Survival Topic Models for Predicting Outcomes for Trauma Patients

Yuanyang Zhang, Richard Jiang and Linda Petzold
Department of Computer Science, University of California, Santa Barbara

***Abstract*—Data mining techniques have been proposed to predict mortality for ICU patients using their demographic data, measurements and notes from doctors and nurses. Most of these techniques suffer from two main drawbacks. First, they model the mortality prediction problem as a binary classification problem, while ignoring the time of death as continuous values. Second, they use topic models to analyze the notes, while ignoring the relationship between measurements, notes and mortality/discharge outcomes. In this paper we propose a novel model called the survival topic model (SVTM), which models patients' measurements, notes and mortality/discharge jointly, and predicts the probability of mortality/discharge as functions of time. The idea is that each patient has a latent distribution of disease conditions, which we call topics. These conditions generate the measurements and notes and determine the patients' mortality. We derive a mean-field variational inference algorithm for this model. We fitted the SVTM with two outcomes on Medical Information Mart for Intensive Care III (MIMIC III) trauma patients data and obtained some important topics. Also, we demonstrated the relationships between these topics.**

***Index Terms*—Mortality Prediction, Survival Analysis, Topic Model**

I. Introduction

Trauma is the leading cause of death from age 15 to 49 worldwide. More than 5 million people die each year as a result of trauma [1]. Most trauma patients either recover after some period of time or they die very quickly. After admitted to intensive care units (ICUs), the majority of these deaths take place in the first hours or days, and much of the initial treatments and decision-making actions occur in the first minutes or hours after injury [2]. Not surprisingly, the ICU has been found to be one of the sites where medical errors are most likely to occur [3], [4]. Tools that can provide quick and accurate assessments of a patient's condition can provide immense value in helping physicians to make well-informed critical decisions, and ultimately to curb the mortality rate for trauma patients.

Several clinical methods have been developed to address this problem. Examples of the most widely used include Apache III [5], and SAPS II [6]. These score-based methods require physicians to gather a set of easily obtainable measurements from a patient and combine them in a specific way, ultimately using a logistic regression type procedure to obtain a mortality score and probability. Though the simplicity of these approaches allow for quick diagnostics of a patient, the results provided have often been found to be unsatisfactory [7], [8]. Ultimately, these methods fail to model the complex dependencies of human physiology, leading to under-fitting in many situations. In addition, by utilizing only easily obtainable measurements, many highly informative pieces of data are lost. For example, textual descriptions such as *bone fracture, strong cough*, and *spine injury*, cannot easily be quantified or included in the above methods but offer significant insight into a patient's condition.

To cover the limitations of the clinical methods, a few data mining approaches have recently been proposed for the same task [9], [10], [11]. These methods use state-of-the-art machine learning and data mining algorithms to incorporate both measurements and notes in predicting patient mortality, achieving much better performance than the score-based methods. However, there are three drawbacks of these methods. Firstly, they consider the mortality prediction problem as a classification problem. These methods predict either whether or not a patient will survive to discharge, or whether or not a patient will die before some pre-specified time after admission, for example 24 hours, 48 hours, etc. By modeling the problem as such, these methods ignore the information embedded in the length of time to death/discharge. Concretely, if three patients died at 25.2, 47.8 and 258 hours from the time of admission respectively, modeling the patient's survival to discharge as a binary classification task would consider them as three identical samples of patients that eventually died. When modeling the patient's mortality at 48 hours, the first two patients are considered the same. Yet, the time to death can often reflect important information about the initial state of the patient, as one can imagine that some conditions are inherently more time-sensitive than others. The continuous nature of this value makes setting any cut-off value as above very difficult. Secondly, these methods use topic models [12], [13] to analyze notes, but only treat the topics distributions generated as features to be used in another classifier. While this approach extracts useful information from the massive quantity of notes, it doesn't account for the inner dependency structure between the notes, measurements and time to death/discharge of a particular patient. We can imagine a patient as generating measurements and notes based on their underlying disease conditions, which will also ultimately contain information about mortality/discharge. Lastly, because they rely on additional classifiers, these methods do not distinguish between exogenous or endogenous features in the models. Exogenous features are the features that are external to the patient's stay and do not change [14] with the disease, such as demographic data and historical data. Endogenous features contain the opposite, namely the patient's in-hospital measurements, tests or doctor's notes. Without clearly distinguishing this difference, some models assume that the disease will impact exogenous features, which can lead to biased predictions.

In this paper, we propose a new graphical model, called the survival topic model (SVTM), which jointly models a patient's measurements, notes, and mortality/discharge. Furthermore, given a patient's measurements and/or notes, the SVTM can predict the probability of mortality/discharge as a function of time. The core idea is inspired by topic models such as latent Dirichlet allocation (LDA) [12], [13], [15], whereas in SVTM, instead of a distribution over textual topics, each patient is represented by a latent distribution of disease conditions, which we denote as *topics*. These topics determine the measurements and notes that a patient will generate as well as the patient's survival function for mortality. The model consists of three parts: a notes submodel, a measurements submodel, and a survival submodel. As we jointly model the notes, we can infer the textual representation of our topics by sampling from the notes submodel, in contrast to other latent disease states models [16]. We applied the SVTM on the publicly available MIMIC III dataset [17], which provides extensive Electronic Medical Records for ICU admissions at the Beth Israel Deaconess Medical Center between 2001 and 2012. Our results show that we obtained some interesting *topics* and their relationships for trauma patients.

The remainder of the paper is organized as follows: In Section II we introduce and formalize our model as well as provide a variational inference algorithm for learning it. In Section III we demonstrate its application to the trauma patient data in the MIMIC III dataset and provide some interesting clinical findings. In Section IV, we provide a discussion and conclusions.

II. Methods

In this section we first describe the key ideas and the framework of the SVTM. Then we introduce some basics of survival analysis, which is also the key to the SVTM. After that, we describe the variational method to estimate the parameters when modeling one survival outcome. Finally, we demonstrate its extension to deal with two or more outcomes, which is the model we have used to predict mortality and discharge for ICU patients. We note that when we talks about mortality/discharge, the term *discharge* usually means survival to discharge.

A. Survival Topic Model

Suppose there are D patients. For each patient d, we define $\mathbf{x}_d$ as the endogenous features, which are the in-hospital measurements and tests, $\mathbf{y}_d$ as the exogenous features, such as demographic data and historical data, W_d as the notes from doctors and nurses, δ_d as the outcome indicator, where $\delta_d = 1$ represents that the patient died in the hospital and $\delta_d = 0$ represents that the patient recovered and was discharged from the hospital, and T_d as the time to death or discharge. We propose that there are some underlying disease conditions, or *topics*, which represent the patient's condition. These topics can be heart disease, brain injury, coagulopathy, etc. From a generative point of view, the topics generate the measurements for the patient, impact the notes about the patients, and eventually, together with the patient's exogenous data, determine the outcome (death/discharge) and time-to-outcome of the patient. By estimating the distribution of these latent topics, the patient's condition can be assessed and the outcome for the patient can be predicted.

Suppose that K topics represent the latent conditions. Each patient has probabilities for the K topics that add up to 1. For a patient, there are N words in the notes and M measurements. The SVTM consists of three submodels, which are the notes submodel, the measurements submodel and the survival submodel. Each patient's condition arises from the following generative process:

1) Draw $\boldsymbol{\eta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2) (a) Notes submodel:
   For each word $W_n, n = 1, 2, \cdots, N$
   - Draw a topic assignment $Z_n \mid \boldsymbol{\eta} \sim$ Multi$(f(\boldsymbol{\eta}))$, where $f(\eta_i) = \frac{\exp \eta_i}{\sum_j \exp \eta_j}$.
   - Draw a word $W_n \mid Z_n, \beta_{1:K} \sim$ Multi(β_{z_n}).

   (b) Measurement submodel:
   For each measurement $x_m, m = 1, 2, \cdots, M$
   - Draw a measurement $x_m \mid \boldsymbol{\theta}_m, \boldsymbol{\eta}, \sigma_m^2 \sim \mathcal{N}(\boldsymbol{\theta}_m^T \boldsymbol{\eta}, \sigma_m^2)$.

   (c) Survival submodel:
   When modeling one outcome:
   - Draw the time to death T from the Cox model $T \mid \mathbf{y}, \boldsymbol{\eta}, h_0, \boldsymbol{\tau}, \boldsymbol{\gamma} \sim \text{Cox}(h_0, \boldsymbol{\tau}^T \mathbf{y} + \boldsymbol{\gamma}^T \boldsymbol{\eta})$.

   When modeling two outcomes:
   - Draw the time to death $T^{(1)}$ and time to discharge $T^{(2)}$ from the revised competing-risk Cox model $T^{(1)}, T^{(2)} \mid \mathbf{y}, \boldsymbol{\eta}, h_0^{(1)}, h_0^{(2)}, \boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)} \sim \text{Cox}(h_0^{(1)}, h_0^{(2)}, \boldsymbol{\tau}^{(1)T} \mathbf{y} + \boldsymbol{\gamma}^{(1)T} \boldsymbol{\eta}, \boldsymbol{\tau}^{(2)T} \mathbf{y} + \boldsymbol{\gamma}^{(2)T} \boldsymbol{\eta})$.
   - The earlier event time determines the outcome δ.

The latent topics for patients are $f(\boldsymbol{\eta})$, which adds up to 1. A diagram of the cure topic model is shown in Figure 1.

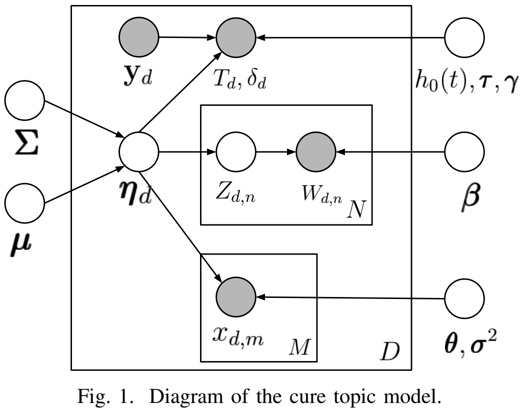

Fig. 1. Diagram of the cure topic model.

There are several benefits of using the SVTM. First, the topics are related not only to the notes, but also related to measurements and mortality/discharge outcomes. Thus the topic set can better represent the nature of the patient's condition. Also, the topics are assumed to be sampled from a multi-variate Normal distribution. With that, the relationships between topics can be learned from the covariance matrix $\boldsymbol{\Sigma}$, which can be useful when studying the relationships between injuries. Third, the measurements and Cox models include linear relationships with topics. By studying the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, the influence of each topic to measurements, mortality/discharge can be learned.

B. Survival Analysis

The key to the SVTM is survival analysis. Survival analysis [18], [14] is a branch of statistics for analyzing the expected duration of time until one or more events occur. The survival function is defined as $S(t) = P(T > t)$, which gives the probability that the time to the event is later than a given time t. We note that the survival function is basically $1 - F(t)$, where $F(t)$ is the cumulative distribution function of the time to the event. One important function in survival analysis is the hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \quad (1)$$

where $f(t)$ is the probability density function. The hazard function can be interpreted as the probability that the event will occur in the time interval $[t, t + \Delta t)$ given that the event has not occurred until time t. The relationship between $S(t)$ and $h(t)$ is

$$S(t) = \exp\left(-\int_0^t h(s)\mathrm{d}s\right). \quad (2)$$

An important survival model is the Cox model [19]. The Cox model defines the hazard function given data to be the product of a baseline function, which is a function of time, and a parametric function of the data. In the case of the survival submodel in the SVTM, the Cox model has the form

$$h(t \mid \mathbf{y}, \boldsymbol{\eta}) = h_0(t) \exp(\boldsymbol{\tau}^T \mathbf{y} + \boldsymbol{\gamma}^T \boldsymbol{\eta}),$$

where $h_0(t)$ is the baseline hazard function, and $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ are the vectors of weights for $\mathbf{y}$ and $\boldsymbol{\eta}$.

When modeling one survival outcome, the likelihood function for the survival submodel is as follows:

$$p(T, \delta) = h(T)^{\delta} S(T). \quad (3)$$

When $\delta = 1$, this means that the death event has been observed and T is the time to death. When $\delta = 0$, it means that the death event has not been observed before T and it will be unknown after T. In this case, the time T is also called the censoring time, which means that the observation is censored.

When modeling two outcomes, we adopt the competing risk framework [14] and define two hazard functions given different outcomes as follow

$$h^{(1)}(t)\Delta t = P(t \leq T < t + \Delta t, \delta = 1 \mid T \geq t),$$
$$h^{(2)}(t)\Delta t = P(t \leq T < t + \Delta t, \delta = 0 \mid T \geq t),$$

when $\Delta t \to 0$. Then we have the overall hazard function as $h(t) = h^{(1)}(t) + h^{(2)}(t)$. The likelihood function for the survival submodel is $p(T, \delta) = h^{(1)}(T)^{\delta} h^{(2)}(T)^{1-\delta} S(T)$. When $\delta = 1$, this means that mortality has been observed and T is the time to death. When $\delta = 0$, it means that the patient has recovered and was discharged from the hospital, and T is the time to discharge. We assume that the hazard function follows the form of the Cox model, which is

$$h^{(1)}(t \mid \mathbf{y}, \boldsymbol{\eta}) = h_0^{(1)}(t) \exp(\boldsymbol{\tau}^{(1)T} \mathbf{y} + \boldsymbol{\gamma}^{(1)T} \boldsymbol{\eta}), \quad (4)$$
$$h^{(2)}(t \mid \mathbf{y}, \boldsymbol{\eta}) = h_0^{(2)}(t) \exp(\boldsymbol{\tau}^{(2)T} \mathbf{y} + \boldsymbol{\gamma}^{(2)T} \boldsymbol{\eta}). \quad (5)$$

The hazard function $h^{(1)}(t)$ models the hazard of one outcome, which is death, and the hazard function $h^{(2)}(t)$ models the hazard of the other outcome, which is being discharged from the hospital.

C. Survival Topic Model with One Outcome

We will begin by discussing the variational method to estimate the parameters when modeling one outcome. In this case, for each patient, we have either the death time or the censoring time. In the model, $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2, h_0(t), \boldsymbol{\tau}, \boldsymbol{\gamma}$ are parameters. To estimate these parameters, we carry out approximate maximization-likelihood estimation using a variational expectation-maximization (EM) procedure.

1) Variational E-step: Given a patient's measurements, notes and outcome, the evidence lower bound has the following form

$$\begin{aligned}
&\log p(T, \delta, x_{1:M}, W_{1:N} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2, h_0(t), \boldsymbol{\tau}, \boldsymbol{\gamma}) \\
\geq &\mathbf{E}_q \left[\log p(\boldsymbol{\eta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \\
&+ \sum_{n=1}^{N} \left(\mathbf{E}_q \left[\log p(Z_n \mid \boldsymbol{\eta})\right] + \mathbf{E}_q \left[\log p(W_n \mid Z_n, \boldsymbol{\beta})\right]\right) \\
&+ \sum_{m=1}^{M} \mathbf{E}_q \left[\log p(x_m \mid \boldsymbol{\eta}, \boldsymbol{\theta}_m, \sigma_m^2)\right] \\
&+ \mathbf{E}_q \left[\log p(T, \delta \mid \boldsymbol{\eta}, h_0(t), \boldsymbol{\tau}, \boldsymbol{\gamma})\right] + H(q),
\end{aligned} \quad (6)$$

where the expectations are taken with respect to a variational distribution of the latent variables, and $H(q)$ denotes the entropy of that distribution. We use a factorized distribution for the variational distribution:

$$\begin{aligned}
&q(\eta_{1:K}, Z_{1:N} \mid \lambda_{1:K}, \nu_{1:K}^2, \phi_{1:N}) \\
&\quad = \prod_{i=1}^{K} q(\eta_i \mid \lambda_i, \nu_i^2) \prod_{n=1}^{N} q(Z_n \mid \phi_n).
\end{aligned} \quad (7)$$

The variational distribution of the discrete variables $Z_{1:N}$ is specified by the K-dimensional multinomial parameters $\phi_{1:N}$. The variational distribution of the continuous variables $\eta_{1:K}$ is described by K independent univariate Gaussian distributions $\mathcal{N}(\lambda_i, \nu_i^2)$.

In the E-step, we maximize the evidence lower bound with respect to the variational parameters $\lambda_{1:K}, \nu_{1:k}$ and $\phi_{1:N}$. The E-step of the first three terms in Equation (6) is the same as for the correlated topic model [13]. The first term in Equation (6) is:

$$\begin{aligned}\mathbf{E}_q\left[\log p(\boldsymbol{\eta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\right] =&\frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{K}{2}\log 2\pi - \\ &\frac{1}{2}\mathbf{E}_q\left[(\boldsymbol{\eta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}-\boldsymbol{\mu})\right],\end{aligned}$$

where $\mathbf{E}_q\left[(\boldsymbol{\eta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}-\boldsymbol{\mu})\right] = \mathrm{Tr}(\mathrm{diag}(\boldsymbol{\nu}^2)\boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\lambda}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}-\boldsymbol{\mu})$. Because $Z_n \mid \boldsymbol{\eta} \sim \mathrm{Multi}(f(\boldsymbol{\eta}))$ and $f(\eta_i) = \frac{\exp\eta_i}{\sum_j \exp\eta_j}$, the second term in Equation (6) is:

$$\mathbf{E}_q\left[\log p(Z_n \mid \boldsymbol{\eta})\right] = \mathbf{E}_q\left[\boldsymbol{\eta}^T Z_n\right] - \mathbf{E}_q\left[\log\left(\sum_{i=1}^{K}\exp(\eta_i)\right)\right].$$

To preserve the lower bound on the log probability, we bound the log normalizer from above with a Taylor expansion,

$$\mathbf{E}_q\left[\log\left(\sum_{i=1}^{K}\exp(\eta_i)\right)\right] \leq \zeta^{-1}\sum_{i=1}^{K}\mathbf{E}_q\left[\exp(\eta_i)\right] - 1 + \log\zeta,$$

where ζ is a new variational parameter. The expectation $\mathbf{E}_q\left[\exp(\eta_i)\right]$ is the mean of a log normal distribution, thus $\mathbf{E}_q\left[\exp(\eta_i)\right] = \exp(\lambda_i + \nu_i^2/2)$ for $i = 1, \cdots, K$. Thus, we have

$$\begin{aligned}\mathbf{E}_q\left[\log p(Z_n \mid \boldsymbol{\eta})\right] \geq &\sum_{i=1}^{K}\lambda_i\phi_{n,i} - \zeta^{-1}\left(\sum_{i=1}^{K}\exp(\lambda_i + \nu_i^2/2)\right) \\ &+ 1 - \log\zeta. \qquad (8)\end{aligned}$$

The third term of Equation (6) is given by

$$\mathbf{E}_q\left[\log p(W_n \mid Z_n, \boldsymbol{\beta})\right] = \sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,W_n},$$

and the fourth term of Equation (6) is:

$$\begin{aligned}\mathbf{E}_q\left[\log p(x_m \mid \boldsymbol{\eta}, \boldsymbol{\theta}_m, \sigma_m^2)\right] = &-\frac{1}{2}\log(2\pi\sigma_m^2) - \\ &\frac{1}{2\sigma_m^2}\mathbf{E}_q\left[(x_m - \boldsymbol{\theta}_m^T\boldsymbol{\eta})^2\right],\end{aligned}$$

where $\mathbf{E}_q\left[(x_m - \boldsymbol{\theta}_m^T\boldsymbol{\eta})^2\right] = (x_m - \boldsymbol{\theta}_m^T\boldsymbol{\lambda})^2 + \boldsymbol{\theta}_m^T(\mathrm{diag}(\boldsymbol{\nu}^2))\boldsymbol{\theta}_m$. Based on Equations (2) and (3), the fifth term of Equation (6) is given by

$$\begin{aligned}&\mathbf{E}_q\left[\log p(T, \delta \mid \boldsymbol{\eta}, h_0(t), \boldsymbol{\tau}, \boldsymbol{\gamma})\right] \\ =&\delta\mathbf{E}_q\left[\log h(T)\right] - \mathbf{E}_q\left[\int_0^T h(s)\mathrm{d}s\right], \qquad (9)\end{aligned}$$

where $\mathbf{E}_q\left[\log h(T)\right] = \log(h_0(T)) + \boldsymbol{\tau}^T\mathbf{y} + \boldsymbol{\gamma}^T\boldsymbol{\lambda}$ and $\mathbf{E}_q\left[\int_0^T h(s)\mathrm{d}s\right] = H_0(T) \cdot \exp(\boldsymbol{\tau}^T\mathbf{y}) \cdot \mathbf{E}_q\left[\exp\left(\sum_{i=1}^K \gamma_i\eta_i\right)\right]$, in which $H_0(t)$ is the cumulative hazard function for $h_0(t)$. Because the η_i's are independent of each other, we have

$$\mathbf{E}_q\left[\exp\left(\sum_{i=1}^{K}\gamma_i\eta_i\right)\right] = \prod_{i=1}^{K}\exp\left(\gamma_i\lambda_i + \frac{{\gamma_i}^2\nu_i^2}{2}\right),$$

based on the mean of a log normal distribution. Thus we have

$$\mathbf{E}_q\left[\int_0^T h(s)\mathrm{d}s\right] = H_0(T) \cdot \exp\left(\boldsymbol{\tau}^T\mathbf{y} + \boldsymbol{\gamma}^T\boldsymbol{\lambda} + \frac{(\boldsymbol{\gamma}^2)^T\boldsymbol{\nu}^2}{2}\right).$$

The last term in Equation (6) is the entropy of the variational distribution:

$$H(q) = \sum_{i=1}^{K}\frac{1}{2}(\log\nu_i^2 + \log 2\pi + 1) - \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i}.$$

In the E-step, we maximize the bound with respect to the variational parameters $\lambda_{1:K}, \nu_{1:k}, \phi_{1:N}$ and ζ. We use a coordinate ascent algorithm to iteratively maximize the bound with respect to the parameters. First, we maximize Equation (6) with respect to ζ, using the second bound in Equation (8). We take the derivative with respect to ζ and set it to 0, to obtain

$$\hat{\zeta} = \sum_{i=1}^{K}\exp(\lambda_i + \nu_i^2/2).$$

Second, we maximize with respect to ϕ_n, to obtain

$$\hat{\phi}_{n,i} \propto \exp(\lambda_i)\beta_{i,W_n}, \quad i = 1, \cdots, K.$$

Third, we maximize with respect to λ_i. Because there is no closed form solution, we use a gradient descent algorithm with derivative

$$\begin{aligned}\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\lambda}} = &-\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}-\boldsymbol{\mu}) + \sum_{n=1}^{N}\phi_{n,1:K} - \frac{N}{\zeta}\exp(\boldsymbol{\lambda} + \boldsymbol{\nu}^2/2) + \\ &\sum_{m=1}^{M}\frac{1}{\sigma_m^2}(x_m - \boldsymbol{\theta}_m^T\boldsymbol{\lambda})\boldsymbol{\theta}_m + \delta\boldsymbol{\gamma} - \\ &H_0(T) \cdot \exp\left(\boldsymbol{\tau}^T\mathbf{y} + \boldsymbol{\gamma}^T\boldsymbol{\lambda} + \frac{(\boldsymbol{\gamma}^2)^T\boldsymbol{\nu}^2}{2}\right) \cdot \boldsymbol{\gamma}. \qquad (10)\end{aligned}$$

Finally, we maximize with respect to ν_i^2. We use Newton's method for each coordinate, with the constraints $\nu_i^2 > 0$:

$$\begin{aligned}\frac{\mathrm{d}L}{\mathrm{d}\nu_i^2} = &-\Sigma_{ii}^{-1}/2 - \frac{N}{2\zeta}\exp(\lambda_i + \nu_i^2/2) + 1/(2\nu_i^2) - \\ &\sum_{m=1}^{M}\left(\frac{\theta_{m,i}^2}{2\sigma_m^2}\right) - \\ &H_0(T)\exp\left(\boldsymbol{\tau}^T\mathbf{y} + \boldsymbol{\gamma}^T\boldsymbol{\lambda} + \frac{(\boldsymbol{\gamma}^2)^T\boldsymbol{\nu}^2}{2}\right) \cdot {\gamma_i}^2/2.\end{aligned}$$

2) Variational M-step: In the M-step, we maximize the bound with respect to the model parameters. The parameters are $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2, h_0(t), \boldsymbol{\tau}$ and $\boldsymbol{\gamma}$. The maximization with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ amounts to the maximum likelihood estimation of the multivariate Gaussian, thus we have $\hat{\boldsymbol{\mu}} = \frac{1}{D}\sum_{d=1}^{D}\boldsymbol{\lambda}_d$, and $\hat{\boldsymbol{\Sigma}} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{diag}(\boldsymbol{\nu}_d^2) + (\boldsymbol{\lambda}_d - \hat{\boldsymbol{\mu}})(\boldsymbol{\lambda}_d - \hat{\boldsymbol{\mu}})^T$. The maximization with respect to β amounts to the maximum likelihood estimation of the multinomial distribution, thus

$$\hat{\beta}_i \propto \sum_{d=1}^{D}\phi_{d,i}n_d.$$

The maximization with respect to $\boldsymbol{\theta}_m$ and σ_m^2 amounts to the maximum likelihood estimation of the linear regression. Letting $\mathbf{H}$ be the $D \times (K+1)$ matrix whose rows are $\boldsymbol{\eta}_d^T$, we have

$$\hat{\boldsymbol{\theta}}_m = \mathbf{E}_q \left[\mathbf{H}^T\mathbf{H}\right]^{-1} \mathbf{E}_q(\mathbf{H})^T \mathbf{x}_m,$$
$$\hat{\sigma_m}^2 = \frac{1}{D}\left\{\mathbf{x}_m^T\mathbf{x}_m - \mathbf{x}_m^T\mathbf{E}_q\left[\mathbf{H}\right]\mathbf{E}_q\left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{E}_q\left[\mathbf{H}\right]^T\mathbf{x}_m\right\},$$

and we have $\mathbf{E}_q\left[\mathbf{H}\right] = \Lambda$ and $\mathbf{E}_q\left[\mathbf{H}^T\mathbf{H}\right] = \sum_{d=1}^D \mathbf{E}_q\left[\boldsymbol{\eta}_d\boldsymbol{\eta}_d^T\right]$ $= \sum_{d=1}^D \left(\boldsymbol{\lambda}_d\boldsymbol{\lambda}_d^T + \text{diag}(\boldsymbol{\nu}_d^2)\right)$, where Λ is the matrix whose rows are $\boldsymbol{\lambda}_d^T$. The maximization with respect to $h_0(t), \boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ amounts to the maximum likelihood estimation of the Cox models. We let $\hat{\boldsymbol{\eta}} = (\mathbf{y}^T, \boldsymbol{\eta}^T)^T$, $\hat{\boldsymbol{\lambda}} = (\mathbf{y}^T, \boldsymbol{\lambda}^T)^T$, $\hat{\boldsymbol{\gamma}} = (\boldsymbol{\tau}^T, \boldsymbol{\gamma}^T)^T$ and $\hat{\boldsymbol{\nu}}^2 = (\mathbf{0}^T, {\boldsymbol{\nu}^2}^T)^T$. We note that $\mathbf{E}_q\left[\hat{\boldsymbol{\eta}}\right] = \hat{\boldsymbol{\lambda}}$. Thus equation (9) becomes

$$\mathbf{E}_q\left[\log p(T, \delta \mid \boldsymbol{\eta}, h_0(t), \boldsymbol{\tau}, \boldsymbol{\gamma})\right] = \delta\left(\log h_0(T) + \hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\lambda}}\right) - H_0^{(1)}(T)\cdot\exp\left(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\lambda}} + \frac{(\hat{\boldsymbol{\gamma}}^2)^T\hat{\boldsymbol{\nu}}^2}{2}\right). \tag{11}$$

Setting the derivative of Equation (11) with respect to $h_0(t)$ to 0, we have

$$\hat{h}_0(t) = \sum_{d=1}^D \frac{\delta_d I(T_d = t)}{\sum_{i=1}^D\left[\exp\left(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\lambda}_i} + \frac{(\hat{\boldsymbol{\gamma}}^2)^T\hat{\boldsymbol{\nu}_i}^2}{2}\right)\right]Y_d(t)}. \tag{12}$$

where $I(\cdot)$ is a indicator function, which is 1 when the input is true and 0 otherwise, and $Y_d(t)$ is $I(T_d \geq t)$. To maximize the likelihood with respect to $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$, we need to maximize $\hat{\boldsymbol{\gamma}}$. There is no closed form solution, thus we use the Newton-Raphson' method. The score function $U(\hat{\boldsymbol{\tau}})$ can be found by taking the derivative of Equation (11) respect to $\hat{\boldsymbol{\tau}}$ and plugging in $\hat{h}_0(t)$. Letting $U_k(\hat{\boldsymbol{\tau}})$ be the k-th item in $U(\hat{\boldsymbol{\tau}})$, we have

$$U_k(\hat{\boldsymbol{\gamma}}) = \sum_{d=1}^D \delta_d\mathbf{E}_q\left[\hat{\eta}_{d,k}\right] - \sum_{d=1}^D \frac{\sum_{i=1}^D \mathbf{E}_q\left[\hat{\eta}_{i,k}\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}{\sum_{i=1}^D \mathbf{E}_q\left[\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}.$$

The information matrix is the negative of the matrix of second derivatives of the Equation (11) and is given by $\mathbf{I}(\hat{\boldsymbol{\gamma}}) = \left[I_{g,b}(\hat{\boldsymbol{\gamma}})\right]_{K\times K}$, with the (g,b)th element given by

$$I_{g,b}(\hat{\boldsymbol{\gamma}}) = \sum_{d=1}^D \frac{\sum_{i=1}^D \mathbf{E}_q\left[\hat{\eta}_{i,g}\hat{\eta}_{i,b}\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}{\sum_{i=1}^D \mathbf{E}_q\left[\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)} - \sum_{d=1}^D \left[\frac{\sum_{i=1}^D \mathbf{E}_q\left[\hat{\eta}_{i,g}\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}{\sum_{i=1}^D \mathbf{E}_q\left[\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}\right] \cdot \left[\frac{\sum_{i=1}^D \mathbf{E}_q\left[\hat{\eta}_{i,b}\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}{\sum_{i=1}^D \mathbf{E}_q\left[\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right]Y_i(T_d)}\right].$$

After having obtained $U(\hat{\boldsymbol{\gamma}})$ and $\mathbf{I}(\hat{\boldsymbol{\gamma}})$, we can find $\hat{\boldsymbol{\gamma}}$ by an iterative process

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\gamma}}_{k-1} + \mathbf{I}(\hat{\boldsymbol{\gamma}}_{k-1})^{-1}U(\hat{\boldsymbol{\gamma}}_{k-1}).$$

Letting $\tilde{\boldsymbol{\lambda}}_i = \hat{\boldsymbol{\lambda}_i} + \hat{\boldsymbol{\gamma}} \cdot \hat{\boldsymbol{\nu}_i}^2$, we have $\mathbf{E}_q\left[\hat{\boldsymbol{\eta}}_i\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right] = \tilde{\boldsymbol{\lambda}}_i\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\lambda}_i} + \frac{(\hat{\boldsymbol{\gamma}}^2)^T\hat{\boldsymbol{\nu}_i}^2}{2})$, and $\mathbf{E}_q\left[\hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i^T\exp(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\eta}}_i)\right] = \left(\tilde{\boldsymbol{\lambda}}_i\tilde{\boldsymbol{\lambda}}_i^T + \text{diag}(\hat{\boldsymbol{\nu}_i}^2)\right)\exp\left(\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\lambda}_i} + \frac{(\hat{\boldsymbol{\gamma}}^2)^T\hat{\boldsymbol{\nu}_i}^2}{2}\right)$.

D. Survival Topic Models with Two or More Outcomes

When using the SVTM to predict two outcomes, we need to include two hazard functions (4) and (5), to reflect two outcomes. The parameters for the survival submodel are $h_0^{(1)}(t)$, $h_0^{(2)}(t)$,$\boldsymbol{\tau}^{(1)}$, $\boldsymbol{\tau}^{(2)}$, $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$. The bound in Equation (9) becomes

$$\begin{aligned}&\mathbf{E}_q\left[\log p(T, \delta \mid \boldsymbol{\eta}, h_0^{(1)}(t), h_0^{(2)}(t), \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)})\right]\\ =&\delta\mathbf{E}_q\left[\log h^{(1)}(T)\right] + (1-\delta)\mathbf{E}_q\left[\log h^{(2)}(T)\right] + \\ &\mathbf{E}_q\left[\log S(T)\right]\\ =&\delta\mathbf{E}_q\left[\log h^{(1)}(T)\right] - \mathbf{E}_q\left[\int_0^T h^{(1)}(s)\mathrm{d}s\right] + \\ &(1-\delta)\mathbf{E}_q\left[\log h^{(2)}(T)\right] - \mathbf{E}_q\left[\int_0^T h^{(2)}(s)\mathrm{d}s\right],\end{aligned} \tag{13}$$

which is the sum of the log-likelihoods of two Cox models. When maximizing the bound with respect to the variational variables in the E-step, we need only to add one more survival term in each of the first order derivatives of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}^2$. When maximizing the bound with respect to the model parameters in the M-step, two Newton-Raphson iterations are needed to update two sets of parameters in two hazard functions. The natural splitting of the survival submodel makes it easy to extend to multiple outcomes. If we want to model more than two outcomes, we can add more hazard functions, each with a indicator term, and ensure the sum of the indicator terms equal to 1. However, it will also cause problems without modifications. For example, for the Cox model with $\delta = 1$, it models the patients' time to death, and it treats the patients' time to discharge as the censoring time. The underlying assumption is that the times for patients who died and for patients who were discharged belong to the same distribution. Similarly, the Cox model with $\delta = 0$ models the patients' time to recovery, and it assumes that the discharge times for patients who died and patients who recovered belong to the same distribution. These underlying assumptions will bias the estimation. Thus we adopted the approach in [20]. In the Cox model with $\delta = 1$, we adjust the censoring time of recovering patients to be T_∞, which is a large number compared with the distribution of time to death. This means that the recovering patients will have a much later death time compared with the patients who died in the hospital. Similarly, in the Cox model with $\delta = 0$, we adjust the censoring time of dying patients to be T_∞, which means that the dying patients have an infinite time to recover.

E. Prediction

For a new patient, with his/her notes and/or measurements, the estimation of $\boldsymbol{\eta}$ is its posterior expectation $\boldsymbol{\lambda}$, and the estimation of topic distributions is $f(\boldsymbol{\lambda})$. If both notes and measurements are obtained, $\boldsymbol{\lambda}$ can be obtained by maximizing the bound using Equation (10) without the survival submodel. If either notes or measurements are observed, $\boldsymbol{\lambda}$ can be obtained by maximizing the bound using Equation (10) with only the notes submodel or measurements submodel. With $\boldsymbol{\lambda}$ and $\mathbf{y}$, $P(T_{death} > t|\boldsymbol{\lambda}, \mathbf{y}) = S^{(1)}(t|\boldsymbol{\lambda}, \mathbf{y})$ and $P(T_{discharge} > t|\boldsymbol{\lambda}, \mathbf{y}) = S^{(2)}(t|\boldsymbol{\lambda}, \mathbf{y})$ can be obtained.

F. Choice of the Number of Topics

The number of topics is a hyper-parameter in the model. To choose the number of topics in supervised problem, cross-validation is a simple technique to use. However, it is not easy to apply to the SVTM directly, which provides the prediction as a function of time. To apply cross-validation, we calculate $\log \frac{1-S^{(1)}(T_\infty)}{1-S^{(2)}(T_\infty)}$, where T_∞ is a very large time compared with the model, as the log-odds ratio. Binary classification results can be obtained by thresholding the log-odds ratio, and cross-validation can be applied to obtain the number of topics and the threshold.

III. Empirical Results

A. Datasets

We applied the SVTM with two outcomes, one for mortality and the other for discharge, to the trauma patient data in the MIMIC-III dataset. The trauma patients were selected using the ICD-9 code. The measurements, clinical notes, demographic data, outcome, and time to outcome were obtained for every admission to train the model. To gauge the initial state of an individual admission, and to eliminate discrepancies and missing data due to different methodologies of administering physicians, we aggregated each of the endogenous measurements over the first 4 hours of an admission. Specifically, we took the average value for each measurement and concatenated all of the clinical notes together. As many of the measurements are redundant and excessively sparse, we selected 21 of the most common features, such as blood pressure, temperature, and respiratory rate, to use as our representative set. This ultimately overlaps 65% of the measurements required by APACHE III [5]. From the demographic data, we extracted 4 key exogenous features: age, gender, race, and weight.

Preprocessing of the notes is very important for good topic performance. Notes were pre-processed by first removing formatting, numbers and non-punctuation symbols. Due to the large presence of negation statements in the clinical notes, such as *has no sign*, and *unchanged*, we implemented a simple negation removal algorithm where phrases that are found between a negation keyword and the next ending punctuation are eliminated, unless that phrase also contains a negation keyword. We followed this by removal of punctuations, morphological affixes [21] and stopwords. Our stopword dictionary includes the most common English stopwords as well as a few specific clinical-related and trauma-related stopwords, which we found occur in many admission notes but under rather uninformative contexts. Examples of clinical-related stopwords include *doctor, nurse, measure*, and *plan*. Examples of clinical-related stopwords include *trauma* and *injury*. As a heuristic, we limited our vocabulary to only bi-grams because we found that most of the meaningful clinical terms fall into this category. To shrink our vocabulary size, we filtered the remaining bi-grams by document frequency, keeping only those that have occurred in at least 5 documents and in at most 90% of all documents. Finally, for each document, we kept only the top 500 bi-grams as measured by term frequency-inverse document frequency (TF-IDF).

Our final trauma dataset consists of 2,471 ICU trauma patients, with 21 endogenous measurements, 4 exogenous measurements and 56,960 bi-grams.

B. Model Outcomes

1) Number of Topics: The number of topics was chosen by the cross-validation described in Section II-F. 3-times 3-fold cross-validation was used to select the best number of topics. For each held-out validation dataset, we calculated the sum of the sensitivity and specificity for the prediction results. The average and the error bar for held-out sums of the sensitivity and specificity for the trauma patient data is shown in Figure 2, where the best number of topics for trauma patient data can be seen to be 25.

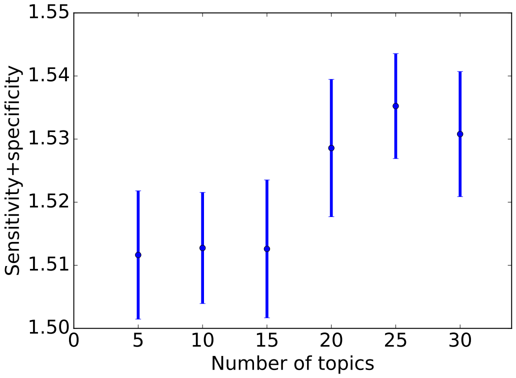

Fig. 2. Average held-out sum of sensitivity and specificity for trauma patient data. When the number of topics is 25, the best performance is achieved.

2) Predictions: When making predictions using the SVTM we obtain two *survival functions* for each admission: one for the outcome of death and the other for the outcome of discharge. These functions can respectively tell us $P(T_{death} \leq t)$ and $P(T_{discharge} \leq t)$ by using $1 - S(t)$. We note that $P(T_{death} \leq t)$ and $P(T_{discharge} \leq t)$ respectively means the probability of death and the probability to discharge before any given time t, which are functions of time. In Figure 3 we have plotted the predicted $P(T_{death} \leq t)$ and $P(T_{discharge} \leq t)$ for 4 of the admissions in the validation set using the SVTM. Each subfigure lists the true outcome and time to outcome for the patient, at the top. We point out that these functions are predicted from the first 4 hours of aggregated measurements. In Figure 3(a), the patient was discharged at hour 717.32. We see that the predicted $P(T_{death} \leq t)$ slowly increases while $P(T_{discharge} \leq t)$ quickly increases. As these functions directly correspond to probabilities, these growth rates can be understood as the model predicting that the probability for patient death is low while the probability for patient discharge is high. In addition, relative to the time of the true outcome at hour 717.32, we see that our model predicts that the patient will be discharged at or before this time with probability of 0.7, while the probability that they die at or before this time is only around 0.1. Similarly, for Figure 3(b), the predicted probability of death increases very slowly while the predicted probability of discharge increases rapidly. This also indicates that while time goes on, the probability of discharge becomes much larger than the probability of death. It predicts the patient to be in good condition and to recover and be discharged soon. The patient actually was discharged after 102.58 hours after admission. Using the same analysis for Figure 3(c), we see that the predicted probability of death at or before the true time of 350 is around 0.7, while the predicted probability of discharge is around 0.1, reflecting the true outcome of death rather nicely. Similarly, Figure 3(d) also provides a good prediction of the probabilities of death and discharge. Also, comparing (a) and (b) in Figure 3, we notice that patient in (b) was discharged earlier than the patient in (a), which corresponds to the behaviors of the prediction functions.

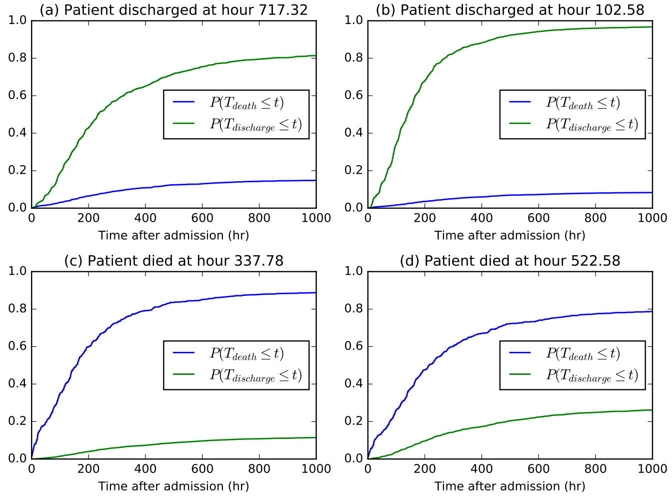

Fig. 3. Prediction of mortality/discharge as a function of time using the SVTM.

We also compared these curves for the basic competing risk Cox model [14]. With the competing risk Cox model, we use the exogenous and endogenous measurements directly in the model, without using topics. We found that the SVTM performs uniformly better than the competing risk Cox model. For the same patients, their predicted probabilities of death and discharge are shown in Figure 4. Comparing with Figure 3, the separation between the two probability functions are much closer.

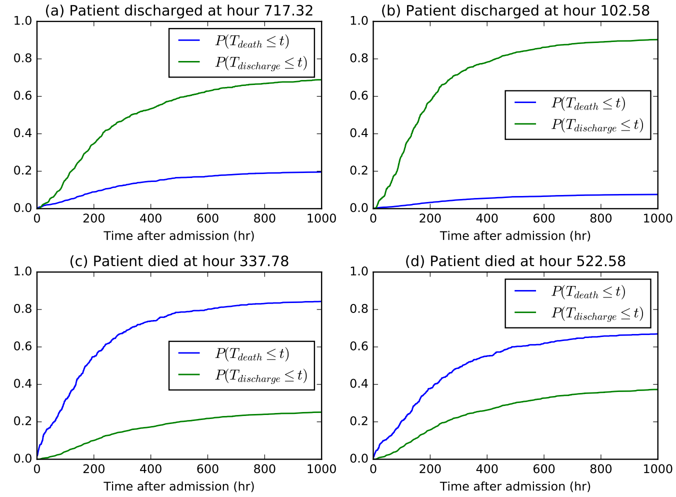

Fig. 4. Prediction of mortality/discharge as a function of time using the competing risk Cox model.

In Figure 5 we show the obtained *hazard functions* for both death and discharge for the same patients. The hazard function is defined in Equation (1). A larger hazard rate for death predicts that the patient is more likely to die, while a larger hazard rate for discharge predicts the patient is more likely to survive. Comparing the subfigures in the first row, we can see that even though both patients were discharged, their predicted hazard rates for death and discharge are quite different. The patient with an earlier discharge time has a larger and more rapidly increasing discharge hazard compared to the patient who was discharged later. Similarly, for the patients who died, the patient with an earlier death time has a larger and faster increasing death hazard.

Such survival and hazard curves can provide physicians with a more comprehensive insight into patients' conditions compared to point estimates provided by most other methods.

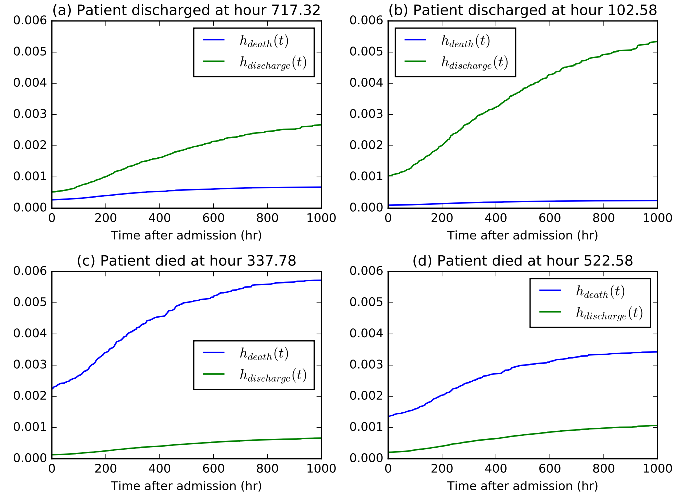

Fig. 5. The hazard function of mortality/discharge as a function of time using the SVTM.

C. Clinical Results

After applying the SVTM with 25 topics on the whole trauma patient data, we identified some interesting topics using the keywords. Some chosen topics with their corresponding keywords are shown in Figure 6. The links between topics indicate that the absolute covariances between topics are higher than the threshold 10^{-5}.

According to the figure, different types of injury have been picked out as topics from the notes. Some injuries are highly correlated with each other. This can be explained that these correlated injuries are located nearby in the human body, thus patients may have these multiple injuries simultaneously, and undergo multiple inspections for these injuries.

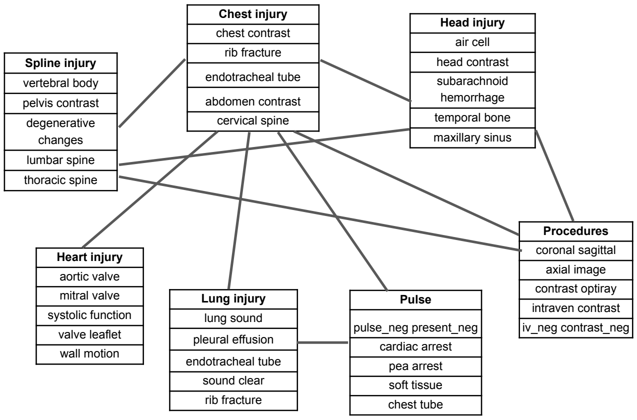

Fig. 6. The chosen topics and corresponding keywords.

IV. Discussion and Conclusions

In this paper we have proposed a new method, the survival topic model (SVTM), to predict the mortality and discharge for trauma patients by combining heterogeneous data, with consideration of their inner relationships. Specifically, the survival topic model jointly models a patient's measurements, notes and mortality/discharge. We claim that there are several major gains from the SVTM. First, it models the mortality/discharge time as continuous values. In contrast to binary classification, there is no need to give ad-hoc cutoffs to time values. Also, when making predictions, the probabilities of mortality and discharge are provided as functions of time. This can provide more information to doctors about the severity of the patient's condition. Second, the topics generated from the SVTM are not only related to the texts, but also related to measurements and to the mortality/discharge outcomes. These topics can better represent the condition of a patient than when only using text. Also, comparing with other state models, such as the hidden Markov model, the topics generated from the SVTM can be easily categorized by their keywords. Third, the SVTM is a supervised model, even though it does not produce binary predictions. With a supervised model framework, we can use cross-validation to select hyper-parameters.

References

[1] World Health Organization *et al.*, "Injuries and violence: the facts 2014," 2014.

[2] O. Blow, L. Magliore, J. A. Claridge, K. Butler, and J. S. Young, "The golden hour and the silver day: Detection and correction of occult hypoperfusion within 24 hours improves outcome from major trauma," *Journal of Trauma and Acute Care Surgery*, vol. 47, no. 5, p. 964, 1999.

[3] D. J. Cullen, B. J. Sweitzer, D. W. Bates, E. Burdick, A. Edmondson, and L. L. Leape, "Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units," *Critical care medicine*, vol. 25, no. 8, pp. 1289–1297, 1997.

[4] L. B. Andrews, C. Stocking, T. Krizek, L. Gottlieb, C. Krizek, T. Vargish, and M. Siegler, "An alternative strategy for studying adverse events in medical care," *The Lancet*, vol. 349, no. 9048, pp. 309–313, 1997.

[5] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, and A. Damiano, "The APACHE III prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults." *Chest Journal*, vol. 100, no. 6, pp. 1619–1636, 1991.

[6] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, 1993.

[7] J.-L. Vincent and R. Moreno, "Clinical review: scoring systems in the critically ill," *Crit Care*, vol. 14, no. 2, p. 207, 2010.

[8] M. T. Keegan, B. Gali, J. Y. Findlay, J. K. Heimbach, D. J. Plevak, and B. Afessa, "APACHE III outcome prediction in patients admitted to the intensive care unit after liver transplantation: a retrospective cohort study," *BMC Surgery*, vol. 9, no. 1, p. 1, 2009.

[9] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 75–84.

[10] K. L. Caballero Barajas and R. Akella, "Dynamically modeling patient's health state from electronic medical records: A time series approach," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 69–78.

[11] L.-W. H. Lehman, M. Saeed, W. J. Long, J. Lee, and R. G. Mark, "Risk stratification of icu patients using topic models inferred from unstructured progress notes." in *AMIA*. Citeseer, 2012.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, pp. 17–35, 2007.

[14] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2011, vol. 360.

[15] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems*, 2008, pp. 121–128.

[16] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 85–94.

[17] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Fen, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, 2016.

[18] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media, 2005.

[19] D. R. Cox and D. Oakes, *Analysis of Survival Data*. CRC Press, 1984, vol. 21.

[20] Y. Zhang, B. J. Daigle, M. Cohen, and L. Petzold, "A cure time model for joint prediction of outcome and time-to-outcome," in *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1117–1122.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215 PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.