PHQ-2 as a screening tool for administration
of the PHQ-9. Likelihood ratios for moderate-severe and minimal depression are
shown in Figure 5f. As the PHQ-2 increases, so does the risk of moderate-severe
depression. A value of $\text{PHQ-2} = 0$ has a likelihood ratio less than 1, whereas
 $\text{PHQ-2} = 6$ has a likelihood ratio of approximately 4. Compared to Cholesterol
310 and CRP, the PHQ-2 has a stronger influence on the PHQ-9. Zooming in to short
time intervals (Figure 6b), we see the likelihood ratios for the less prevalent 1,
3, and 5 PHQ-2 scores.

The number of components determined by the model selection procedure
for each model is shown in Table 4. This number varies for each model as the
315 complexity required to model the data is different in each case. The variability
is due in part to the sample size, and in part due to the underlying nonlinear
relationships that the model is attempting to capture.

Using all three models, we can formulate a single joint probability distribu-
tion using (3). This can be used to evaluate prevalence and risk for any
320 combination of Cholesterol, CRP, and PHQ-9. In general, a single risk score

(likelihood ratio) can be produced by many different combinations of these variables. In Table 5, a single combination for given likelihood ratios are given. In general, the Cholesterol values increase as risk increases, but in some cases decreased Cholesterol is traded for increased values in other dimensions. For example, moving from a risk of 1.0 to 2.0, the sample provided in the Table shows Cholesterol decreasing, but the PHQ-2 increasing from 1 to 2. As we have seen, the PHQ-2 is a stronger indicator of PHQ-9 risk than Cholesterol. A PHQ-2 of 2 carries a risk of approximately 2.0 (Figure 5f), requiring the other variables to have a risk around 1.0 for the total risk to be 2.0.

A situation of interest is when there are a large number of variables. In this case, we expect many of the variables may have limited effect on the target variable Y . This can be expressed through independence of a variable X_i with Y , $f(x_i, y) = f(x_i)f(y)$. When this occurs for a subset of the variables $X_{P+1}, X_{P+2}, \dots, X_M$, then from the model definition (3), $f(x|x_1, \dots, x_M, \delta_1, \dots, \delta_M) \propto \prod_{i=1}^P f(y|x_i, \delta_i)$. Therefore, only the dependent variables will have an impact under this model. This analysis, however, does not include estimation error, which would contribute to inference error.

Limitations. The structure defined above is designed to capture the joint distribution between variables and the time difference between them. Within each mixture component, the choice of Gaussian distributions for the variables and an Exponential distribution for the time difference may impart model bias. The use of mixtures of these distributions mitigates this limitation to some extent.

Estimation using the EM algorithm is sensitive to the parameter initialization. Several approaches can be used to reduce the variability caused by this that use multiple initializations. These include selecting the best fit model from the set of trained models or averaging resulting models. Model selection performed by the BO method trains multiple models, each of which can be initialized differently. Increasing a larger number of BO iterations may also help in increasing robustness.

5. Conclusions

In this work, we have developed a probabilistic model that captures the joint distribution between two measurements and the time interval between them. The model utilizes a latent variable to control the dependence between the measurements and time interval. Estimation algorithms using EM and inference equations are derived. A trained model can be used to calculate future risk at arbitrary time points. In addition, multiple models can be composed to weigh the contributions of risk from each model.

We show results on data derived from the VHA. Our focus for this work is on diabetes and depression, in particular the impact of A1C on 250.XX diagnoses and that of Cholesterol, CRP, and PHQ-2 on the PHQ-9. We show likelihood ratios for two classes of outcomes corresponding to moderate-severe and minimal depression as measured by the PHQ-9 tool.

This method allows for continuous-time modeling of longitudinal EHR data. It provides a model-based approach to calculating risk as a function of time. In contrast to existing methods, this model is a full likelihood probabilistic model that can be used to infer any outcome class with a single model. The model complexity can change to fit the complexity of the underlying data, as opposed to methods with a fixed number of coefficients.

There are many avenues for expanding this work. In these results, we do not utilize any control groups. All subjects under consideration have taken the PHQ-9 in the depression models, or had a 250.XX diagnosis in the diabetes model. It would be beneficial to train additional models using right-censored data that have never taken the PHQ-9, or have never had a 250.XX diagnosis. This would enable evaluation of risk for not having these outcomes and thereby form a more complete picture of possible outcomes.

The composition of models does not take into account interaction between variables (for example Cholesterol and CRP), but rather uses conditional independence of these variables given the outcome PHQ-9. In this way, they each make conditionally independent contributions to the risk. It is possible to expand the model into higher dimensions and taking into account higher order interactions. However, this would greatly increase the complexity of the model and computational cost of estimating the model. Depending on the application this may not or not be worth the added expressiveness gained.

These models can be trained on much larger collections of variables. Since they are trained in a pairwise fashion, this can be done in a scalable manner. The estimation of each model can also be parallelized if needed, as the EM algorithm fits neatly in the map-reduce framework. This method is geared towards the incorporation of entire longitudinal EHR data streams that could enable probabilistic inference and quantification of risk odds for diverse subjects. The analysis of such models is left as future work.

6. Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This research is based on data from the Million Veteran Program, Office of Research and Development (ORD), Veterans Health Administration (VHA), and was supported by award #I01CX001729 from the Clinical Science Research and Development (CSR&D) Service of VHA ORD. This publication does not represent the views of the Department of Veteran Affairs or the United States Government. J.C. Beckham was also supported by a Senior Research Career Scientist Award (#IK6BX003777) from CSR&D.

The MVP Suicide Exemplar Workgroup for this publication includes Khushbu Agarwal, Allison E. Ashley-Koch, Mihaela Aslan, Jean C. Beckham, Edmond Begoli, Tanmoy Bhattacharya, Ben Brown, Patrick S. Calhoun, Mikaela Cashman McDevitt, Kei-Hoi Cheung, Sutanay Choudhury, Ashley M. Cliff, Judith D. Cohn, Silvia Crivelli, Leticia Cuellar-Hengartner, Haedi E. Deangelis, Michelle F. Dennis, Sayera Dhaubhadel, Patrick D. Finley, Kumkum Ganguly, Michael

Table A.6: Notation

symbol	description
X, Y	measurement random variables
Δ	time elapsed between measurements random variable
x, y	realization of X, Y
δ	realization of Δ
N_Z	number of components
z	component index
f	joint distribution between (Δ, X, Y)
f_{exp}	exponential distribution
f_G	Gaussian distribution
α_z	mixing coefficient for component z
$\boldsymbol{\alpha}$	collection of $\alpha_1, \dots, \alpha_{N_Z}$
λ_z	rate parameter for component z
$\boldsymbol{\lambda}$	collection of $\lambda_1, \dots, \lambda_{N_Z}$
μ_z, σ_z^2	mean and variance parameters of X for component z
$\boldsymbol{\mu}, \boldsymbol{\sigma}^2$	collection of μ_i, \dots, μ_{N_Z} and $\sigma_1^2, \dots, \sigma_{N_Z}^2$
ν_z, ξ_z^2	mean and variance parameters of Y for component z
$\boldsymbol{\nu}, \boldsymbol{\xi}^2$	collection of ν_i, \dots, ν_{N_Z} and $\xi_1^2, \dots, \xi_{N_Z}^2$
N	number of samples
$\delta^{(i)}, x^{(i)}, y^{(i)}$	i th sample
$\boldsymbol{\delta}, \boldsymbol{x}, \boldsymbol{y}$	collection of samples $\delta^{(1)}, \dots, \delta^{(N)}$, $x^{(1)}, \dots, x^{(N)}$, and $y^{(1)}, \dots, y^{(N)}$
t	time

R. Garvin, Joel E. Gelernter, Lauren P. Hair, Phillip D. Harvey, Elizabeth R. Hauser, Michael A. Hauser, Nick W. Hengartner, Daniel A. Jacobson, Piet C. Jones, David Kainer, Alan D. Kaplan, Ira R. Katz, Rachel L. Kember, Nathan
410 A. Kimbrel, Angela C. Kirby, John C. Ko, Beauty Kolade, John Lagergren, Matthew Lane, Daniel F. Levey, Drew Levin, Jennifer H. Lindquist, Xianlian Liu, Ravi K. Madduri, Carrie Manore, Susana B. Martins, John F. McCarthy, Benjamin H. McMahon, J. Izaak Miller, Destinee Morrow, David W. Oslin, Mirko Pavicic, John P. Pestian, Saiju Pyarajan, Xue J. Qin, Nallakkandi Rajeevan, Christine M. Ramsey, Ruy Ribeiro, Jonathon Romero, Alex Rodriguez,
415 Daniel Santel, Noah Schaefferkoetter, Yunling Shi, Murray B. Stein, Kyle A. Sullivan, Ning Sun, Suzanne R. Tamang, Alice Townsend, Jodie A. Trafton, Angelica Walker, Xiange Wang, Victoria Wangia-Anderson, Renji Yang, Shin-jae Yoo, Hong-Jun Yoon, Rafael Zamora-Resendiz, and Hongyu Zhao.

420 Appendix A. Notation

Table A.6 contains descriptions of the notation used in this work.

Appendix B. Estimation Details

The log-likelihood of the data is $L = \sum_{n=1}^N \log f(\delta_n, x_n, y_n)$. Using the variable k_n to indicate the component membership of sample n , the complete data log-likelihood is $\log f(\delta, x, y, k) = \sum_{n=1}^N \sum_{z=1}^Z I(k_n = z)(\log \alpha_z + \log f_{exp}(\delta; \lambda_z) + \log f_G(x; \mu_z, \sigma_z^2) + \log f_G(y; \nu_z, \xi_z^2))$. Taking the expected value of the complete data log-likelihood with respect to $k|\delta, x, y$, we have

$$E_{k|\delta, x, y} \log f(\delta, x, y, k) = \sum_{n=1}^N \sum_{z=1}^Z \gamma_i(z)(\log \alpha_z + \log f_{exp}(\delta; \lambda_z) + \log f_G(x; \mu_z, \sigma_z^2) + \log f_G(y; \nu_z, \xi_z^2)),$$

where $\gamma_i(z) = f(k_i = z|\delta_i, x_i, y_i)$. The estimation algorithm (Algorithm 1) iterates between calculating $\gamma_i(z)$ and maximizing the expected complete data log-likelihood function.

Appendix C. Bayesian Optimization Parameters

We use the Bayesian Optimization Python package [31]. The bounds on Z are initially set to $[0, 1000]$. Sequential domain reduction is used to shrink the domain as the iterations progress [32]. The default values of the parameters are used: shrinkage=0.7, panning=1, and zoom=0.9). We start with 4 initial points and proceed with 20 iterations.

References

- [1] G. S. Ginsburg, K. A. Phillips, Precision medicine: From science to value, *Health Aff.* 37 (5) (2018) 694–701.
- [2] M. R. Kosorok, E. B. Laber, Precision medicine, *Annu. Rev. Stat. Appl.* 6 (1) (2019) 263–286.
- [3] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel, J. L. Warner, The evolving use of electronic health records (EHR) for research, *Semin. Radiat. Oncol.* 29 (4) (2019) 354–361.
- [4] J. C. Weiss, S. Natarajan, P. L. Peissig, C. A. McCarty, D. Page, Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records, *AIMag* 33 (4) (2012) 33–33.
- [5] R. J. Ellis, Z. Wang, N. Genes, A. Ma’ayan, Predicting opioid dependence from electronic health records with machine learning, *BioData Min.* 12 (2019) 3.
- [6] F. Rahimian, G. Salimi-Khorshidi, A. H. Payberah, J. Tran, R. Ayala Solares, F. Raimondi, M. Nazarzadeh, D. Canoy, K. Rahimi, Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records, *PLoS Med.* 15 (11) (2018) e1002695.

- [7] F. Xie, B. Chakraborty, M. E. H. Ong, B. A. Goldstein, N. Liu, AutoScore: A machine Learning-Based automatic clinical score generator and its application to mortality prediction using electronic health records, *JMIR Medical Informatics* 8 (10) (2020) e21798.
- 455 [8] S.-M. Zhou, F. Fernandez-Gutierrez, J. Kennedy, R. Cooksey, M. Atkinson, S. Denaxas, S. Siebert, W. G. Dixon, T. W. O’Neill, E. Choy, C. Sudlow, UK Biobank Follow-up and Outcomes Group, S. Brophy, Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis, *PLoS One* 460 11 (5) (2016) e0154515.
- [9] C. Ye, J. Li, S. Hao, M. Liu, H. Jin, L. Zheng, M. Xia, B. Jin, C. Zhu, S. T. Alfreds, F. Stearns, L. Kanov, K. G. Sylvester, E. Widen, D. McElhinney, X. B. Ling, Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm, *Int. J. Med. Inform.* 465 137 (2020) 104105.
- [10] H. J. Kam, H. Y. Kim, Learning representations for the early detection of sepsis with deep neural networks, *Comput. Biol. Med.* 89 (2017) 248–255.
- [11] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. J. Zheng, F. Wang, H. Wu, H. Xu, D. Zhi, A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous 470 EHR data set, *J. Biomed. Inform.* 84 (2018) 11–16.
- [12] C. Su, R. Aseltine, R. Doshi, K. Chen, S. C. Rogers, F. Wang, Machine learning for suicide risk prediction in children and adolescents with electronic health records, *Transl. Psychiatry* 10 (1) (2020) 413.
- 475 [13] M. E. Levine, D. J. Albers, G. Hripcsak, Methodological variations in lagged regression for detecting physiologic drug effects in EHR data, *J. Biomed. Inform.* 86 (2018) 149–159.
- [14] J. Wu, J. Roy, W. F. Stewart, Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches, *Med. Care* 48 (6 Suppl) (2010) S106–13. 480
- [15] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, *Nature* 521 (7553) (2015) 452–459.
- [16] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
- 485 [17] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis, *IEEE J Biomed Health Inform* 22 (5) (2018) 1589–1604.

- [18] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, X. Wei, Predicting the risk of heart failure with EHR sequential data modeling, *IEEE Access* 6 (2018) 9256–9261.
- 490 [19] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, Doctor AI: Predicting clinical events via recurrent neural networks, *JMLR Workshop Conf. Proc.* 56 (2016) 301–318.
- [20] Z. Huang, Z. Ge, W. Dong, K. He, H. Duan, Probabilistic modeling personalized treatment pathways using electronic health records, *J. Biomed. Inform.* 86 (2018) 33–48.
- 495 [21] Y.-Y. Liu, S. Li, F. Li, L. Song, J. M. Rehg, Efficient learning of Continuous-Time hidden markov models for disease progression, *Adv. Neural Inf. Process. Syst.* 28 (2015) 3599–3607.
- 500 [22] F. Stella, Y. Amer, Continuous time bayesian network classifiers, *J. Biomed. Inform.* 45 (6) (2012) 1108–1119.
- [23] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, N. Elhadad, Learning probabilistic phenotypes from heterogeneous EHR data, *J. Biomed. Inform.* 58 (2015) 156–165.
- 505 [24] M. B. Mayhew, B. K. Petersen, A. P. Sales, J. D. Greene, V. X. Liu, T. S. Wasson, Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models, *J. Biomed. Inform.* 78 (2018) 33–42.
- [25] Z. Huang, W. Dong, H. Duan, A probabilistic topic model for clinical risk stratification from electronic health records, *J. Biomed. Inform.* 58 (2015) 28–36.
- 510 [26] L. Ohno-Machado, Modeling medical prognosis: survival analysis techniques, *J. Biomed. Inform.* 34 (6) (2001) 428–439.
- [27] T. Martinussen, T. H. Scheike, *Dynamic Regression Models for Survival Data*, Springer, New York, NY, 2006.
- 515 [28] T. K. Moon, The expectation-maximization algorithm, *IEEE Signal Process. Mag.* 13 (6) (1996) 47–60.
- [29] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Science & Business Media, 2009.
- 520 [30] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc., 2012.
- 525 [31] F. Nogueira, *Bayesian Optimization: Open source constrained global optimization tool for Python* (2014).

- [32] N. Stander, K. J. Craig, On the robustness of a simple domain reduction scheme for simulation-based optimization, *Eng. Comput.* 19 (4) (2002) 431–450.