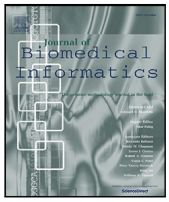
Journal of Biomedical Informatics 153 (2024) 104638

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

MixEHR-SurG: A joint proportional hazard and guided topic model for inferring mortality-associated topics from electronic health records

Yixuan Li [a,b], Archer Y. Yang [a,b,d], Ariane Marelli [c], Yue Li [b,d,*]

[a] Department of Mathematics and Statistics, McGill University, Montreal, Canada
[b] Mila - Quebec AI institute, Montreal, Canada
[c] McGill Adult Unit for Congenital Heart Disease (MAUDE Unit), McGill University of Health Centre, Montreal, Canada
[d] School of Computer Science, McGill University, Montreal, Canada

ARTICLE INFO

Keywords:
Electronic health records
Survival analysis
Topic modeling

ABSTRACT

Survival models can help medical practitioners to evaluate the prognostic importance of clinical variables to patient outcomes such as mortality or hospital readmission and subsequently design personalized treatment regimes. Electronic Health Records (EHRs) hold the promise for large-scale survival analysis based on systematically recorded clinical features for each patient. However, existing survival models either do not scale to high dimensional and multi-modal EHR data or are difficult to interpret. In this study, we present a supervised topic model called MixEHR-SurG to simultaneously integrate heterogeneous EHR data and model survival hazard. Our contributions are three-folds: (1) integrating EHR topic inference with Cox proportional hazards likelihood; (2) integrating patient-specific topic hyperparameters using the PheCode concepts such that each topic can be identified with exactly one PheCode-associated phenotype; (3) multi-modal survival topic inference. This leads to a highly interpretable survival topic model that can infer PheCode-specific phenotype topics associated with patient mortality. We evaluated MixEHR-SurG using a simulated dataset and two real-world EHR datasets: the Quebec Congenital Heart Disease (CHD) data consisting of 8211 subjects with 75,187 outpatient claim records of 1767 unique ICD codes; the MIMIC-III consisting of 1458 subjects with multi-modal EHR records. Compared to the baselines, MixEHR-SurG achieved a superior dynamic AUROC for mortality prediction, with a mean AUROC score of 0.89 in the simulation dataset and a mean AUROC of 0.645 on the CHD dataset. Qualitatively, MixEHR-SurG associates severe cardiac conditions with high mortality risk among the CHD patients after the first heart failure hospitalization and critical brain injuries with increased mortality among the MIMIC-III patients after their ICU discharge. Together, the integration of the Cox proportional hazards model and EHR topic inference in MixEHR-SurG not only leads to competitive mortality prediction but also meaningful phenotype topics for in-depth survival analysis. The software is available at GitHub: https://github.com/li-lab-mcgill/MixEHR-SurG.

1. Introduction

The rapid adoption of Electronic Health Records (EHRs) [1] enables systematic investigation of phenotypes and their comorbidity [2–5]. EHR include rich phenotypic observations of patient subjects from physician and nursing notes to diagnostic codes and prescription. One important application of EHR is to *detect* and *understand* the risk of adverse events such as death based on the recent health history of the patient [6]. Accurate detection will enable efficient resource allocation for the high-risk patients and can cost-effectively save many lives [7]. Understanding the mortality risk is equally important as it can inform practitioners for subsequent intervention. Many machine learning methods were developed recently for predicting adverse events such as mortality and unplanned emergency re-admission [8–11]. However, the progress on this front has been hindered by the lack of an interpretable approach that can distill interpretable phenotypic concepts relevant to the outcome of interest while having competitive detection precision on those events.

Predicting mortality events using EHRs has been a long-standing challenge due to the large search space of causal events. Survival analysis models have evolved beyond traditional Cox proportional hazards (PH) models [12] to include sophisticated techniques capable of handling complex, high-dimensional data. For instance, the kernel Cox regression method [13] extends the Cox model by incorporating kernel methods, allowing for a nuanced understanding of patient

* Corresponding author at: School of Computer Science, McGill University, Montreal, Canada.
E-mail addresses: archer.yang@mcgill.ca (A.Y. Yang), ariane.marelli@mcgill.ca (A. Marelli), yueli@cs.mcgill.ca (Y. Li).

https://doi.org/10.1016/j.jbi.2024.104638
Received 20 December 2023; Received in revised form 7 March 2024; Accepted 3 April 2024
Available online 15 April 2024

survival in relation to a broader range of clinical factors. Random Survival Forests [14] and LASSO-penalized Cox models [15] were developed for high-dimensional data, enhancing the predictive accuracy and interpretability of survival outcomes. These advancements represent significant strides in survival analysis, enabling more precise and comprehensive evaluations of patient data. While these have set the foundational benchmark, they sometimes sidestep the complex, patient-specific nuances. Recently, deep learning methods like DeepSurv [9] and neural multi-task logistic regression [8] have entered the fray, harnessing the power of neural networks to predict patient survival with greater accuracy. However, these methods are hard to interpret and often require external approaches to explain their prediction [16–18].

Topic models are a family of Bayesian models [19]. In our context, we treat patients as documents and their EHR codes as tokens. Topic models infer the topic mixture of each document, the latent topic for each token, and a set of latent topic distributions. Here the topic mixture represents the mixture of phenotype of the patient and the set of topic distributions represent the set of phenotypic distributions over the EHR codes. Despite the simple generative process, topic models are effective in distilling phenotype concepts from the EHR data [20–22]. Recently, we developed a guided topic model called Mixture of EHR Guided (MixEHR-G) [23], which specifies the topic hyperparameters based on the high-level phenotype codes (i.e., PheCodes) observed in the patients. As a result, each topic is identifiable with known phenotype codes, thereby improving the down-stream analysis. However, MixEHR-G does not have the ability to predict mortality. To address this challenge, we aim to develop a model that leverages EHR data for two primary purposes: (1) inferring mortality risk of patients from their multi-modal EHRs; (2) identifying phenotype concepts of disease comorbidity in order to explain high risk of mortality.

In this study, we present MixEHR Survival Guided (MixEHR-SurG; Fig. 1). MixEHR-SurG is an extension of MixEHR [20] and MixEHR-G [23] and designed to integrate survival information and high-dimensional EHRs data via a supervised topic model framework [24]. Our contributions are three-folds: (1) integrating EHR topic inference with Cox proportional hazards likelihood; (2) integrating patient-specific topic hyperparameters using the PheCode concepts such that each topic can be identified with exactly one PheCode-associated phenotype; (3) multi-modal survival topic inference. As a result, MixEHR-SurG can perform guided phenotype topic inference and survival risk analysis simultaneously. We perform comprehensive evaluations of MixEHR-SurG, benchmarking on both its predictive accuracy for patient survival times and its ability to generate meaningful survival-related phenotype topics. In our simulation study, MixEHR-SurG not only accurately predicts survival times but also identifies true survival topics. When applied to the real-world Quebec Congenital Heart Disease (CHD) dataset and the MIMIC-III ICU dataset, MixEHR-SurG excels in predicting survival times and produces meaningful mortality-related phenotype topics. In the CHD dataset, MixEHR-SurG reveals cardiac-related phenotypes as significant mortality risk factors after the first onset of heart failure. In the MIMIC-III dataset, MixEHR-SurG identifies critical neurological conditions as one of the key mortality indicators.

2. Methods

2.1. MixEHR

This section briefly reviews MixEHR [20]. EHR includes a collection of medical documents of M types, indexed by $m = 1, \ldots, M$, such as ICD codes, drug codes, and clinic notes, etc. These documents provide a comprehensive overview of patients' clinical histories and examination results, which reflects personal health conditions. For document type m, a list of EHR features, indexed by $v = 1, \ldots, V^{(m)}$, encompasses all potential unique EHR features that are collected for that specific document type present in the dataset. For patient $j \in \{1, \ldots, P\}$, the EHR document of type m contains $N_j^{(m)}$ tokens, and each token is represented as $x_{ji}^{(m)}$, for $i = 1, \ldots, N_j^{(m)}$. In the context of topic modeling, the feature distribution of document type m under topic k is denoted as $\boldsymbol{\phi}_k^{(m)} = [\phi_{kv}^{(m)}]_{v=1}^{V^{(m)}} \in \mathbb{R}^{V^{(m)}}$. These weights are derived from a Dirichlet distribution, with an unknown hyperparameter $\boldsymbol{\beta}^{(m)} \in \mathbb{R}^{V^{(m)}}$. Additionally, the model assumes a specific topic assignment, represented as $\boldsymbol{\theta}_j \in \mathbb{R}^K$, for each patient j, which is also derived from a Dirichlet distribution, with a K-dimensional hyperparameter $\boldsymbol{\alpha}$. For every EHR token $x_{ji}^{(m)}$, with a latent topic assignment represented as $z_{ji}^{(m)}$, MixEHR has the following generative process (Fig. 2a):

1. Generate the feature distribution $\boldsymbol{\phi}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(m)}\right)$ for topic $k = 1, \ldots, K$ and type $m = 1, \ldots, M$.
2. For each patient $j = 1, \ldots, P$, sample a K-dimensional topic mixture: $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha})$.
   (a) For each of the EHR token $x_{ji}^{(m)}$ for $i = 1, \ldots, N_j^{(m)}$, $j = 1, \ldots, P$ and $m = 1, \ldots, M$:
      i. Sample a latent topic for token i: $z_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\theta}_j\right)$.
      ii. Sample a word for token i: $x_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\phi}_{z_{ji}^{(m)}}^{(m)}\right)$.

The posterior distributions of $\boldsymbol{\theta}_j$ and $\boldsymbol{\phi}_k^{(m)}$ are approximated by the collapsed mean-field variational inference method [25]. Although MixEHR is useful for multi-modal topic inference, it does not directly predict a target phenotype of interest. [21] proposed the MixEHR-S model in their study, which enables supervised topic inference for predicting a binary phenotype label. Nevertheless, time-to-event outcomes for survival analysis are crucial in medical research and clinical applications. Consequently, we seek to expand the MixEHR family to survival-supervised disease topic learning.

2.2. MixEHR-G

The data generative process (Fig. 2b) assumes that for each patient j, a set of noisy phenotype label are observed based on a phenotype reference such as the Phenotype Code or PheCode [26]. Let $\mathbf{u}_j \in \{0, 1\}^K$ be a binary vector of observed phenotype labels in patient j. The topic mixture is sampled from a Dirichlet distribution $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\pi}_j)$, where $\pi_{jk} \equiv p(y_{jk} = 1 \mid u_{jk})$, where y_{jk} is a binary latent variable indicating presence or absence of phenotype k for patient j. We infer the posterior distribution of y_{jk} using two-component univariate mixture models as described in **Supplementary Section** S3. In a nutshell, topic k with the observed phenotype label support in patient j will have relatively higher mixture proportion of θ_{jk} than those topics without the phenotype label support. The rest of the data generative process is identical to MixEHR.

2.3. MixEHR-SurG

Our objective is to identify phenotype topics that are informative of patient survival time. To this end, we extend MixEHR-G to integrate survival information. Let Y be the survival time for a patient, i.e., the time until a specific event occurs. In many applications, such as clinical studies, the survival time of a patient may not be known exactly. For example, a patient may not experience the event before the study ends or dropout during the study period (i.e., censored). Let C be the censoring time. The actual observed time T is either the survival time or the censoring time, whichever comes first, i.e., $T = \min(Y, C)$. Let $\delta = \mathbb{I}(Y \leq C) \in \{0, 1\}$ be the censoring status, where $\delta = 1$ indicates that Y is observed and 0 otherwise.

The survival function $S(t) = P(T > t)$ outputs the probability of survival beyond time t:

$$S(t) = \exp\left[-H(t)\right]$$

where $H(t)$ is the cumulative hazard function, defined as $H(t) = \int_0^t h(u)\mathrm{d}u$. This function accumulates the hazard function $h(u)$, over

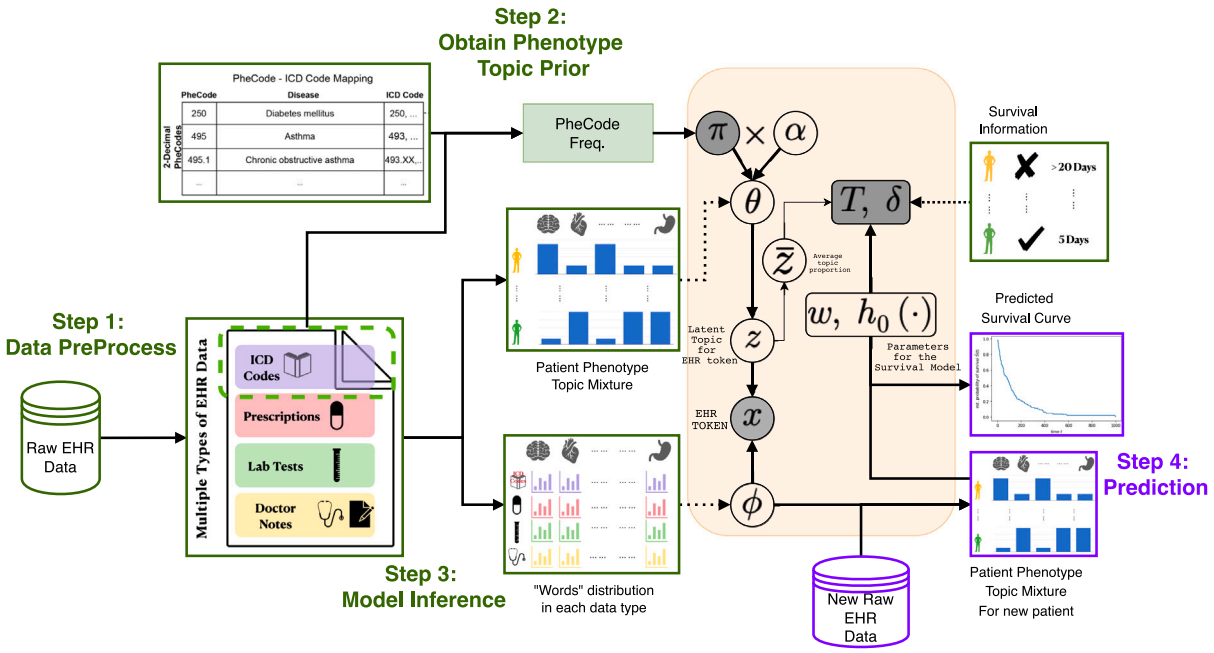


Fig. 1. MixEHR-SurG overview. MixEHR-SurG consists of four main steps. The training process is highlighted in green, and the prediction process is depicted in purple. In Step 1, we prepossess and aggregate raw EHR data for each patient j. Step 2 involves determining a K-dimensional phenotype topic prior, $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$, for each patient. Step 3 infers phenotype topic distribution $\boldsymbol{\phi}_k^{(m)} \in \mathbb{R}^{V^{(m)}}$ for EHR type m in topic k (i.e., the model parameters of MixEHR-SurG). This requires inferring the latent topic assignment $z_{ji} \in \{1, \ldots, K\}$ for each EHR token i in patient j. In Step 4, the trained model is applied to predict personalized survival function for new patient. The details of the probabilistic graphical model is depicted in Fig. 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the interval $[0, t]$. The hazard function $h(u)$ typically represents the instantaneous risk of the event (such as failure or death) occurring at time u.

Here we use a semi-supervised Cox PH model [12] for the hazard function. We further assume that the latent topic assignments can influence the survival response T_j (i.e., observed survival time for patient j). Specifically, we first compute the topic proportion $\bar{\mathbf{z}}_j$ for each patient $j = 1, \ldots, P$:

$$\bar{\mathbf{z}}_j = [\bar{z}_{jk}]_{k=1}^K = \left[\frac{\sum_{m=1}^M \sum_{i=1}^{N_j^{(m)}} \mathbb{I}(z_{ji}^{(m)} = k)}{\sum_{m=1}^M N_j^{(m)}} \right]_{k=1}^K$$

where $\bar{\mathbf{z}}_j$ can be viewed as the estimate of $\boldsymbol{\theta}_j$.

Next, the survival time T_j corresponds to the Cox proportional hazards (PH) model with a system-wide K-dimensional Cox PH regression coefficients $\mathbf{w}$ (fixed but unknown). The baseline hazard function is defined as: $h_0(\cdot)$, for each patient $j = 1, \ldots, P$:

$$h\left(T_j | \bar{\mathbf{z}}_j\right) = h_0\left(T_j\right) \exp\left\{\mathbf{w}^\top \bar{\mathbf{z}}_j\right\}$$

The preceding generative process of MixEHR-SurG is the same as MixEHR-G (Fig. 2d). Lastly, as one of the simplified model, we also implemented MixEHR-Surv (Fig. 2c), which is a survival-supervised MixEHR without using the PheCode guide.

2.4. MixEHR-SurG model inference

As depicted in Fig. 1, MixEHR-SurG combines MixEHR-G and survival topic model into a single model. The joint-likelihood function is:

$$p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\theta}, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right)$$
$$= p(\boldsymbol{\pi} \mid \mathbf{U}) p(\mathcal{X}, \mathcal{Z}, \boldsymbol{\theta}, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}) p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right)$$

The first term is the prior term $\boldsymbol{\pi}$ for the phenotype topic, which we separately infer using 2-component mixture univariate model on the Phecode counts matrix $\mathbf{U}$ for each PheCode-guided topic as detailed in the **Supplementary Section** S3. The second term is the unsupervised part of the likelihood and the same as the MixEHR-G [20]. The third term is the survival supervised component of the model. We use the Cox PH model with elastic net penalization (i.e., L1 + L2 norm) [27] for the survival coefficients.

While we fit the first term separately and fix $\boldsymbol{\pi}$ to the expected value, we jointly fit the second and the third term of the joint likelihood. Specifically, the full likelihood function of the penalized Cox PH model is obtained by incorporating Breslow's estimate of the baseline hazard function and the penalty term with the hyperparameter λ_1 for L1 norm and λ_2 for L2 norm.

$$p(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w})$$
$$= \prod_{j=1}^P \left\{ \left[h_0\left(T_j\right) \exp\left(\mathbf{w}^\top \bar{\mathbf{z}}_j\right)\right]^{\delta_j} \exp\left[-H_0\left(T_j\right) \exp\left(\mathbf{w}^\top \bar{\mathbf{z}}_j\right)\right] \right\}$$
$$\times \exp\left\{-\lambda_2 \|\mathbf{w}\|_2^2 - \lambda_1 \|\mathbf{w}\|_1\right\}$$

Note that here $\mathbf{w}$ is not a variable and there is no prior distribution. The second term was added only for regularization purpose using Elastic Net (Eq. (9); S4).

We will first integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ to achieve more accurate and efficient inference, due to the conjugacy of Dirichlet variables $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ to the multinomial likelihood variables $\mathcal{X}$ and $\mathcal{Z}$ [25]. Then, the ELBO for the current marginal distribution for the observed data is:

$$\mathcal{L}_{ELBO} = \mathbb{E}_q\left[\log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot)\right), \mathbf{w}\right] - \mathbb{E}_q\left[\log q(\mathcal{Z})\right] \quad (1)$$

where we assume a mean-field variational distribution for the topic assignments:

$$q(\mathcal{Z}) = \prod_{m=1}^M \prod_{j=1}^P \prod_{i=1}^{N_j^{(m)}} q(z_{ji}^{(m)}) = \prod_{m=1}^M \prod_{j=1}^P \prod_{i=1}^{N_j^{(m)}} \prod_{k=1}^K \left(\gamma_{jik}^{(m)}\right)^{[z_{ji}^{(m)}=k]}$$

Maximizing the Evidence Lower Bound (ELBO) with respect to $\gamma_{jik}^{(m)}$ is equivalent to computing the conditional expectation of the variable $z_{ji}^{(m)} = k$ given the estimates for other tokens. There exists an efficient

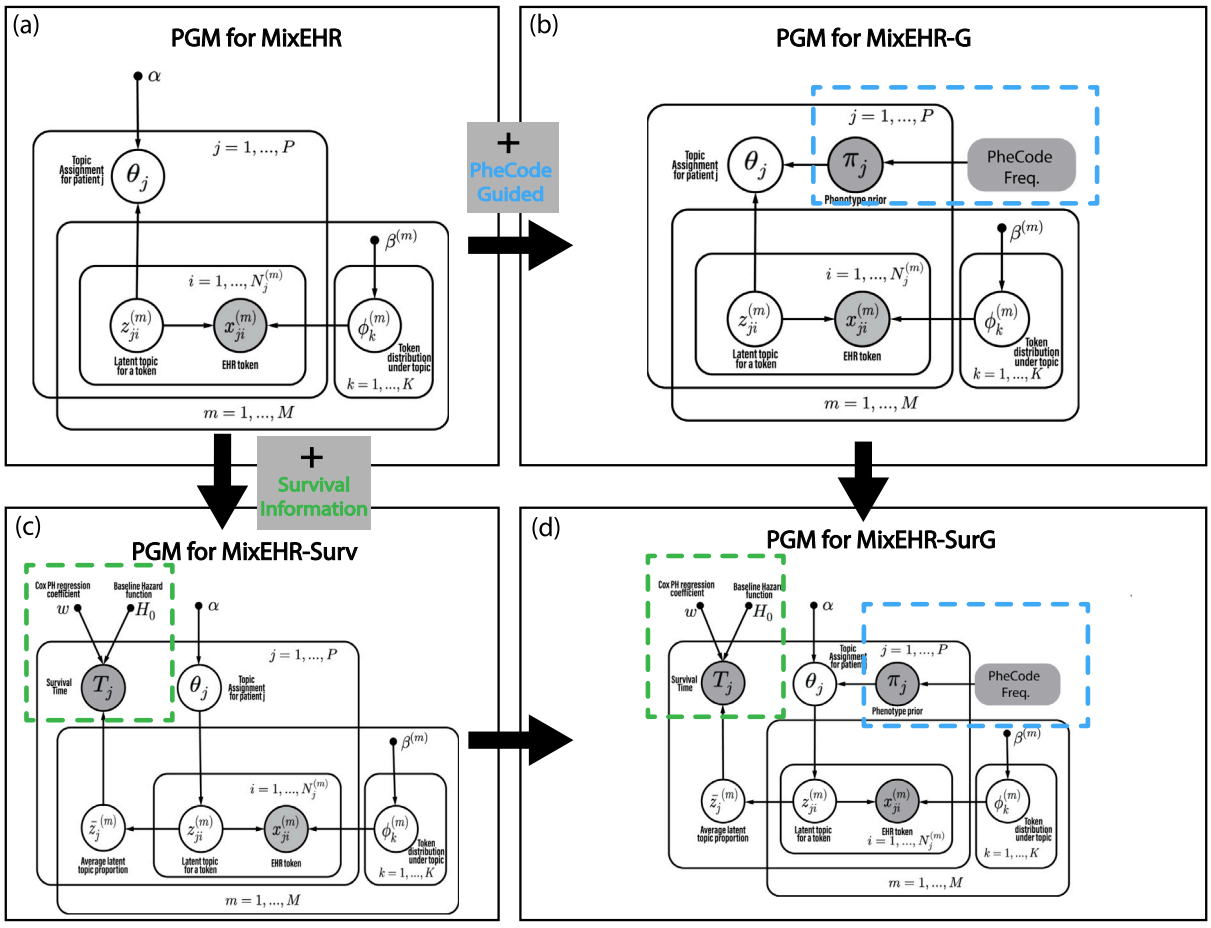


Fig. 2. Probabilistic graphical model (PGM) illustration of four models variants. (a) PGM for MixEHR. We first generate topic distributions $\boldsymbol{\phi}_k^{(m)}$ for each topic k and document type m, then we generate of a K-dimensional topic proportion $\boldsymbol{\theta}_j$ for every patient j. Finally, we generate latent topics $z_{ji}^{(m)}$ and corresponding words $x_{ji}^{(m)}$ for each EHR token. (b) PGM for MixEHR-G. We infer patient specif PheCode-Guided topic prior $\boldsymbol{\pi}_j$ for each patient j and used it as Dirichlet hyperparameters for the patient topic mixture $\boldsymbol{\theta}_j$ enclosed by a blue dashed rectangular. (c) PGM for MixEHR-Surv. For each patient j, we obtained the survival time T_j and employed the Cox proportional hazards (PH) model with coefficient $\mathbf{w}$ and baseline hazard function $h_0(\cdot)$ to guide the learning of topics, as enclosed by a green dashed rectangular. (d) PGM for the proposed MixEHR-SurG. We combine both PheCode-Guided prior and survival information into one single model. The resulting model can use the guided phenotype topics to model the Cox PH of survival likelihood. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

closed-form update expression (**Supplementary Section** S4):

$$\gamma_{jik}^{(m)} \propto \exp\left\{\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(T_j, \delta_j \mid z_{(j,-i)}^{(m)}, z_{ji}^{(m)}=k, h_0(T_j), \mathbf{w}\right)\right]\right\}$$
$$\times\left(\alpha_k \pi_{jk} + \left[n_{j\bullet k}^{(\bullet)}\right]_{(j,-i)}\right) \frac{\beta_{x_{ji}^{(m)}}^{(m)} + \left[n_{\bullet x_{ji}^{(m)} k}^{(m)}\right]_{(-j,-i)}}{\sum_v \beta_v^{(m)} + \left[n_{\bullet vk}^{(m)}\right]_{(-j,-i)}}, \tag{2}$$

where the subscript $(j,-i)$ indicates excluding token i of patient j when calculating its own expectation and the coordinate sufficient statistics are:

$$n_{\bullet vk}^{(m)} = \sum_{j=1}^{P}\sum_{i=1}^{N_j^{(m)}} \mathbb{I}\left[x_{ji}^{(m)}=v, z_{ji}^{(m)}=k\right],$$
$$n_{j\bullet k}^{(\bullet)} = \sum_{m=1}^{M}\sum_{i=1}^{N_j^{(m)}} \mathbb{I}\left[z_{ji}^{(m)}=k\right].$$

This equation is derived under the principle that the Kullback–Leibler (KL) divergence reaches its minimum when the approximation of the variational parameter matches the expectation under all other known latent variables. Additionally, the hyperparameters $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}^{(m)}$'s updates are refined through empirical Bayes by optimizing ELBO given the variational estimates of the topic assignments $\gamma_{jik}^{(m)}$'s. Detailed derivation are described in **Supplementary Section** S4.

Given the variational expectation of $\mathcal{Z}$, the Cox regression coefficients $\mathbf{w}$ are fit via penalized log likelihood of $\log p(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w})$ via Cox elastic net regression, which was originally implemented in the `glmnet` R package [27]. In our MixEHR-SurG implementation, we use the Python wrapper of the `glmnet` (https://pypi.org/project/glmnet/).

Upon training MixEHR-SurG, we obtain $\hat{\boldsymbol{\Phi}}^{(m)}$, where $m = 1, \ldots, M$ indexes modalities, the hyperparameters $\hat{\boldsymbol{\alpha}}$ and $\hat{B}$, and the point estimates of the Cox regression coefficients $\hat{\mathbf{w}}$ along with the two-component mixture models trained for each PheCode-guided topic (**Supplementary Section** S3).

2.5. Inferring personalized survival probabilities

For a new patient j' with EHR documents denoted as $x_{1:N^{(m)},j'}^{(m)}$, $m = 1, \ldots, M$. We first compute the topic assignments by the following steps:

1. For every token $i = 1, \ldots, N_{j'}^{(m)}$ over every modality $m = 1, \ldots, M$:

$$\gamma_{j'ik}^{(m)} \propto \left(\hat{\alpha}_k \pi_{j'k} + \left[n_{j'\bullet k}^{(\bullet)}\right]_{(j',-i)}\right) \hat{\boldsymbol{\Phi}}_{x_{j'ik}^{(m)}}^{(m)}$$

where $\pi_{j'k}$ is inferred using the trained 2-component mixture model on the training data, $\hat{\alpha}_k$ is the estimated global hyperparameter, and $\hat{\boldsymbol{\Phi}}_{1:K}^{(m)}$ is the estimated topic distributions from the training data.

2. Update sufficient statistics:

$$n^{(\bullet)}_{j'\bullet k} = \sum_{m=1}^{M}\sum_{i=1}^{N^{(m)}_{j'}} \gamma^{(m)}_{j'ik}$$

3. Evaluate the log marginal likelihood:

$$\mathcal{L}_{j'} = \sum_{m=1}^{M}\sum_{i=1}^{N^{(m)}_{j'}}\sum_{k=1}^{K} \log \theta_{j'k}\phi^{(m)}_{kx^{(m)}_{j'i}}$$

where $\theta_{j'k} = n^{(\bullet)}_{j'\bullet k} / \sum_{k'=1}^{K} n^{(\bullet)}_{j'\bullet k'}$.

4. Repeat 1 and 2 until 3 converges.

Finally, we compute the mean of the variational estimates of the topic assignments:

$$\bar{\gamma}_{j'k} = \sum_{m=1}^{M} \frac{1}{N^{(m)}_{j'}} \sum_{i=1}^{N^{(m)}_{j'}} \gamma^{(m)}_{j'ik}$$

Using $\bar{\boldsymbol{\gamma}}_{j'} = \left[\bar{\gamma}_{j'k}\right]_{k=1}^{K}$ and survival coefficients $\hat{\mathbf{w}}$, we can calculate the estimated hazard ratio for the new patient:

$$\widehat{\mathrm{HR}}_{j'} = \exp\left(\hat{\mathbf{w}}^{\top}\bar{\boldsymbol{\gamma}}_{j'}\right)$$

Moreover, we can compute the survival function for patient j up to time t:

$$P(T_{j'} > t) = \exp[-H_0(t)\widehat{\mathrm{HR}}_{j'}]$$

where $H_0(t) = \int_0^t h_0(u)\mathrm{d}u$ denoted as the baseline cumulative hazard function. With this survival function, we can generate personalized survival curve for every patient and estimate their survival probability at specific time points [12].

3. Data processing and experiments

3.1. Simulation design 1

We designed a simulation study to evaluate MixEHR-SurG in terms of the accuracy of (1) identifying topics associated with patient survival and (2) predicting patient survival times. We set the vocabulary to be 1000 words ($V = 1000$) and simulated 500 distinct topics ($K = 500$). We sampled 8000 patients and each patient consists of 100 tokens using the data generative process of MixEHR-SurG. For each patient $j \in \{1, \ldots, 8000\}$, the topic proportions $\boldsymbol{\theta}_j$ was sampled from a Dirichlet distribution with the hyperparameter $\boldsymbol{\alpha}$ sampled from a Gamma distribution with a shape parameter of 10 and a scale parameter of 1. The topic distributions $\boldsymbol{\phi}_k$ was sampled from a V-dimensional Dirichlet distribution with the hyperparameter $\mathcal{B}$ sampled from a Gamma distribution with shape and scale parameters of 2 and 500, respectively. The topic assignment $z_{ji} \in \{1, \ldots, K\}$ was sampled from a Categorical distribution at the rate set to $\boldsymbol{\theta}_j$, and word assignments $x_{ji} \in \{1, \ldots, V\}$ were sampled from another Categorical distribution at the rate $\boldsymbol{\phi}_{z_{ji}}$.

To evaluate whether our model can identify mortality-related topics, we set the survival coefficients $\mathbf{w}$ to be a sparse vector. Specifically, we set 50 out of the 450 coefficients to 6 and the rest to 0. We then computed the topic proportion $\bar{\mathbf{z}}_j$ for each patient j.

Survival time T_j were simulated via the Cox model:

$$T_j = H_0^{-1}\left(-\log(U)\exp(-\mathbf{w}^{\top}\bar{\mathbf{z}}_j)\right)$$

where U is a uniformly distributed random variable on the interval $[0, 1]$, and the transformation is done through the inverse of the baseline hazard function H_0^{-1} [28]. We chose H_0^{-1} based on the distribution of the survival times. For simplicity, we adopt the Exponential distribution, a common choice in survival analysis. In this scenario, the cumulative baseline hazard function is expressed as $H_0(t) = \lambda t$, with λ being a hyperparameter set to 1 for simplicity, leading to the inverse function $H_0^{-1}(t) = \lambda^{-1}t$.

3.2. Simulation design 2

To create a simulated dataset that closely replicates real-world data, we focused on the CHD dataset, utilizing diagnosis codes documented prior to the first ICU discharge of CHD patients for predicting mortality time. Our simulation was tailored to mirror the CHD dataset's specific attributes, including a total of 8211 patients (P = 8211) and maintaining consistency with the actual count of phenotype topics found in the CHD dataset (K = 490).

For the simulation of topic distributions $\boldsymbol{\phi}_k$, we drew from a V-dimensional Dirichlet distribution, with each hyperparameter $\boldsymbol{\beta}_k$ determined by the relationship between ICD-9 codes and PheCodes. Specifically, for a given ICD-9 code v and PheCode k, we set $\beta_{kv} = 1$ if there exists a mapping between v and k; otherwise, $\beta_{kv} = 0$. To satisfy the Dirichlet distribution requirement that $\boldsymbol{\beta}_k > 0$, we transformed the mapping to $\beta_{kv} = \beta_{kv} \times 3 + 0.6$. This adjustment ensures that the simulated topic distributions closely approximate those learned from the CHD dataset by reflecting the distribution of words within a topic and highlighting dataset-specific signals and differences.

For each patient, the topic proportions $\boldsymbol{\theta}_j$ were directly sampled based on the observed patient's PheCode frequency from the CHD dataset. The number of records N_j for each patient was also matched to the CHD dataset to preserve information density and sparsity. We then simulated the ICD codes for the patient as follows: for each code i, the topic assignment $z_{ji} \in \{1, \ldots, K\}$ was sampled from a Categorical distribution parameterized by $\boldsymbol{\theta}_j$. The ICD code for $x_{ji} \in \{1, \ldots, V\}$ was then sampled from the Categorical distribution parameterized by $\boldsymbol{\phi}_{z_{ji}}$. We then randomly designated 10% of the Phenocode topics to have survival coefficients w_k set to 6, with the remainder set to 0. Survival times T_j for each patient was sampled the same way as in Simulation Design 1.

3.3. Preprocessing of the Quebec CHD data

We leveraged the inpatient and outpatient ICD codes from a patient's first documented heart failure episode to predict their subsequent time to death, measured in days. The cohort for this study was the Congenital Heart Disease (CHD) claim database. This dataset combines the Physician Services and Claims spanning from 1983 to 2010, the Hospital Discharge Summaries from 1987 to 2010, and the Vital Status records from 1983 to 2010. The dataset contains 8211 CHD subjects, who experienced at least one heart failure and had a recorded death date. The data were constructed by collating all the ICD codes recorded during the hospitalization of the first heart failure episode. In total, there are 75,187 records and 1767 unique ICD9 codes. We mapped the ICD9 codes to 498 unique PheCodes in one-decimal code format for the guided topic inference (**Supplementary Section** S3). We selected PheCodes that appeared in over 25% of the patient population within the dataset. The survival time of each CHD patient is the time difference between the death date and the discharge date of the first heart failure hospitalization.

3.4. Preprocessing of MIMIC-III data

To demonstrate the generalization of MixEHR-SurG and the ability for inferring multi-modal EHR topics, we made use of the Medical Information Mart for Intensive Care III (MIMIC-III) dataset [29]. MIMIC-III is a comprehensive dataset originating from the Beth Israel Deaconess Medical Center in Boston, MA, encompassing 53,423 distinct hospital admissions across 38,597 adult patients and 7870 neonates from 2001 to 2012. The dataset was downloaded from the PhysioNet database (mimic.physionet.org) under its user agreement. We carried out the same preprocessing as described in [23]. We then selected patients who had multiple inpatient records and a documented time of death. We utilized all available EHR information up to the discharge time of the first inpatient stay to predict the time lapse the patient survived since

the ICU discharge. To refine our dataset for more accurate predictions, we specifically filtered out patients whose discharge date from their first inpatient admission coincided with their date of death. The final dataset consisted of 1458 patients, of which 1168 were used for training the model and 290 for testing. Among these patients, there are 55,529 unique features among five EHR types including clinical notes (47,383), ICD-9 codes (3293), lab tests (588), prescriptions (3444), and DRG codes (821).

3.5. Evaluation

To evaluate the MixEHR-SurG's ability of predicting patient survival time, we utilized dynamic area under the ROC (AUC) curve [30–32], a modification of the traditional ROC curve particularly suited for survival data analysis. Dynamic AUC extends the concept of AUC to survival data by defining time-dependent sensitivity (true positive rate) and specificity (true negative rate). In this context, cumulative cases include individuals who experienced an event by or before a specific time $\{j \mid T_j \le t,\ j = 1, \ldots, P\}$, while cumulative controls are those for whom the event occurs after this time $\{j \mid T_j > t,\ j = 1, \ldots, P\}$. The corresponding cumulative/dynamic AUC evaluates the model's ability to distinguish between subjects who experienced an event by a given time $(T_j \le t)$ and those who experienced it later $(T_j > t)$.

Given an estimated risk ratio $\widehat{\mathrm{HR}}_j$ for the jth individual, the cumulative/dynamic AUC at time t is defined as:

$$\widehat{\mathrm{AUC}}(t) = \frac{\sum_{i=1}^{P}\sum_{j=1}^{P} I\left(T_j > t\right) I\left(T_i \le t\right) I\left(\widehat{\mathrm{HR}}_j \le \widehat{\mathrm{HR}}_i\right)}{\left(\sum_{j=1}^{P} I\left(T_i > t\right)\right)\left(\sum_{j=1}^{P} I\left(T_i \le t\right)\right)}$$

Building on this, we define a sequence of time points and calculate the cumulative/dynamic AUC at each point in this series, thereby constructing the Dynamic AUC curve.

4. Results

4.1. Simulation

MixEHR-SurG demonstrates high sensitivity and specificity in detecting the 50 true survival-associated phenotypes out of the 450 phenotypes (Fig. 3a,b; **Supplementary Section** S5). Notably, the true effect size is 6, and the model estimates is between 2 and 3, which is due to the L1/2-regularization (i.e., elastic net) on $\mathbf{w}$ via the regularized Cox regression (**Supplementary Section** S4). We then evaluated MixEHR-SurG in terms of predicting survival times in comparison to pipeline approach that ran MixEHR-G followed by Cox regression. This comparison was made using dynamic AUC curves (Fig. 3c), which provide a nuanced measure of sensitivity and specificity over time for survival data. MixEHR-SurG slightly improved over MixEHR-G with mean AUC of 0.89 versus mean AUC of 0.88, respectively. To assess whether the improvement is statistically significant, we repeated the simulations 10 times and computed the Wilcoxon signed-rank test, which yielded a p-value of 0.0488 (**Supplementary Fig. S1**). We conducted another simulation closely based on the real-world MIMIC-III and CHD datasets (**Methods** 3.2). As expected, we observed lower AUCs but the relative performance between MixEHR-SurG and Coxnet-MixEHR-G are similar (**Supplementary Fig. S2**).

4.2. Application to the CHD dataset

We evaluated the survival models on CHD patient survival time predictions after their initial HF hospitalization. MixEHR-SurG conferred the highest mean AUC (0.645) compared to MixEHR-G with the Coxnet pipeline (0.623), MixEHR-Surv (0.576), and MixEHR with the Coxnet pipeline (0.556) and DeepSurv (0.64) (Fig. 4). To further ascertain the benefit of joint topic inference and survival regression, we sampled from the test patients with replacement 10,000 times and calculated the difference of the mean AUCs for each bootstrap between MixEHR-SurG and MixEHR-G+Coxnet: ΔAUC = AUC(MixEHR-SurG) - AUC(MixEHR-G+Coxnet). We used the 10,000 ΔAUCs to construct an empirical distribution of the performance difference between the two methods. The 75% confidence intervals (CIs) of the empirical distribution is [0.00364 0.0370], corresponding to the 12.5% quantile and 87.5% quantile of the empirical ΔAUC distribution, respectively. Furthermore, 9260 out of the 10,000 bootstrap ΔAUC are positive, which is statistically significant at the p-value < 2.5e−324 based on one-sided Binomial test with the null hypothesis being at the equal chance of producing positive and negative ΔAUC over the 10,000 bootstrap samples.

Based on the Cox regression learned simultaneously by MixEHR-SurG, we identified the most predictive phenotypes of the post-HF survival time (Fig. 5a). Among these phenotypes, nonrheumatic pulmonary valve disorder (NPVD) (395.4) is the most prominent phenotype. Indeed, the CHD subjects who exhibit high topic proportion for NPVD tend to have much shorter survival time compared to the rest of the CHD subjects (Fig. 5b; **S6**). We then obtained the p-values and confidence intervals of the six phenotypes selected for their large absolute value of w_k, through a Cox proportional hazards model [12]. Based on the results (Fig. 5c), it is evident that "Nonrheumatic pulmonary valve disorders", "Postoperative shock", and "Cardiogenic shock" have emerged as significant factors contributing to the occurrence of mortality. These phenotypes are characterized by substantial positive coefficients and statistically significant p-values, underscoring their strong association with increased risk. Interestingly, "Complication due to other implant and internal device" (859.0) is associated with longer survival time, which perhaps imply the deficiency of healthcare among those high risk patient group. We then examined the underlying top ICD9 codes under the predictive phenotype topics (Fig. 5d). In particular, topic for NPVD includes several cardiac-related ICD codes with pulmonary valve disorders being the most prominent one as expected. Phenotype topics "Postoperative shock" (958.1) and"Cardiogenic shock" (797.1) were also associated with the relevant ICD-9 codes, implying high topic coherence. The 3 negative topics are not heart-specific but nonetheless semantically coherent. We further validated the topic coherence based on the mutual information (MI) between the top ICD codes for the top 6 survival phenotype topics (**Supplementary Fig.** S4). Indeed, we observe a clear 5×5 block pattern corresponding to the top 5 ICD codes for the corresponding phenotype topic along the diagonal of the MI matrix. Furthermore, the top ICD codes that are not part of the PheCode definition exhibit high MI with the PheCode-defining ICD code, implying that they are not only related to the phenotype but also co-occur with the PheCode-defining ICD code in the actual patient records. This also suggests that MixEHR-SurG does not completely rely on the PheCode guide not also driven by the CHD data in characterizing the phenotype topic distributions.

4.3. Application to MIMIC-III dataset

We then benchmarked each method on the mortality prediction using the MIMIC-III dataset (**Supplementary Fig.** S5). Among the four model variants, MixEHR-SurG achieved the highest mean AUC (0.54), closely followed by the Coxnet-MixEHR-G pipeline (0.53). MixEHR-Surv and the Coxnet-MixEHR pipeline conferred mean AUCs of 0.48 and 0.39, respectively. Moreover, MixEHR-SurG significantly outperformed the runner up baseline MixEHR-G+Coxnet with the 75% CI estimated from 10,000 bootstrap equal to [0.000913, 0.0360] (**Supplementary Fig.** S3b), and 8967 out of the 10,000 bootstrap ΔAUC = AUC(MixEHR-SurG) - AUC(MixEHR-G+Coxnet) being positive (p-value <2.5e−324 based on one-sided Binomial test rejecting the null hypothesis at the equal chance of getting positive and negative ΔAUC).

Nonetheless, the absolute AUC level is lower than the CHD data, which may be due to the smaller sample size and more diverse causes of death. In addition, the relationship between patient data and survival

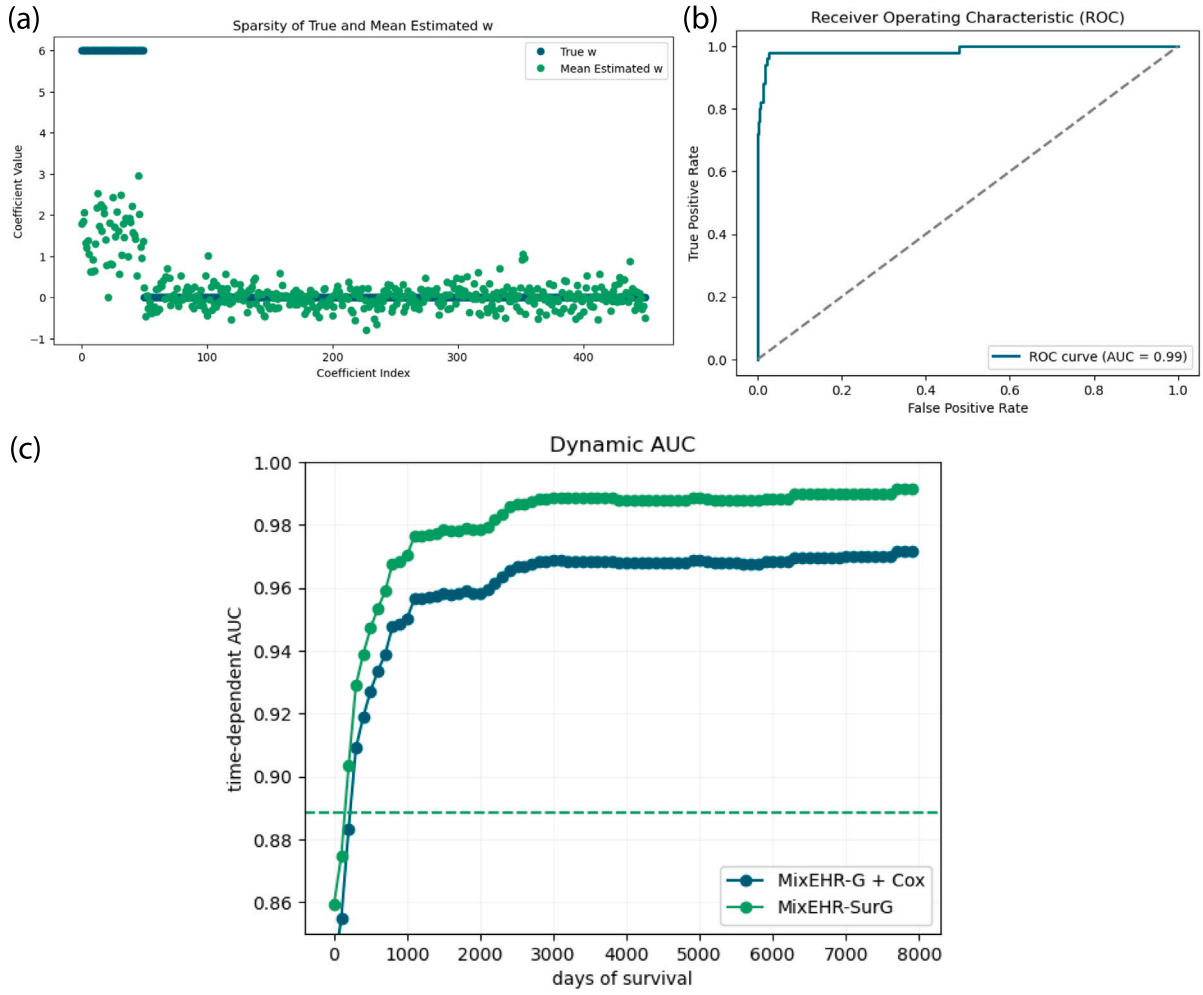


Fig. 3. Simulation Results for MixEHR-SurG. (a) Scatter plot comparing the estimated coefficients **w** (in green) with their true values (in blue). (b) ROC curve for predicting zero coefficients. (c) Dynamic AUC curves to evaluate survival time prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outcomes in MIMIC-III is influenced by the dataset's heterogeneity and the emergency nature of many admissions, where acute conditions can overshadow chronic illness history in predicting mortality. The CHD dataset, by contrast, lends itself to more accurate predictions due to the focused nature of the cohort. Patients with CHD often have extensive medical histories and a narrower range of complications, providing a stronger and more direct signal for predicting mortality.

We then sought to identify phenotype topics that are indicative of the short-term mortality based on the Cox regression coefficients that were jointly fit with the EHR data by our MixEHR-SurG (Fig. 6a). The most prominent phenotype topic is "Cerebral laceration and contusion (816.0)", which also separates patients into high and low risk groups (Fig. 6b). We subsequently assessed the p-values and confidence intervals of six phenotypes with the largest absolute values of w_k, through a Cox proportional hazards model [12] (Fig. 6c). The results confirm the significance of "Cerebral laceration and contusion", "Cerebral edema and compression of brain" and "Dysthymic disorder". Specifically, "Cerebral laceration and contusion" and "Cerebral edema and compression of brain" show a positive correlation with an increased risk of mortality, while "Dysthymic disorder" indicates a negative correlation, suggesting a potential protective effect against mortality. Indeed, traumatic brain injuries often lead to servere morbidity and ultimately death. Conversely, MixEHR-SurG reveals conditions such as "Retinoschisis and retinal cysts" and "Dysthymic disorder" with large negative survival coefficients, suggesting a low immediate threat to life. Retinoschisis and retinal cysts typically do not directly impact survival unless complicated by additional factors, and "Dysthymic disorder" while affecting quality of life, generally does not shorten life expectancy in the absence of other comorbid conditions.

We further performed Kaplan–Meier (KM) survival analysis and computed the p-values using one-sided log-rank tests for the top ICD codes under the top 6 survival phenotype topics (**Supplementary Fig.** S6a). We observe that the top codes associated with the first three phenotypes, which have higher survival coefficients, display significant marginal effects of increased hazard risks. Conversely, the top codes linked to the last three phenotypes exhibit significant marginal effects of reduced hazard risks. These findings confirm that our model can effectively pinpoint terms with substantial impacts on survival. Some ICD-9 codes, such as '8080 Closed fracture of acetabulum' and '72889 Disorder of muscle, ligament, and fascia', are not significant by themselves but contribute in aggregate to the survival phenotype suhc as '816.0'. Under the topics of traumatic brain injuries, namely "816.0: Cerebral laceration and contusion" and "348.2: Cerebral edema and compression of brain", the top ICD codes are semantically coherent (Fig. 6d). Quantitatively, we computed the mutual information (MI) between the top ICD codes (**Supplementary Fig.** S6b). As expected, the top codes under the same phenotype topic exhibit high MI, implying a high topic coherence. This may not be surprising as some of the ICD codes were used to define the PheCode, which were then used to build the topic prior.

To gain further insights to these topics, we examined the top EHR codes from non-ICD modality topics (Fig. 6e-I). In clinical notes

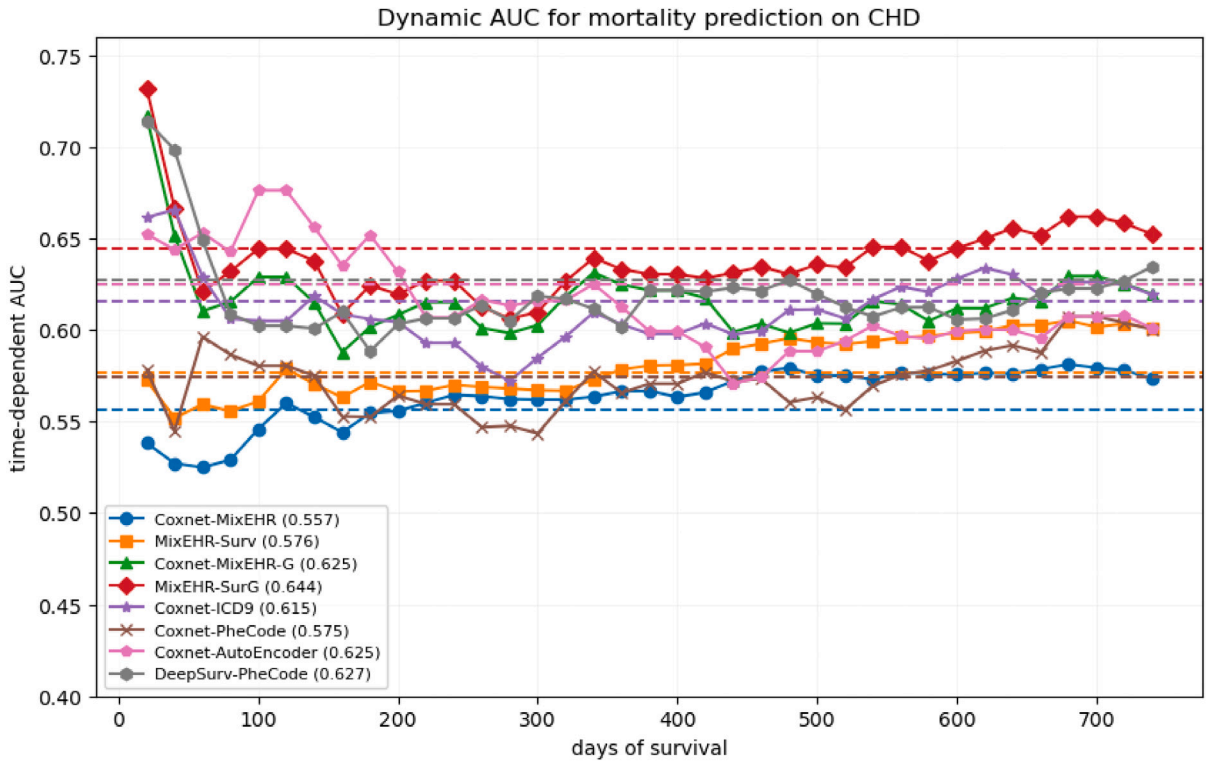


Fig. 4. Dynamic AUC curves for predicting time to death in CHD patients. We built a series of time points starting from 20 and incrementing by 20 up to 755. For each of these time points, we computed the cumulative AUC, which then formed the Dynamic AUC curve. The mean AUC over time for each method was indicated as dash lines and in the bracket after each method in the legend. The compared methods are: **Coxnet-MixEHR**: A pipeline approach by training MixEHR first and then training a Cox elastic net (Coxnet) using the topic mixture from MixEHR as the input features; **MixEHR-Surv**: MixEHR with the Cox supervision but without the phecode guided prior for the topic inference; **Coxnet-MixEHR-G**: A pipeline approach by training MixEHR-G first and then training a Cox elastic net (Coxnet) using the topic mixture from MixEHR-G as the input features; **MixEHR-SurG**: the proposed method in this paper; **Coxnet-ICD9**: Cox elastic net (Coxnet) using ICD9 code as input features; **Coxnet-PheCode**: Coxnet using PheCode as input features; **Coxnet-AutoEncoder**: Coxnet using the output of an autoencoder as input features; **DeepSurv-PheCode**: Deep survival model using PheCode as input features.

(Fig. 6e), we identified top words related to mannitol, a diuretic used to reduce intracranial pressure [33]. Mannitol is the treatment of cerebral edema (accumulation of excessive fluid in the brain). The fact that it is the top drug code for the top risk mortality phenotype 816.0 suggests the severity of the condition. Indeed, we also found "Osmolality, Measured" to be the top term under the laboratory modality and "Mannitol 20%" as the top term under the same topic from prescription modality (Fig. 6I). In addition, the DRG (Diagnosis-Related Group) topic modality exhibit strong connection with the ICD-modality topic despite the fact that DRG codes are not part of the PheCode definition. Most of these top codes also exhibit consistently significant marginal effect size based on the KM test and coherence in terms of mutual information (**Supplementary Fig.** S7, S8, S9, S10). Together, these results showcases the MixEHR-SurG's ability to harness non-ICD modality to enrich the phenotyping, which is consistent to what we observed in MixEHR-G [23].

Our results suggested that brain injuries are the strong mortality-indicators and the topic coherence across modalities provide the fine-grained markers for screening high-risk patients in the future.

5. Discussion

Effective utilization of EHR data holds the promise to automate phenotyping [3] and identifying prognosis markers [34]. MixEHR-SurG extends EHR topic modeling to survival topic model with the identifiable topics by utilizing the patient survival time and PheCode definitions, respectively. We demonstrated the utility of MixEHR-SurG via both simulation and real-world EHR data including the Quebec CHD and MIMIC-III datasets [29]. The results from these rigorous experiments highlights our contribution in MixEHR-SurG as an effective approach to identify clinically meaningful phenotypes that implicate mortality.

Despite this advance, there are several limitations in our method. First, EHR data often contain hierarchical structures of phenotypes that are yet to be unraveled. For instance, leveraging advanced hierarchical topic modeling [35,36] could shed light on sub-phenotypes and their interactions within broader disease categories. While we have harnessed cross-sectional data effectively, the longitudinal nature of EHRs, characterized by patient trajectories and time-stamped health events, presents an opportunity to explore temporal patterns and trends [37] in future studies.

Although MixEHR-SurG showcases predictive prowess, it does not have the same level of expressiveness as deep neural networks. The integration of deep learning with topic model [38,39] could potentially enhance predictive performance by capturing non-linear relationships and complex interaction effects within EHR data [22,40–42]. Furthermore, the challenge in distinguishing between high-mortality risk phenotypes and confounding factors remains and calls for causality-driven models [43–46]. Future study will be dedicated to not only predict outcomes but also discern the underlying causal mechanisms, offering a more granular understanding of patient risk profiles. Causal inference that discern direct from indirect influences of phenotypes on survival outcomes will bring a step closer to effective clinical interventions.

In summary, MixEHR-SurG is a novel topic model that leverages EHR data for both interpretive and predictive modeling of patient survival outcomes. By successfully mapping EHR data to relevant phenotypes and delineating those with high mortality risks, MixEHR-SurG serves as a prototype for future systems that could offer nuanced insights into patient care. The current study lays the groundwork for subsequent research that could incorporate hierarchical data structures and temporal dynamics within EHRs [35–37], potentially utilizing advanced machine learning techniques such as deep learning and causal

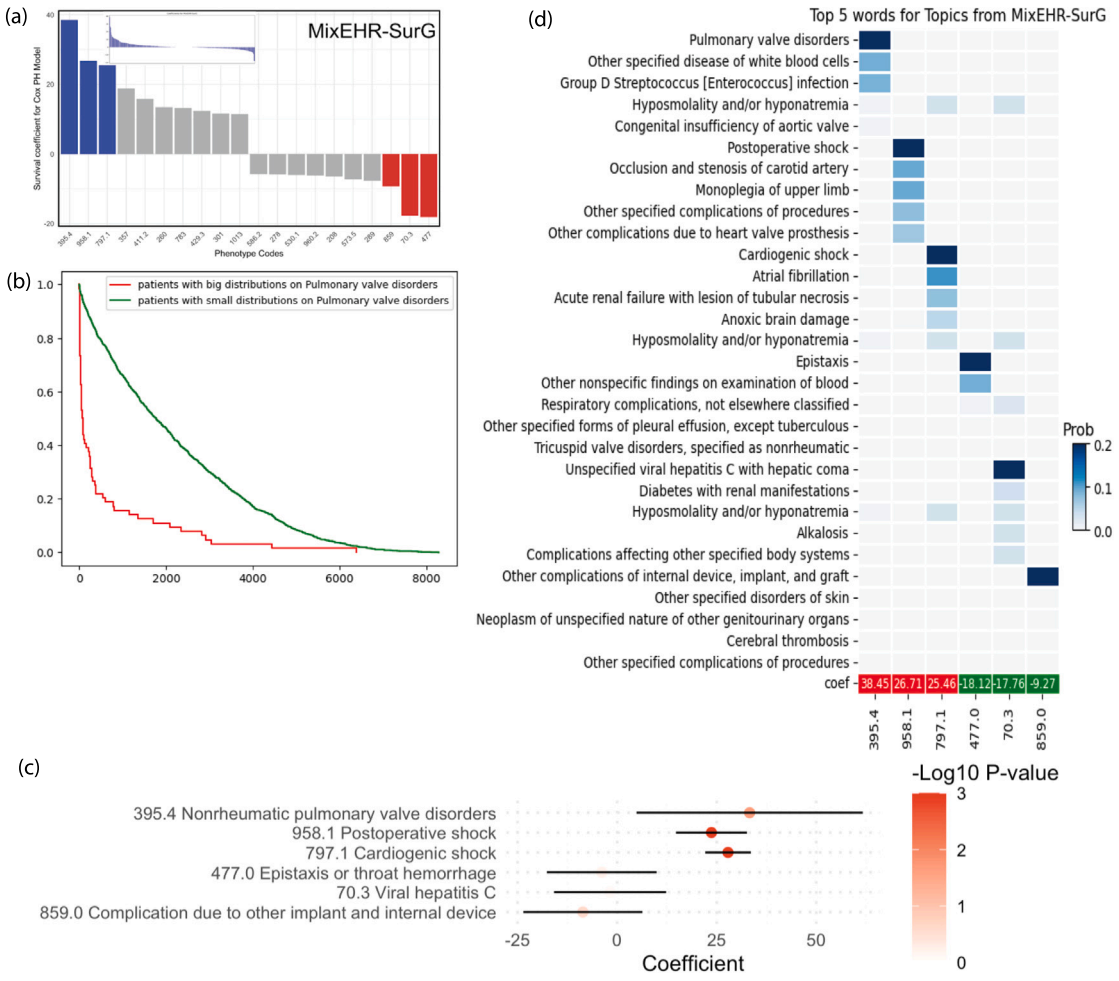


Fig. 5. Mortality-related phenotypes for CHD patients who experienced first heart failure hospitalization. (a) Bar plot of the survival regression coefficients **w**. The effect size of the 10 most positive and the 10 most negative phenotypes are displyaed as barplot. The positive value refers to phenotypes that are associated with high risk of mortality and the negative value refers to phenotypes associated with low mortality risk. The inset at the up-left corner contains the bar plot for all the estimated $w_k, k = 1, \ldots, K$ ranked from the largest value to the smallest value. The top 3 and bottom phenotypes were colored in blue and red, respectively. (b) The survival curves of patient with high and low risk of nonrheumatic pulmonary valve disorder (NPVD) (395.4). Patients were divided into two groups based on their topic proportions. The red curve represents patients with a higher topic proportion (top 30%) in NPVD as shown by a significantly steeper decline and lower survival probability over time. The green curve, representing patients with lower topic proportions of NPVD phenotype, shows a more gradual decline, reflecting a comparatively lower risk of mortality. (c) Effect size of the mortality-related phenotypes. We ran simple Cox regression per phenotype topic to obtain their marginal effect size and 95% confidence interval of the top 3 high risk and bottom 3 low risk mortality-associated phenotypes as identified by MixEHR-SurG in panel (a). Points indicate the coefficient values, Error bars show the 95% confidence intervals, and colors represent the significance levels of these coefficients. (d) Heatmap featuring the top ICD-9 codes from the three most positively predictive and three most negatively predictive phenotypes as determined by from MixEHR-SurG. The intensity of the colors indicates the topic probability in under each topic. The magnitude of the Cox coefficients are displayed in the last row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

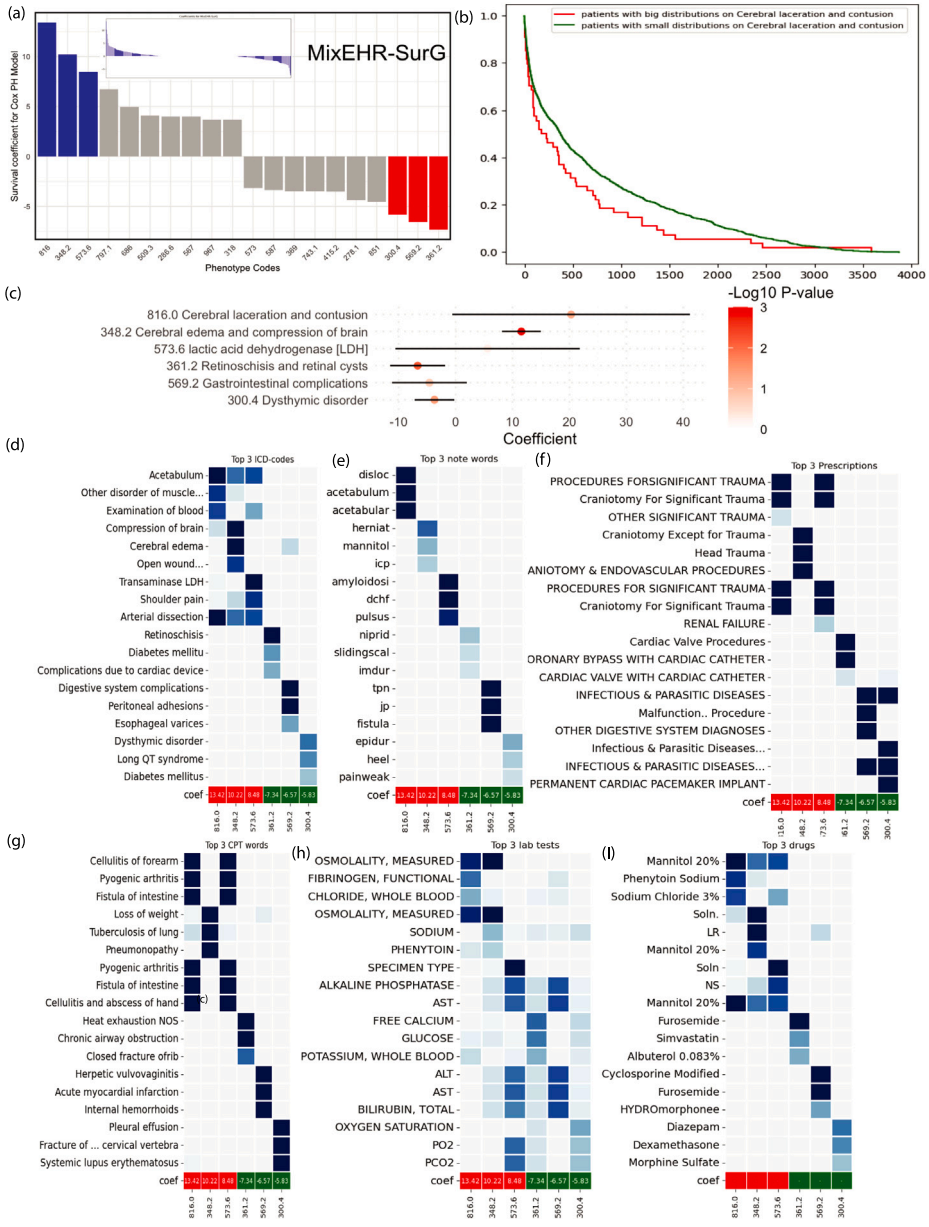


Fig. 6. Mortality-related multi-modal phenotypes for the ICU patients from MIMIC-III dataset. (a) Bar plot of survival regression coefficients $\mathbf{w}$. Inset in the upper-left corner that displays all estimated coefficients ω_k for $k = 1, \ldots, K$, organized from the largest to the smallest. We highlighted the top 10 coefficients with the largest effect sizes of positive and negative values. The positive coefficients are linked to phenotypes that elevate the risk of mortality, while the negative coefficients are associated with phenotypes that are predictive of lower mortality risk. The blue and red color highlight the top 3 and bottom 3 phenotypes, respectively, that we analyzed in-depth below. (b) Survival curves delineating two patient groups based on their distribution levels within the "Cerebral laceration and contusion" topic, which is the most significant predictor of high mortality risk. The red curve illustrates patients with a higher distribution in this topic, exhibiting a more pronounced decline in survival; the green curve, indicative of patients with a lower distribution, depicts a more gradual decrease in survival, pointing to a lower mortality risk. (c) Effect size of the mortality-related phenotypes. We ran simple Cox regression per phenotype topic to obtain their marginal effect size and 95% confidence interval of the top 3 high risk and bottom 3 low risk mortality-associated phenotypes as identified by MixEHR-SurG in panel (a). Points indicate the coefficient values, Error bars show the 95% confidence intervals, and colors represent the significance levels of these coefficients. (d)-(I) Heatmap showing the top ICD-9 codes, clinical note terms, CPT descriptors, medications, and lab tests under the moratality-related phenotypes. The color gradation indicate the prevalence of each feature within each phenotype topic. The last row indicates the Cox regression coefficients. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

inference models [38,39,43–46]. Such developments could further refine the precision of survival predictions and enhance the interpretability of complex healthcare data, ultimately leading to more informed and personalized medical decision-making. As the field advances, we anticipate that the integration of these sophisticated methodologies will yield models that not only predict but also disentangle the intricate network of disease causality within patient health trajectories [47–49].

6. Ethic

The use of the CHD data for this research was approved by the McGill University Health Centre Research Ethics Board.

CRediT authorship contribution statement

Yixuan Li: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation. **Archer Y. Yang:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Ariane Marelli:** Supervision, Resources, Funding acquisition, Data curation. **Yue Li:** Visualization, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Dr. Yue Li is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare, Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2016-05174). Dr. Archer Yi Yang is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2019-0621). Both Dr. Yang and Dr. Li are by the FRQNT Team Research Project Grant (FRQ-NT 327788). Dr. Ariane Marelli is supported by the Heart and Stroke Foundation grant and award (John Day, MD, Excellence Award for Heart Failure Trajectory Along the Care Continuum for Congenital Heart Disease Patients Across the Lifespan).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104638.

References

[1] J. Jiang, K. Qi, G. Bai, K. Schulman, Pre-pandemic assessment: a decade of progress in electronic health record adoption among US hospitals, Health Affairs Scholar 1 (5) (2023) qxad056.
[2] J.W. Smoller, The use of electronic health records for psychiatric phenotyping and genomics, Am. J. Med. Genet. B Neuropsychiatr. Genet. 177 (7) (2018) 601–612.
[3] H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, M. Alazab, A review of automatic phenotyping approaches using electronic health records, Electronics 8 (11) (2019) 1235.
[4] C. Shivade, P. Raghavan, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, A.M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inf. Assoc. 21 (2) (2014) 221–230.
[5] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Rev. Genet. 13 (6) (2012) 395–405.
[6] K. Jensen, C. Soguero-Ruiz, K. Oyvind Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S. Olav Skrovseth, K.M. Augestad, Analysis of free text in electronic health records for identification of cancer patient trajectories, Sci. Rep. 7 (1) (2017) 46226.
[7] M. Javaid, A. Haleem, R.P. Singh, R. Suman, S. Rab, Significance of machine learning in healthcare: Features, pillars and applications, Int. J. Intell. Netw. 3 (2022) 58–73.
[8] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (1) (2016) 1–10.
[9] R. Ranganath, A. Perotte, N. Elhadad, D. Blei, Deep survival analysis, in: Machine Learning for Healthcare Conference, PMLR, 2016, pp. 101–114.
[10] C. Lee, W. Zame, J. Yoon, M. Van Der Schaar, Deephit: A deep learning approach to survival analysis with competing risks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.
[11] S. Shin, P.C. Austin, H.J. Ross, H. Abdel-Qadir, C. Freitas, G. Tomlinson, D. Chicco, M. Mahendiran, P.R. Lawler, F. Billia, et al., Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality, ESC Heart Failure 8 (1) (2021) 106–115.
[12] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2) (1972) 187–202.
[13] H. Li, Y. Luan, Kernel cox regression models for linking gene expression profiles to censored survival data, in: Biocomputing 2003, World Scientific, 2002, pp. 65–76.
[14] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forests, 2008.
[15] R. Tibshirani, The lasso method for variable selection in the Cox model, Stat. Med. 16 (4) (1997) 385–395.
[16] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[17] H. Chen, S.M. Lundberg, S.-I. Lee, Explaining a series of models by propagating Shapley values, Nature Commun. 13 (1) (2022) 4512.
[18] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Mach. Intell. 2 (1) (2020) 56–67.
[19] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[20] Y. Li, P. Nair, X.H. Lu, Z. Wen, Y. Wang, A.A.K. Dehaghi, Y. Miao, W. Liu, T. Ordog, J.M. Biernacka, et al., Inferring multimodal latent topics from electronic health records, Nature Commun. 11 (1) (2020) 2536.
[21] Z. Song, X.S. Toral, Y. Xu, A. Liu, L. Guo, G. Powell, A. Verma, D. Buckeridge, A. Marelli, Y. Li, Supervised multi-specialist topic model with applications on large-scale electronic health record data, in: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2021, pp. 1–26.
[22] Z. Song, Y. Hu, A. Verma, D.L. Buckeridge, Y. Li, Automatic phenotyping by a seed-guided topic model, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4713–4723.
[23] Y. Ahuja, Y. Zou, A. Verma, D. Buckeridge, Y. Li, MixEHR-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record, J. Biomed. Inf. 134 (2022) 104190.
[24] J.A. Dawson, C. Kendziorski, Survival-supervised latent Dirichlet allocation models for genomic analysis of time-to-event outcomes, 2012, arXiv preprint arXiv:1202.5999.
[25] Y. Teh, D. Newman, M. Welling, A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, Adv. Neural Inf. Process. Syst. 19 (2006).
[26] W.-Q. Wei, L.A. Bastarache, R.J. Carroll, J.E. Marlo, T.J. Osterman, E.R. Gamazon, N.J. Cox, D.M. Roden, J.C. Denny, Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record, PLoS One 12 (7) (2017) e0175508.
[27] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for Cox's proportional hazards model via coordinate descent, J. Stat. Softw. 39 (5) (2011) 1.
[28] R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate Cox proportional hazards models, Stat. Med. 24 (11) (2005) 1713–1723.
[29] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[30] H. Uno, T. Cai, L. Tian, L.-J. Wei, Evaluating prediction rules for t-year survivors with censored regression models, J. Amer. Statist. Assoc. 102 (478) (2007) 527–537.
[31] H. Hung, C.-T. Chiang, Estimation methods for time-dependent AUC models with survival data, Canad. J. Statist. 38 (1) (2010) 8–26.
[32] J. Lambert, S. Chevret, Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves, Stat. Methods Med. Res. 25 (5) (2016) 2088–2102.
[33] A. Wakai, I.G. Roberts, G. Schierhout, Mannitol for acute traumatic brain injury, Cochrane Database Syst. Rev. (4) (2005).
[34] Q. Yuan, T. Cai, C. Hong, M. Du, B.E. Johnson, M. Lanuti, T. Cai, D.C. Christiani, Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer, JAMA Netw. Open 4 (7) (2021) e2114723.

[35] I.M. Baytas, K. Lin, F. Wang, A.K. Jain, J. Zhou, Phenotree: Interactive visual analytics for hierarchical phenotyping from large-scale electronic health records, IEEE Trans. Multimed. 18 (11) (2016) 2257–2270.
[36] R. Pivovarov, A.J. Perotte, E. Grave, J. Angiolillo, C.H. Wiggins, N. Elhadad, Learning probabilistic phenotypes from heterogeneous EHR data, J. Biomed. Inf. 58 (2015) 156–165.
[37] G. Defossez, A. Rollet, O. Dameron, P. Ingrand, Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer, BMC Med. Inf. Decis. Mak. 14 (1) (2014) 1–15.
[38] M.R. Bhat, M.A. Kundroo, T.A. Tarray, B. Agarwal, Deep LDA: A new way to topic model, J. Inf. Optim. Sci. 41 (3) (2020) 823–834.
[39] Z. Cao, S. Li, Y. Liu, W. Li, H. Ji, A novel neural topic model and its supervised extension, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29, No. 1, 2015.
[40] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, W. Buntine, Topic modelling meets deep neural networks: A survey, 2021, arXiv preprint arXiv:2103.00498.
[41] Y. Wang, R. Benavides, L. Diatchenko, A.V. Grant, Y. Li, A graph-embedded topic model enables characterization of diverse pain phenotypes among UK biobank individuals, Iscience 25 (6) (2022).
[42] Y. Zou, A. Pesaranghader, Z. Song, A. Verma, D.L. Buckeridge, Y. Li, Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model, Sci. Rep. 12 (1) (2022) 17868.
[43] V. Veitch, D. Sridhar, D. Blei, Adapting text embeddings for causal inference, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 919–928.
[44] H.D. Kim, M. Castellanos, M. Hsu, C. Zhai, T. Rietz, D. Diermeier, Mining causal topics in text data: iterative topic modeling with time series feedback, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 885–890.
[45] G.J. Rosa, B.D. Valente, G. de los Campos, X.-L. Wu, D. Gianola, M.A. Silva, Inferring causal phenotype networks using structural equation models, Genet. Select. Evol. 43 (1) (2011) 1–13.
[46] L.J. Corbin, V.Y. Tan, D.A. Hughes, K.H. Wade, D.S. Paul, K.E. Tansey, F. Butcher, F. Dudbridge, J.M. Howson, M.W. Jallow, et al., Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference, Nature Commun. 9 (1) (2018) 711.
[47] I.F. do Valle, B. Ferolito, H. Gerlovin, L. Costa, S. Demissie, F. Linares, J. Cohen, D.R. Gagnon, J.M. Gaziano, E. Begoli, et al., Network-medicine framework for studying disease trajectories in US veterans, Sci. Rep. 12 (1) (2022) 12018.
[48] X. Han, C. Hou, H. Yang, W. Chen, Z. Ying, Y. Hu, Y. Sun, Y. Qu, L. Yang, U.A. Valdimarsdóttir, et al., Disease trajectories and mortality among individuals diagnosed with depression: a community-based cohort study in UK biobank, Mol. Psychiatry 26 (11) (2021) 6736–6746.
[49] W. Oh, M.S. Steinbach, M.R. Castro, K.A. Peterson, V. Kumar, P.J. Caraballo, G.J. Simon, A computational method for learning disease trajectories from partially observable EHR data, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2476–2486.

Supplementary Materials
MixEHR-SurG: a joint proportional hazard and guided topic model for inferring mortality-associated topics from electronic health records

Yixuan Li[1,2], **Archer Y. Yang**[1,2,4,*], **Ariane Marelli**[3,*], **Yue Li**[2,4,*]

[1]Department of Mathematics and Statistics, McGill University, Montreal, Canada
[2]Mila - Quebec AI institute, Montreal, Canada
[3]McGill Adult Unit for Congenital Heart Disease (MAUDE Unit), McGill University of Health Centre, Montreal, Canada
[4]School of Computer Science, McGill University, Montreal, Canada
[*]Correspondence:
ariane.marelli@mcgill.ca, archer.yang@mcgill.ca, yueli@cs.mcgill.ca

S1. Notation table

Notation	Description
K	Total number of topics
M	Total number of EHR types
P	Total number of patients
$m \in \{1, \dots, M\}$	Index for EHR types
$V^{(m)}$	Total number of unique EHR features for document type m
$k \in \{1, \dots, K\}$	Index for topics
$j \in \{1, \dots, P\}$	Index for patient
$N_j^{(m)}$	Number of tokens in the EHR document of type m for patient j
$i \in \{1, \dots, N_j^{(m)}\}$	Index for tokens for patient j and document type m
$\boldsymbol{\pi}_j \in [0,1]^K$	Phenotype prior for patient j
$\boldsymbol{\theta}_j \in [0,1]^K$	Topic assignment for patient j
$\boldsymbol{\alpha} \in \mathbb{R}_+^K$	Hyperparameter for Dirichlet distribution of θ_j
$\phi_{kv}^{(m)} \in [0,1]$	Feature distribution of token with index v for topic k and document type m
$\boldsymbol{\phi}_k^{(m)} \in [0,1]^{V^{(m)}}$	Feature distribution for topic k and document type m
$\boldsymbol{\beta}^{(m)} \in \mathbb{R}_+^{V^{(m)}}$	Hyperparameter for Dirichlet distribution of $\boldsymbol{\phi}_k^{(m)}$
$x_{ji}^{(m)} \in \{1, \dots, V^{(m)}\}$	Word index of token i in the EHR document of type m for patient j
$z_{ji}^{(m)} \in \{1, \dots, K\}$	Latent topic assignment for token i in document m for patient j
$\gamma_{jik}^{(m)} \in [0,1]$	Variational probability of the k^{th} topic assignment for token i of EHR type m for patient j
$\bar{\mathbf{z}}_j \in [0,1]^K$	Average topic weight for patient j
$T_j \in \mathbb{R}_+$	Observed time for patient j
$\delta_j \in \{0,1\}$	Censoring status for patient j
$h_0(T_j)$	Baseline hazard function for patient j
$H_0(T_j)$	Baseline cumulative hazard function for patient j
$\mathbf{w} \in \mathbb{R}^K$	Cox PH regression coefficient
$\mathbf{T} \in \mathbb{R}_+^P$	Vector of observed times for all patients
$\boldsymbol{\delta} \in \{0,1\}^P$	Vector of censoring status for all patients
$\mathcal{X}^{(m)} = \left\{ \left\{ x_{ji}^{(m)} \right\}_{i=1}^{N_j} \right\}_{j=1}^{P}$	A set of P lists of word indices for all tokens of EHR type m for all patients
$\mathcal{X} = \{\mathcal{X}^{(m)}\}_{m=1}^{M}$	The entire EHR data over the M EHR types
$\mathcal{Z}^{(m)} = \left\{ \left\{ z_{ji}^{(m)} \right\}_{i=1}^{N_j} \right\}_{j=1}^{P}$	A set of P lists of topic indices for all tokens of EHR type m for all patients
$\mathcal{Z} = \{\mathcal{Z}^{(m)}\}_{m=1}^{M}$	The topic assignments of the entire EHR data over the M EHR types
$\boldsymbol{\pi} \in [0,1]^{P \times K}$	Matrix of phenotype priors for all patients
$\boldsymbol{\theta} \in [0,1]^{P \times K}$	Matrix of topic assignments for all patients
$\boldsymbol{\phi}^{(m)} \in [0,1]^{K \times V^{(m)}}$	Matrix of feature distributions for all topics of EHR type m
$\boldsymbol{\Phi} = \{\boldsymbol{\phi}^{(m)}\}_{m=1}^{M}$	List of feature distribution over the M EHR types
$\mathcal{B} = \{\boldsymbol{\beta}^{(m)}\}_{m=1}^{M}$	List of hyperparameters for Dirichlet distribution of $\boldsymbol{\phi}_k^{(m)}$
$\mathbf{U} \in \mathbb{R}^{P \times K}$	Matrix of PheCode counts for all P patients and K PheCodes
u_{jk}	Count of the k-th PheCode for the j-th patient

S2. Generative process the model variants

S2.1. Generative process for MixEHR

MixEHR follows the following generative process as illustrated in **Fig. ??**a:

1. Generate patient-specific topic assignment $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha}), j = 1, \ldots, P$
2. Generate the feature distribution $\boldsymbol{\phi}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(m)}\right)$ for topic $k = 1, \ldots, K$ and type $m = 1, \ldots, M$.
3. For each of the EHR token $x_{ji}^{(m)}$, $i = 1, \ldots, N_j^{(m)}$:
   (a) Generate a latent topic $z_{ji}^{(m)} \sim \text{Mul}(\boldsymbol{\theta}_j)$
   (b) Generate a specific token $x_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\phi}_{z_{ji}^{(m)}}^{(m)}\right)$

Generative process for MixEHR-G

The generative process for MixEHR-G is illustrated in **Fig. ??**b:

1. Obtain the phenotype prior $\boldsymbol{\pi}_j$ by a modified MAP [1] algorithm
2. Draw patient specific topic assignment $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha} \odot \boldsymbol{\pi}_j)$
3. Generate the feature distribution $\boldsymbol{\phi}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(m)}\right)$ for topic $k = 1, \ldots, K$ and type $m = 1, \ldots, M$.
4. For each of the EHR token $x_{ji}^{(m)}$, $i = 1, \ldots, N_j^{(m)}$:
   (a) Generate a latent topic $z_{ji}^{(m)} \sim \text{Mul}(\boldsymbol{\theta}_j)$
   (b) Generate a specific token $x_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\phi}_{z_{ji}^{(m)}}^{(m)}\right)$

Generative process for MixEHR-Surv

The generative process for MixEHR-Survival is illustrated in **Fig. ??**c:

1. Generate patient-specific topic assignment $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha})$
2. Generate the feature distribution $\boldsymbol{\phi}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(m)}\right)$ for topic $k = 1, \ldots, K$ and type $m = 1, \ldots, M$.
3. For each of the EHR token $x_{ji}^{(m)}$, $i = 1, \ldots, N_j^{(m)}$:
   (a) Generate a latent topic $z_{ji}^{(m)} \sim \text{Mul}(\boldsymbol{\theta}_j)$
   (b) Generate a specific token $x_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\phi}_{z_{ji}^{(m)}}^{(m)}\right)$
4. Compute the average topic proportion for each patient: $\bar{\mathbf{z}}_j = [\bar{z}_{jk}]_{k=1}^K = \left[\frac{\sum_{m=1}^M \sum_{i=1}^{N_j^{(m)}} \mathbb{I}(z_{ji}^{(m)}=k)}{\sum_{m=1}^M N_j^{(m)}}\right]_{k=1}^K$
5. Calculate the patient's hazard through the Cox proportional hazards model $h(T_j|\bar{\mathbf{z}}_j) = h_0(T_j)\exp\left\{\mathbf{w}^\top \bar{\mathbf{z}}_j\right\}$, and we could further visualize the survival curve or estimate survival time using the median survival time.

Generative process for MixEHR-SurG

The generative process for MixEHR-SurG is illustrated in **Fig. ??**d:

1. Obtain the phenotype prior $\boldsymbol{\pi}_j$ by a modified MAP [1] algorithm
2. Draw patient specific topic assignment $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha} \odot \boldsymbol{\pi}_j)$
3. Generate the feature distribution $\boldsymbol{\phi}_k^{(m)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(m)}\right)$ for topic $k = 1, \ldots, K$ and type $m = 1, \ldots, M$.
4. For each of the EHR token $x_{ji}^{(m)}$, $i = 1, \ldots, N_j^{(m)}$:

(a) Generate a latent topic $z_{ji}^{(m)} \sim \text{Mul}\left(\theta_j\right)$

(b) Generate a specific token $x_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\Phi}_{z_{ji}^{(m)}}^{(m)}\right)$

5. Compute the average topic weight for each patient:

$$\bar{\mathbf{z}}_j = [\bar{z}_{jk}]_{k=1}^K = \left[\frac{\sum_{m=1}^M \sum_{i=1}^{N_j^{(m)}} \mathbb{I}(z_{ji}^{(m)} = k)}{\sum_{m=1}^M N_j^{(m)}}\right]_{k=1}^K$$

6. Calculate the patient's hazard through the Cox proportional hazards model $h\left(T_j|\bar{\mathbf{z}}_j\right) = h_0\left(T_j\right)\exp\left\{\mathbf{w}^\top \bar{\mathbf{z}}_j\right\}$, we could further visualize the survival curve or estimate survival time using the median survival time.

S3. Computing PheCode topic priors

We compute $\pi_{jk} = p(y_{jk} = 1 \mid u_{jk})$ for each patient j and topic k in 3 steps:

- Step 1: After mapping each ICD code to its corresponding PheCode (`https://phewascatalog.org/phecodes`), we calculate the PheCode counts u_{jk} for each patient, denoted by j, where $j = 1, \ldots, P$, across each PheCode, denoted by k, where $k = 1, \ldots, K$. It's important to note that for a patient who encounters the same PheCode multiple times, either due to repeated ICD code mappings or multiple healthcare visits, each instance is individually accounted for. This approach results in the possibility of accruing multiple counts for the same PheCode for a single patient. As a result, we convert the $P \times V^{(\text{ICD})}$ to a $P \times K$ matrix $\boldsymbol{U} = [u_{jk}]_{P\times K}$. We then infer the posterior distribution of y_{jk} in two parallel ways.

- Step 2A (Model A): Assuming that the counts for a PhenoCode k follows a Poisson distribution with parameters π_{jk}, ρ_0 and ρ_1. The Poisson likelihood takes the following form:

$$P\left(u_{jk}\right) = \pi_{jk}\frac{\left(\rho_1\right)^{u_{jk}} e^{-\rho_1}}{u_{jk}!} + \left(1 - \pi_{jk}\right)\frac{\left(\rho_0\right)^{u_{jk}} e^{-\rho_0}}{u_{jk}!}, \tag{1}$$

  where π_{jk} corresponds to the foreground Poisson component with larger mean ρ_1 and and $1 - \pi_{jk}$ corresponds to the population background Poisson with lower mean ρ_0. Given data $\{u_{jk}\}_{j=1}^P$, we perform expectation-maximization (EM) algorithm: in the E-step, we infer the posterior probability $\widehat{\pi}_{jk} = \hat{p}(y_{jk} = 1|u_{jk})$ and in the M-step, we maximize the likelihood with respect to ρ_1 and ρ_0.

- Step 2B (Model B): Alternatively, we can assume that for each PheCode k, the log-transformed count data $g(u_{1k}), \ldots, g(u_{Pk})$, with $g(u) = \log(u) + 1$ follows a two-component univariate Gaussian mixture model:

$$P\left(g(u_{jk}) = x\right) = \frac{\pi'_{jk}}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-\pi'_{jk}}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right) \tag{2}$$

  We then perform EM algorithm to alternate between inferring $\widehat{\pi}'_{jk} = \hat{p}(y'_{jk} = 1|u_{jk})$ and computing maximum likelihood estimates for the Gaussian parameters.

- Step 3: The prior probability for a patient j having phenotype k is set to $\pi_{jk} = \frac{1}{2}\left(\widehat{\pi}_{jk} + \widehat{\pi}'_{jk}\right)$.

In the application of the MIMIC-III data, as it is not a longitudinal dataset, each PheCode was documented no more than once for each patient. In this case, we assigned the hyperparameters π_{jk} for each phenotype k as either one or zero, based on whether the corresponding PheCode was observed or not for patient j, respectively.

S4. Details of stochastic joint collapsed variational Bayesian inference

First, we derive the joint-likelihood function of all the parameters for observational data and latent variables conditioned on priors and survival regression coefficients for MixEHR-SurG (Fig **??**d) model:

$$\begin{aligned} & p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z}, \theta, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right) \\ = & \underbrace{p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right)}_{\text{supervised part}} \underbrace{p\left(\mathcal{X}, \mathcal{Z}, \theta, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right)}_{\text{unsupervised part}} \end{aligned}$$

where for the survival supervised part, we use the Cox proportional hazards (PH) model with elastic net penalization for the survival coefficients. The full likelihood function of the penalized Cox PH model is obtained by incorporating Breslow's estimate of the baseline hazard function.

$$\begin{aligned} & p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) \\ = & \prod_{j=1}^{P} p\left(T_j, \delta_j \mid \bar{\mathbf{z}}_j, h_0(T_j), \mathbf{w}\right) \\ = & \prod_{j=1}^{P}\left[h\left(T_j, \bar{\mathbf{z}}_j\right)\right]^{\delta_j} S\left(T_j, \bar{\mathbf{z}}_j\right) \exp\left\{-\lambda_2\|\mathbf{w}\|_2^2-\lambda_1\|\mathbf{w}\|_1\right\} \\ = & \prod_{j=1}^{P}\left\{\left[h_0\left(T_j\right) \exp\left(\mathbf{w}^{\top} \bar{\mathbf{z}}_j\right)\right]^{\delta_j} \times \exp\left[-H_0\left(T_j\right) \exp\left(\mathbf{w}^{\top} \bar{\mathbf{z}}_j\right)\right]\right\} \exp\left\{-\lambda_2\|\mathbf{w}\|_2^2-\lambda_1\|\mathbf{w}\|_1\right\}. \end{aligned}$$

Here $H_0(t)$ denotes the cumulative baseline hazard function, obtained by the integral of the baseline hazard function between integration limits of 0 and t as $H_0(t)=\int_0^t h_0(u)\, du$. The elastic net penalty terms including $\|\mathbf{w}\|_2^2=\sum_k w_k^2$ and $\|\mathbf{w}\|_1=\sum_k|w_k|$ consist of the L2 and L1 regularization term weighted by the hyperparameters λ_2 and λ_1, respectively.

We will use the collapsed variational inference algorithm to integret out θ and $\boldsymbol{\Phi}$ in the joint likelihood function to achieve more accurate and efficient inference [2]. This is due to the conjugacy of Dirichlet variables θ and $\boldsymbol{\Phi}$ to the multinomial likelihood variables $\mathcal{X}$ and $\mathcal{Z}$.

$$\begin{aligned} & p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right) \\ = & p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right) \\ = & p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) \int\int p\left(\mathcal{X}, \mathcal{Z}, \theta, \boldsymbol{\Phi} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right) d\boldsymbol{\Phi} d\theta \\ = & p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) \int\int p\left(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Phi}\right) p\left(\boldsymbol{\Phi} \mid \mathcal{B}\right) p\left(\mathcal{Z} \mid \theta\right) p\left(\theta \mid \boldsymbol{\alpha}, \boldsymbol{\pi}\right) d\boldsymbol{\Phi} d\theta \\ = & p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) \int p\left(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Phi}\right) p\left(\boldsymbol{\Phi} \mid \mathcal{B}\right) d\boldsymbol{\Phi} \times \int p\left(\mathcal{Z} \mid \theta\right) p\left(\theta \mid \boldsymbol{\alpha}, \boldsymbol{\pi}\right) d\theta \end{aligned}$$

Upon substituting the distributions outlined in the generative process of MixEHR-SurG, as

detailed in **Methods** S2, the integral can be evaluated as follows:

$$
\begin{aligned}
&\int p\left(\mathcal{Z} \mid \theta\right) p\left(\theta \mid \boldsymbol{\alpha}, \boldsymbol{\pi}\right) d\theta \\
=&\int \left(\prod_{j=1}^{P}\prod_{k=1}^{K} \theta_{jk}^{n_{j\bullet k}^{(\bullet)}}\right) \times \left(\prod_{j=1}^{P} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)} \prod_{k=1}^{K} \theta_{jk}^{\alpha_k \pi_{jk}-1}\right) d\theta \\
=&\prod_{j=1}^{P} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)} \int \left(\prod_{k=1}^{K} \theta_{jk}^{\alpha_k \pi_{jk}-1+n_{j\bullet k}^{(\bullet)}}\right) d\theta \\
=&\prod_{j=1}^{P} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}{\Gamma\left(\sum_{k=1}^{K} \boldsymbol{\alpha}_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}
\end{aligned}
$$

$$
\begin{aligned}
&\int p\left(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Phi}\right) p\left(\boldsymbol{\Phi} \mid \mathcal{B}\right) d\boldsymbol{\Phi} \\
=&\int \left(\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{v=1}^{V^{(m)}} \phi_{vk}^{(m) n_{\bullet vk}^{(m)}}\right) \times \left(\prod_{m=1}^{M}\prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}} \Gamma\left(\beta_v^{(m)}\right)} \prod_{v=1}^{V^{(m)}} \phi_{vk}^{(m)\beta_v^{(m)}-1}\right) d\boldsymbol{\Phi} \\
=&\prod_{m=1}^{M}\prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}} \Gamma\left(\beta_v^{(m)}\right)} \int \left(\prod_{v=1}^{V^{(m)}} \phi_{vk}^{(m)\beta_v^{(m)}-1+n_{\bullet vk}^{(m)}}\right) d\boldsymbol{\Phi} \\
=&\prod_{k=1}^{K}\prod_{m=1}^{M} \frac{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)}\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v^{(m)}\right)} \frac{\prod_{v=1}^{V^{(m)}} \Gamma\left(\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}
\end{aligned}
$$

where the coordinate sufficient statistics are:

$$
n_{\bullet vk}^{(m)} = \sum_{j=1}^{P}\sum_{i=1}^{N_j^{(m)}} \mathbb{I}\left[x_{ji}^{(m)} = v, z_{ji}^{(m)} = k\right]
$$

$$
n_{j\bullet k}^{(\bullet)} = \sum_{m=1}^{M}\sum_{i=1}^{N_j^{(m)}} \mathbb{I}\left[z_{ji}^{(m)} = k\right]
$$

Thus, we have:

$$
\begin{aligned}
&p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right) \\
=&\prod_{k=1}^{K}\prod_{m=1}^{M} \frac{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}} \Gamma\left(\beta_v^{(m)}\right)} \frac{\prod_{v=1}^{V^{(m)}} \Gamma\left(\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(m)}} \beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)} \prod_{j=1}^{P} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)} \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}{\Gamma\left(\sum_{k=1}^{K} \boldsymbol{\alpha}_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}
\end{aligned}
$$

Then, we will derive the evidence lower bound (ELBO) for the current marginal distribution for the observational data as follows:

$$\begin{aligned}\mathcal{L}_{ELBO} &\equiv \mathbb{E}_{q(\mathcal{Z})} \log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right) - \mathbb{E}_{q(\mathcal{Z})} \log q\left(\mathcal{Z}\right) \\ &= \sum_{\mathcal{Z}} q\left(\mathcal{Z}\right) \log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right) - \sum_{\mathcal{Z}} q\left(\mathcal{Z}\right) \log q\left(\mathcal{Z}\right)\end{aligned}$$

Maximizing $\mathcal{L}_{ELBO}$ is equivalent to minimizing the Kullback–Leibler (KL) divergence, as they sum up as the joint distribution of the observational data which is a constant:

$$\begin{aligned}\mathcal{KL}[q(\mathcal{Z}) \| p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z}\right)] &= \mathbb{E}_{q(\mathcal{Z})} \log q\left(\mathcal{Z}\right) - \mathbb{E}_{(\mathcal{Z})} \log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right) + \log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}\right) \\ &= -\mathcal{L}_{ELBO} + \log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}\right)\end{aligned}$$

The mean-field assumption pertains only to word-specific topic assignments $\mathcal{Z}$, which have the proposed distribution under the variational parameter $\gamma_{jik}^{(m)}$ as defined below:

$$q(\mathcal{Z}) = \prod_{m=1}^{M} \prod_{j=1}^{P} \prod_{i=1}^{N_j^{(m)}} q(z_{ji}^{(m)} \mid \gamma_{ji1}^{(m)}, \ldots, \gamma_{jiK}^{(m)}) = \prod_{m=1}^{M} \prod_{j=1}^{P} \prod_{i=1}^{N_j^{(m)}} \prod_{k=1}^{K} {\gamma_{jik}^{(m)}}^{\mathbb{I}[z_{ji}^{(m)}=k]}$$

Under the mean-field assumption, maximizing the ELBO with respect to $\gamma_{jik}^{(m)}$ is equivalent to calculating the variational expectation $\mathbb{E}_{q(\mathcal{Z})}[z_{ji}^{(m)} = k]$ conditioned on the variational expected value for other tokens [3, 4]. The coordinate ascent update has an approximate closed-form expression as derived below:

$$\begin{aligned}\gamma_{jik}^{(m)} &= \frac{\exp\left\{\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right)\right]\right\}}{\exp\left\{\int \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right)\right] dz_{ji}^{(m)}\right\}} \\ &\propto \exp\left\{\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right)\right]\right\}\end{aligned}$$

Then we aximizing the ELBO with respect to $\gamma_{jik}^{(m)}$,

$$
\begin{aligned}
\log \gamma_{jik}^{(m)} &= \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(\mathbf{T}, \boldsymbol{\delta}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}, h_0(\cdot), \mathbf{w}\right)\right] + \text{const} \\
&= \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log\left(p\left(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}\right) p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right)\right)\right] + \text{const} \\
&= \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(T_j, \delta_j \mid z_{(j,-i)}^{(m)}, z_{ji}^{(m)} = k, h_0(\cdot), \mathbf{w}\right)\right] \\
&\quad + \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right)\right] + \text{const} \\
&= \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(T_j, \delta_j \mid z_{(j,-i)}^{(m)}, z_{ji}^{(m)} = k, h_0(\cdot), \mathbf{w}\right)\right] \\
&+ \mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log\left(\prod_{k=1}^{K}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)}\right)}\frac{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}\right.\right. \\
&\quad \left.\left.\prod_{j=1}^{P}\frac{\Gamma\left(\sum_k \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)}\frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}\right)\right] + \text{const}
\end{aligned}
$$

Thus, we calculate the expontential spontaneously at both side

$$
\begin{aligned}
\gamma_{jik}^{(m)} \propto \exp&\left\{\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log p\left(T_j, \delta_j \mid z_{(j,-i)}^{(m)}, z_{ji}^{(m)} = k, h_0(\cdot), \mathbf{w}\right)\right]\right\} \\
&\exp\left\{\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log\left(\prod_{k=1}^{K}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)}\right)}\frac{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)} + n_{\bullet vk}^{(m)}\right)}\right.\right.\right. \\
&\left.\left.\left.\prod_{j=1}^{P}\frac{\Gamma\left(\sum_k \alpha_k \boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k \boldsymbol{\pi}_j\right)}\frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k \boldsymbol{\pi}_j + n_{j\bullet k}^{(\bullet)}\right)}\right)\right]\right\}
\end{aligned}
$$

where the footnote $(j, -i)$ denote when we calculating the coordinate sufficient statistics, we exclude the variable with index ji.

We choose the survival model as the Cox proportional hazards model. The corresponding hazard function and survival function could be written as

$$
h\left(T_j, \bar{\mathbf{z}}_j\right) = h_0\left(T_j\right)\exp\left(\mathbf{w}^\top \bar{\mathbf{z}}_j\right)
$$

and

$$
S\left(T_j, \bar{\mathbf{z}}_j\right) = \exp\left[-H_0\left(T_j\right)\exp\left(\mathbf{w}^\top \bar{\mathbf{z}}_j\right)\right]
$$

respectively. The vector $\mathbf{w} \in \mathbb{R}^K$ contains the survival coefficients, and $h_0\left(T_j\right)$ is the baseline hazard at time T_j. $H_0\left(T_j\right)$ denotes the cumulative hazard at time T_j, which is obtained by the integral of the baseline hazard function between integration limits of 0 and t as $H_0(t) = \int_0^t h_0(u)\, du$.

Under those settings, we could further derive the supervised part as follows:

$$
\begin{aligned}
&\mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\log p\left(T_j, \delta_j \mid z^{(m)}_{(j,-i)}, z^{(m)}_{ji} = k, h_0(\cdot), \mathbf{w}\right)\right] \\
&\overset{(i)}{=} \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\log p\left(T_j, \delta_j \mid \bar{\mathbf{z}}^{(m)}_{(j,-i)}, \bar{\mathbf{z}}^{(m)}_{ji}, h_0(\cdot), \mathbf{w}\right)\right] \\
&= \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\log\left(h\left(T_j, \bar{\mathbf{z}}^{(m)}_{(j,-i)}, \bar{\mathbf{z}}^{(m)}_{ji}\right)^{\delta_j} S\left(T_j, \bar{\mathbf{z}}^{(m)}_{(j,-i)}, \bar{\mathbf{z}}^{(m)}_{ji}\right)\right)\right] \\
&= \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\delta_j \log h_0\left(T_j\right) + \delta_j \mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)} + \delta_j \mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{ji} - H_0\left(T_j\right)\exp\left(\mathbf{w}^\top\left(\bar{\mathbf{z}}^{(m)}_{(j,-i)} + \bar{\mathbf{z}}^{(m)}_{ji}\right)\right)\right] \\
&\overset{(ii)}{=} \delta_j \log h_0\left(T_j\right) + \delta_j \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)}\right] + \delta_j \frac{w_k}{N^{(m)}_j} - H_0\left(T_j\right)\mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\exp\left(\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)} + \frac{w_k}{N^{(m)}_j}\right)\right] \\
&= \delta_j \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)}\right] + \delta_j \frac{w_k}{N^{(m)}_j} - H_0\left(T_j\right)\mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\exp\left(\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)}\right)\right]\exp\left(\frac{w_k}{N^{(m)}_j}\right) + \text{const} \\
&\overset{(iii)}{\approx} \delta_j \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)}\right] + \delta_j \frac{w_k}{N^{(m)}_j} - H_0\left(T_j\right)\mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{(j,-i)} + 1\right]\exp\left(\frac{w_k}{N^{(m)}_j}\right) + \text{const} \\
&\approx \delta_j \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{j}\right] + \delta_j \frac{w_k}{N^{(m)}_j} - H_0\left(T_j\right)\mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbf{w}^\top \bar{\mathbf{z}}^{(m)}_{j} + 1\right]\exp\left(\frac{w_k}{N^{(m)}_j}\right) + \text{const} \\
&\overset{(iv)}{=} \delta_j \mathbf{w}^\top \bar{\boldsymbol{\gamma}}^{(m)}_j + \delta_j \frac{w_k}{N^{(m)}_j} - H_0\left(T_j\right)\left(\mathbf{w}^\top \bar{\boldsymbol{\gamma}}^{(m)}_j + 1\right)\exp\left(\frac{w_k}{N^{(m)}_j}\right) + \text{const}
\end{aligned}
$$

The equation (i) follows by defining

$$
\bar{\mathbf{z}}^{(m)}_{ji} = \left[\frac{\mathbb{I}\left(z^{(m)}_{ji} = k\right)}{N^{(m)}_j}\right]_{k=1}^{K},
$$

and

$$
\bar{\mathbf{z}}^{(m)}_{(j,-i)} = \left[\frac{\sum_{i'=1}^{N^{(m)}_j}\mathbb{I}\left(\left(z^{(m)}_{ji'} = k\right) \cap \left(i' \neq i\right)\right)}{N^{(m)}_j}\right]_{k=1}^{K}.
$$

The equation (ii) follows by

$$
\left[\bar{\mathbf{z}}^{(m)}_{ji}\right]_k = \frac{\mathbb{I}\left(z^{(m)}_{ji} = k\right)}{N^{(m)}_j} = \frac{1}{N^{(m)}_j}
$$

and

$$
\left[\bar{\mathbf{z}}^{(m)}_{ji}\right]_{k'} = \frac{\mathbb{I}\left(z^{(m)}_{ji} = k'\right)}{N^{(m)}_j} = 0,
$$

for $k' \neq k$, since $z^{(m)}_{ji} = k$.

The approximation (iii) is due to the first-order Taylor series of the exponential term $\exp\left(\mathbf{w}^\top \left[\bar{\mathbf{z}}^{(m)}_{(j,-i)}\right]_j\right)$. Note that the exponential function can be approximated by Taylor series as $\exp(x) = 1 + x + x^2/2! + x^2/3! + \ldots$. For computational efficiency, we only took the first order of the Taylor series, which correspond to the first two terms $1 + x$.

The equation (iv) follows by defining:

$$
\begin{aligned}
\bar{\boldsymbol{\gamma}}^{(m)}_j = [\bar{\gamma}^{(m)}_{jk}]_{k=1}^K &= \left[\frac{\sum_{i=1}^{N_j^{(m)}} \gamma^{(m)}_{jik}}{N_j^{(m)}}\right]_{k=1}^K \\
&= \left[\frac{\sum_{i=1}^{N_j^{(m)}} \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\mathbb{I}\left(z^{(m)}_{ji} = k\right)\right]}{N_j^{(m)}}\right]_{k=1}^K \\
&= \mathbb{E}_{q\left(z^{(m)}_{(j,-i)}\right)}\left[\bar{\mathbf{z}}^{(m)}_j\right]
\end{aligned}
$$

And the expectation of the unsupervised part could be derived as:

$$
\begin{aligned}
&\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log\left(\prod_{k=1}^{K}\prod_{m=1}^{M}\frac{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}\right)}{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)}\right)}\frac{\prod_{v=1}^{V^{(m)}}\Gamma\left(\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)}{\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)}\right.\right.\\
&\left.\left.\times\prod_{j=1}^{P}\frac{\Gamma\left(\sum_k\alpha_k\boldsymbol{\pi}_j\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_k\boldsymbol{\pi}_j\right)}\frac{\prod_{k=1}^{K}\Gamma\left(\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)}\right)\right]\\
&=\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\sum_{k=1}^{K}\sum_{m=1}^{M}\log\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}\right)-\sum_{v=1}^{V^{(m)}}\log\Gamma\left(\beta_v^{(m)}\right)\right.\\
&\quad\left.+\sum_{v=1}^{V^{(m)}}\log\Gamma\left(\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)-\log\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)\right]\\
&\quad+\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\sum_{j=1}^{P}\log\Gamma\left(\sum_k\alpha_k\boldsymbol{\pi}_j\right)-\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\boldsymbol{\pi}_j\right)\right.\\
&\quad\left.+\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)-\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)\right]\\
&=\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\sum_{v=1}^{V^{(m)}}\log\Gamma\left(\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)-\log\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)\right.\\
&\quad\left.+\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)-\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)\right]+\text{const}\\
&=\mathbb{E}_{q\left(z_{(j,-i)}^{(m)}\right)}\left[\log\Gamma\left(\beta_{x_{ji}^{(m)}}^{(m)}+n_{\bullet x_{ji}^{(m)}k}^{(m)}\right)-\log\Gamma\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+n_{\bullet vk}^{(m)}\right)\right.\\
&\quad\left.+\log\Gamma\left(\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)-\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\boldsymbol{\pi}_j+n_{j\bullet k}^{(\bullet)}\right)\right]+\text{const}\\
&\overset{(i)}{=}\log\left(\beta_{x_{ji}^{(m)}}^{(m)}+\left[n_{\bullet x_{ji}^{(m)}k}^{(m)}\right]_{(-j,-i)}\right)-\log\left(\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+\left[n_{\bullet vk}^{(m)}\right]_{(-j,-i)}\right)\\
&\quad+\log\left(\alpha_k\boldsymbol{\pi}_j+\left[n_{j\bullet k}^{(\bullet)}\right]_{(j,-i)}\right)-\log\left((\sum_{k=1}^{K}\alpha_k)\pi_j+\sum_{k=1}^{K}\left[n_{j\bullet k}^{(\bullet)}\right]_{(j,-i)}\right)\\
&\quad+\text{const}\\
&=\log\left(\left(\alpha_k\boldsymbol{\pi}_j+\left[n_{j\bullet k}^{(\bullet)}\right]_{(j,-i)}\right)\frac{\left(\beta_{x_{ji}^{(m)}}^{(m)}+\left[n_{\bullet x_{ji}^{(m)}k}^{(m)}\right]_{(-j,-i)}\right)}{\sum_{v=1}^{V^{(m)}}\beta_v^{(m)}+\left[n_{\bullet vk}^{(m)}\right]_{(-j,-i)}}\right)+\text{const}
\end{aligned}
$$

The equation (i) follows by defining the first term as

$$\left[n^{(m)}_{\bullet x^{(m)}_{ji} k}\right]_{(-j,-i)} = \sum_{j'=1}^{P}\sum_{i'=1}^{N_j^{(m)}} \mathbb{I}\left[(x^{(m)}_{j'i'} = x^{(m)}_{ji}, z^{(m)}_{j'i'} = k) \cap (j' \neq j, i' \neq i)\right],$$

the second term as

$$\left[n^{(m)}_{\bullet vk}\right]_{(-j,-i)} = \sum_{j'=1}^{P}\sum_{i'=1}^{N_j^{(m)}} \mathbb{I}\left[(x^{(m)}_{j'i'} = v, z^{(m)}_{j'i'} = k) \cap (j' \neq j, i' \neq i)\right],$$

the third and the forth term as

$$\left[n^{(\bullet)}_{j\bullet k}\right]_{(j,-i)} = \sum_{m=1}^{M}\sum_{i'=1}^{N_j^{(m)}} \mathbb{I}\left[(z^{(m)}_{ji'} = k) \cap (i' \neq i)\right].$$

Finally we will get the estimation of the closed-form latent variational expectation update of $\gamma^{(m)}_{jik}$ after calculating the following and normalizing afterwards:

$$\begin{aligned}
\gamma^{(m)}_{jik} \propto & \exp\left(\left(\delta_j \mathbf{w}^\top \bar{\boldsymbol{\gamma}}^{(m)}_j\right)\left(\delta_j \frac{w_k}{N_j^{(m)}}\right)\right) \\
& \times \exp\left[-H_0\left(T_j\right)\left(\mathbf{w}^\top \bar{\boldsymbol{\gamma}}^{(m)}_j + 1\right)\exp\left(\frac{w_k}{N_j^{(m)}}\right)\right] \\
& \times \left(\alpha_k \boldsymbol{\pi}_j + \left[n^{(\bullet)}_{j\bullet k}\right]_{(j,-i)}\right)\frac{\left(\beta^{(m)}_{x^{(m)}_{ji}} + \left[n^{(m)}_{\bullet x^{(m)}_{ji} k}\right]_{(-j,-i)}\right)}{\sum_{v=1}^{V^{(m)}} \beta^{(m)}_v + \left[n^{(m)}_{\bullet vk}\right]_{(-j,-i)}}
\end{aligned}$$

Furthermore, we update the hyperparameters $\boldsymbol{\alpha}$ and $\mathcal{B}$ by maximizing the marginal log likelihood function under the estimate of the expectation of the variational parameter. Noting that $\boldsymbol{\alpha}$ and $\mathcal{B}$ only participate in the unsupervised term of the ELBO, the closed-form update can be derived by the fixed point process [5]:

$$\alpha^*_k = \arg\max_{\alpha_k} \mathbb{E}_{q(\mathcal{Z})}\left[p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right)\right] \tag{3}$$

$$= \frac{a_\alpha - 1 + \alpha_k \sum_{j=1}^{P} \Psi\left(\alpha_k + n^{(\bullet)}_{j\bullet k}\right) - \Psi\left(\alpha_k\right)}{b_\alpha + \sum_{j=1}^{P} \Psi\left(\sum_{k=1}^{K} \alpha_k + n^{(\bullet)}_{j\bullet k}\right) - \Psi\left(\sum_{k=1}^{K} \alpha_k\right)} \tag{4}$$

$$\beta^{(m)*}_v = \arg\max_{\beta^{(m)}_v} \mathbb{E}_{q(\mathcal{Z})}\left[p\left(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathcal{B}\right)\right] \tag{5}$$

$$= \frac{a_\beta - 1 + \beta^{(m)}_v \left(\sum_{k=1}^{K} \Psi\left(\beta^{(m)}_v + n^{(m)}_{\bullet vk}\right)\right) - K V^{(m)} \Psi\left(\beta^{(m)}_v\right)}{b_\beta + \sum_{k=1}^{K} \Psi\left(V^{(m)} \beta^{(m)}_v + \sum_{v=1}^{V^{(m)}} n^{(m)}_{\bullet vk}\right) - K \Psi\left(V^{(m)} \beta^{(m)}_v\right)} \tag{6}$$

To update the survival-relevant parameters $\boldsymbol{w}$ and $h_0(\cdot)$, we focus on maximizing the components related to these parameters within the ELBO. This maximization is conditioned on the expected values of the latent variables $\mathcal{Z}$:

$$(\mathbf{w}, h_0(\cdot)) = \underset{\mathbf{w}, h_0(\cdot)}{\arg\max} \mathbb{E}_{q(\mathcal{Z})} p(\mathbf{T}, \boldsymbol{\delta} \mid \mathcal{Z}, h_0(\cdot), \mathbf{w}) \tag{7}$$

$$= \underset{\mathbf{w}, h_0(\cdot)}{\arg\max} \sum_{j=1}^{P} \Bigg\{ \delta_j \log h_0(T_j) + \delta_j \mathbf{w}^\top \mathbb{E}_{q(\mathcal{Z})} [\bar{\mathbf{z}}_j] \tag{8}$$

$$- H_0(T_j) \exp\left(\mathbf{w}^\top \mathbb{E}_{q(\mathcal{Z})} [\bar{\mathbf{z}}_j] \right) \Bigg\} - \lambda_2 \|\mathbf{w}\|_2^2 - \lambda_1 \|\mathbf{w}\|_1 \tag{9}$$

$$= \underset{\mathbf{w}, h_0(\cdot)}{\arg\max} \sum_{j=1}^{P} \Bigg\{ \delta_j \log h_0(T_j) + \delta_j \mathbf{w}^\top \bar{\boldsymbol{\gamma}}_j \tag{10}$$

$$- H_0(T_j) \exp\left(\mathbf{w}^\top \bar{\boldsymbol{\gamma}}_j \right) \Bigg\} - \lambda_2 \|\mathbf{w}\|_2^2 - \lambda_1 \|\mathbf{w}\|_1 \tag{11}$$

Above formula mirrors the coefficients estimates employed in the Cox proportional hazards regression with elastic net penalization, which combines both L1 and L2 norms for regularization [6] . In this context, $\bar{\boldsymbol{\gamma}}_j$ function as covariates, while $[T_j, \delta_j]_{j=1}^P$ provide the survival information. The update of $\boldsymbol{w}$ and $h_0(\cdot)$ is facilitated using the scikit-survival [7] Python module, a tool specifically designed for handling such statistical computations in survival analysis.

The whole collapsed variational Inference algorithm for MixEHR-SurG is in Algorithm 1.

Algorithm 1: Collapsed Variational Inference for MixEHR-SurG

Initialization:

$\alpha_k \sim \text{Gamma}(a, b)$ for $k = 1, \ldots, K$

$\beta_v^{(m)} \sim \text{Gamma}(c, d)$ for $v = 1, \ldots, V$ and $m = 1, \ldots, M$

$\gamma_{jik}^{(m)} \sim \text{Unif}(0, 1)$ for all i, j, k, m

Normalize $\gamma_{jik}^{(m)}$ to sum to 1 over k

repeat

E-Step:

for $m = 1, \ldots, M$ **do**

for $j = 1, \ldots, P$ **do**

for $i = 1, \ldots, N_j^{(m)}$ **do**

for $k = 1, \ldots, K$ **do**

Update $\gamma_{jik}^{(m)}$ using Eq. (**??**)

end

Normalize $\gamma_{jik}^{(m)}$ to sum to 1 over k

end

end

end

M-Step:

for $k = 1, \ldots, K$ **do**

Update α_k using Eq. (3)

end

for $m = 1, \ldots, M$ **do**

for $v = 1, \ldots, V^{(m)}$ **do**

Update $\beta_v^{(m)}$ using Eq. (5)

end

end

Estimate $\boldsymbol{w}, h_0(\cdot)$ by Eq. (7) using Coxnet with updated $\bar{\gamma}_j$ as covariates, and survival data $[T_j, \delta_j]_{j=1}^P$.

until ***Converge***;

S5. Evaluating causal phenotypes in simulation study

For the quantitative evaluation of MixEHR-SurG, we first focused on assessing its capability to identify mortality-related topics. In the simulation section, we used Receiver Operating Characteristic (ROC) curve, a widely-used metric in machine learning to evaluate the variable selection performance of our models. The ROC curve is the true positive rate TPR=TP/(TP+FN) as a function of the false positive rate FPR=FP/(FP+TN) in variable selection, where TP, FP, FN, TN are true positive, false positive, false negative, and true negative, respectively. In our context, this involves comparing the estimated survival coefficients of the simulation data set with the ground truth coefficients we predefined (i.e., 50 survival-related topics with a coefficient of 6, and all others set to 0).

S6. Survival analysis

From $\mathbf{w}$ learned by MixEHR-SurG, we selected the top 3 and bottom 3 survival-related phenotypes with the largest positive and negative coefficients, respectively. To assess the statistical significance of each coefficient w_k, we conducted chi-square tests against the null hypothesis that $w_k = 0$ [8]. Specifically, we divided patients into two groups based on their topic proportion. For the phenotype with the highest survival coefficient, denoted as $k_{\max} = \arg\max_k w_k$, we empirically determined the threshold to be the top 30% percentile of the topic mixture probabilities such that patients above the percentile were assigned to one group and the rest of the patients were assigned to the other group (**Fig. ??**b and **Fig. ??**b). We then computed the chi-squared test p-values using the `survival` R package [9] (**Fig. ??**c and **Fig. ??**c).

S7. Supplementary Figures

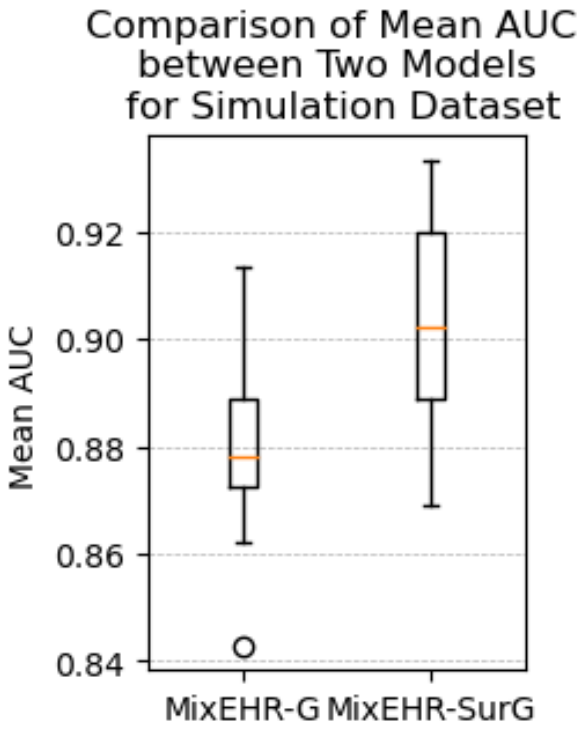


Figure S1: Comparison of the mean AUC between the pipeline MixEHR-G+Coxnet and MixEHR-SurG based on 10 simulated datasets.

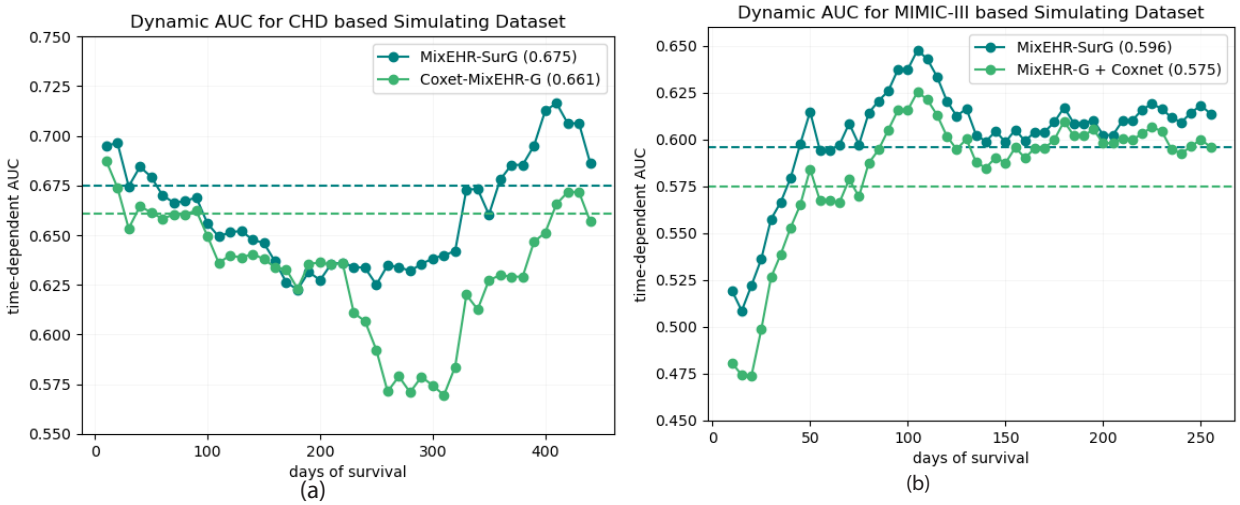


Figure S2: Dynamic AUC curves for predicting time to death in patients from the simulated data. (a) Dynamic AUC curves for predicting time to death in patients from simulating dataset based on the CHD dataset. (b) Dynamic AUC curves for predicting time to death in patients from simulating dataset based on the MIMIC-III dataset.

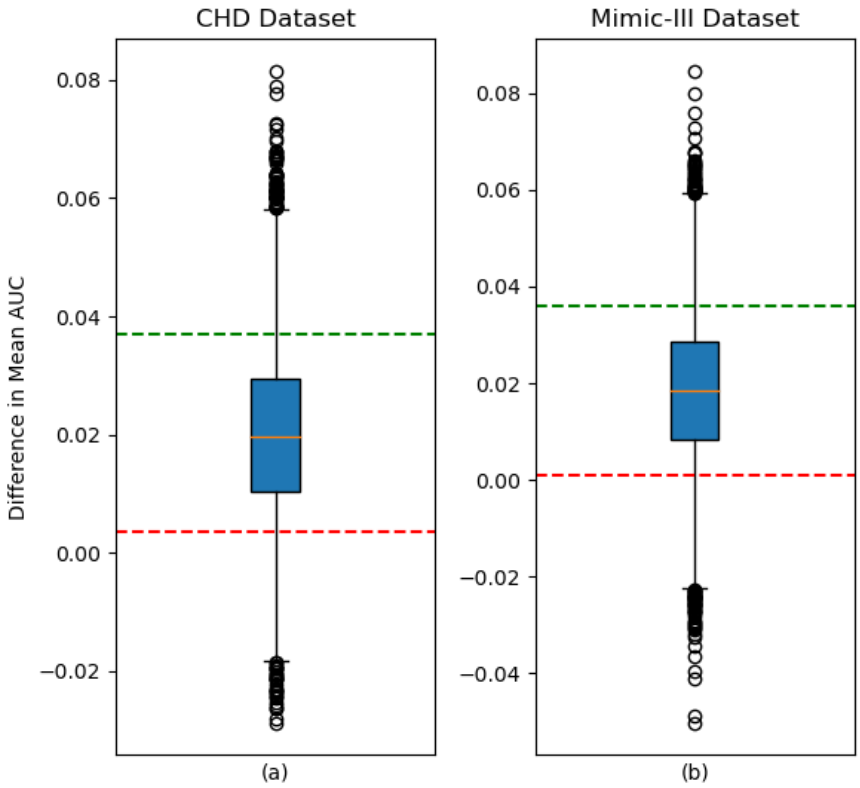


Figure S3: Comparison of mean AUC differences for mortality time prediction between MixEHR-SurG and MixEHR-G+Coxnet (ΔAUC = AUC(MixEHR-SurG) - AUC(MixEHR-G+Coxnet)), based on 10,000 bootstrap datasets for (a) CHD and (b) MIMIC-III dataset. The 75% confidence intervals are indicated by the dashed lines.

Figure S4: Mutual information between the top ICD codes from the top 6 survival phenotype topics identified from the CHD dataset. ICD codes in red are the ones that define the corresponding PheCode. The diagonal entries as well as mutual information between the same ICD codes were intentionally masked out for the ease of viewing.

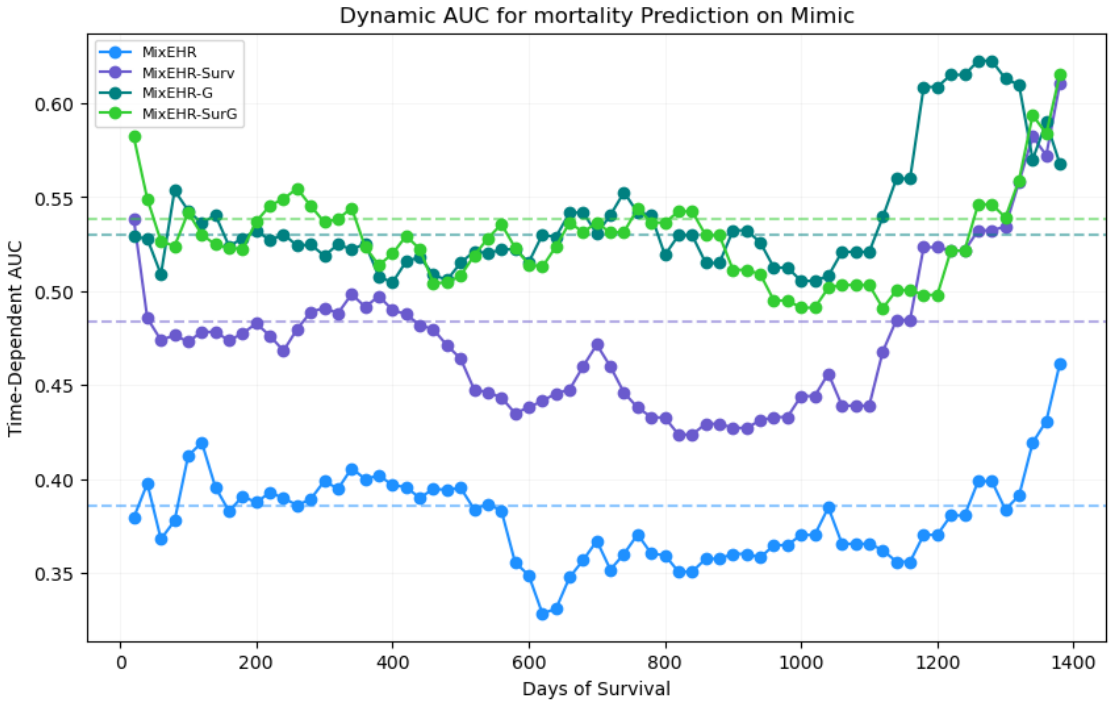


Figure S5: Dynamic AUC curves for predicting time to death in patients from the MIMIC-III dataset. We set a series of time points beginning at 20 and increasing in steps of 20, extending to 1400. At each of these intervals, we calculate the cumulative AUC, which is then used to construct the Dynamic AUC curve.

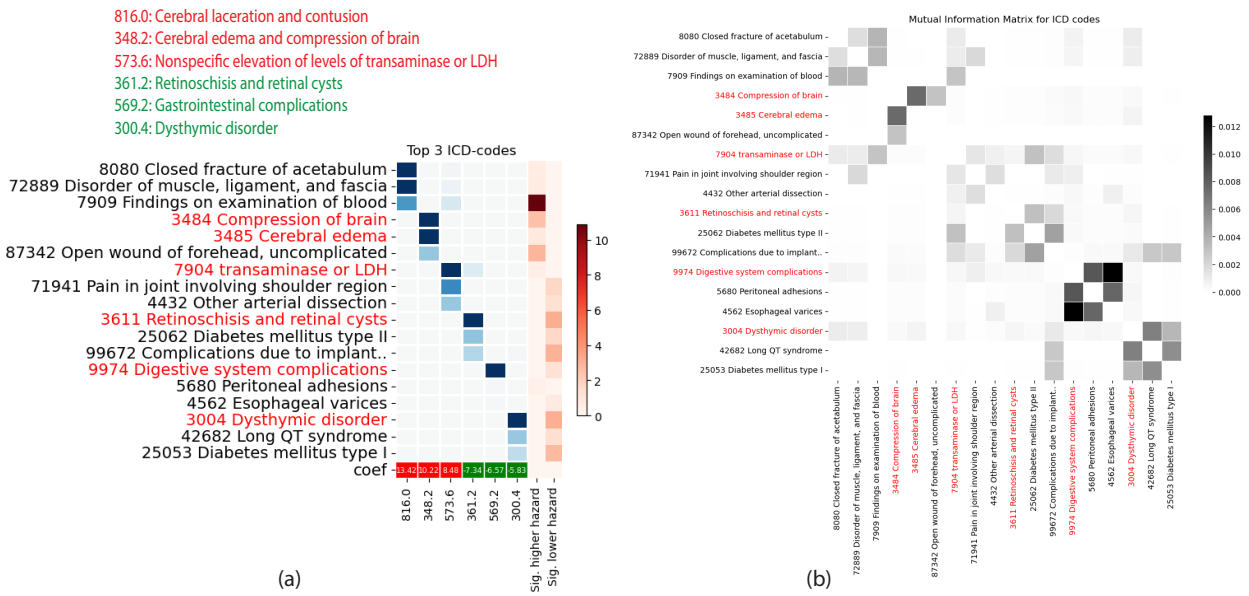


Figure S6: Comorbidity analysis of the top ICD codes for survival phenotype topics identified from the MIMIC-III data. (a) Heatmap displaying the top 3 ICD-9 codes per survival phenotype topics for the top 3 and bottom 3 phenotypes. The color gradation indicates the prevalence of each feature within each phenotype topic. The last row indicates the Cox regression coefficients. The last two columns display the color intensities proportional to the -log p-value from the log-rank test for high mortality risk and low mortality risk, respectively. (b) Mutual information between the top ICD codes from the top 6 survival phenotype topics. ICD codes in red are the ones that define the corresponding PheCode. The diagonal entries were intentionally masked out for the ease of viewing.

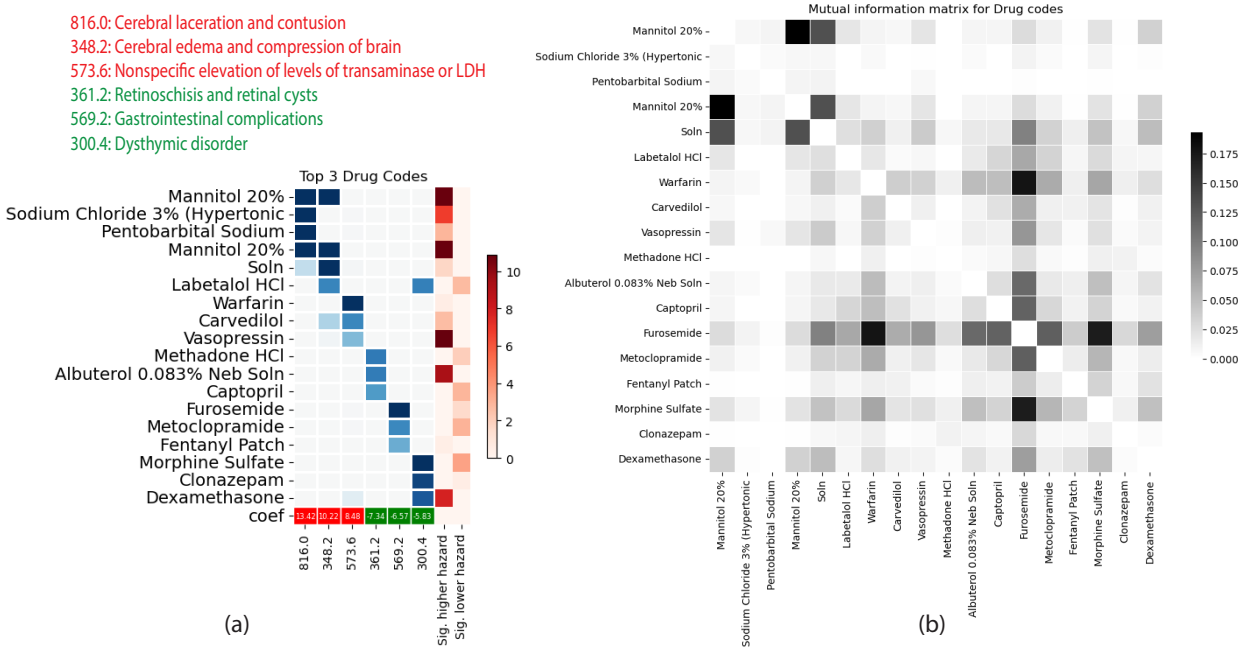


Figure S7: Comorbidity analysis of the top drug codes for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

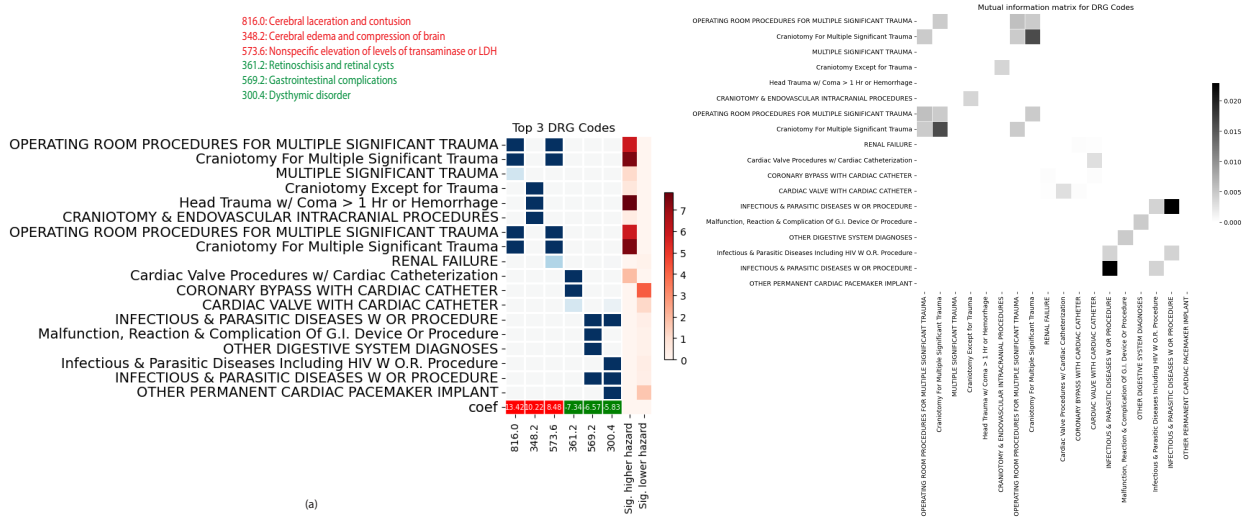


Figure S8: Comorbidity analysis of the top DRG codes for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

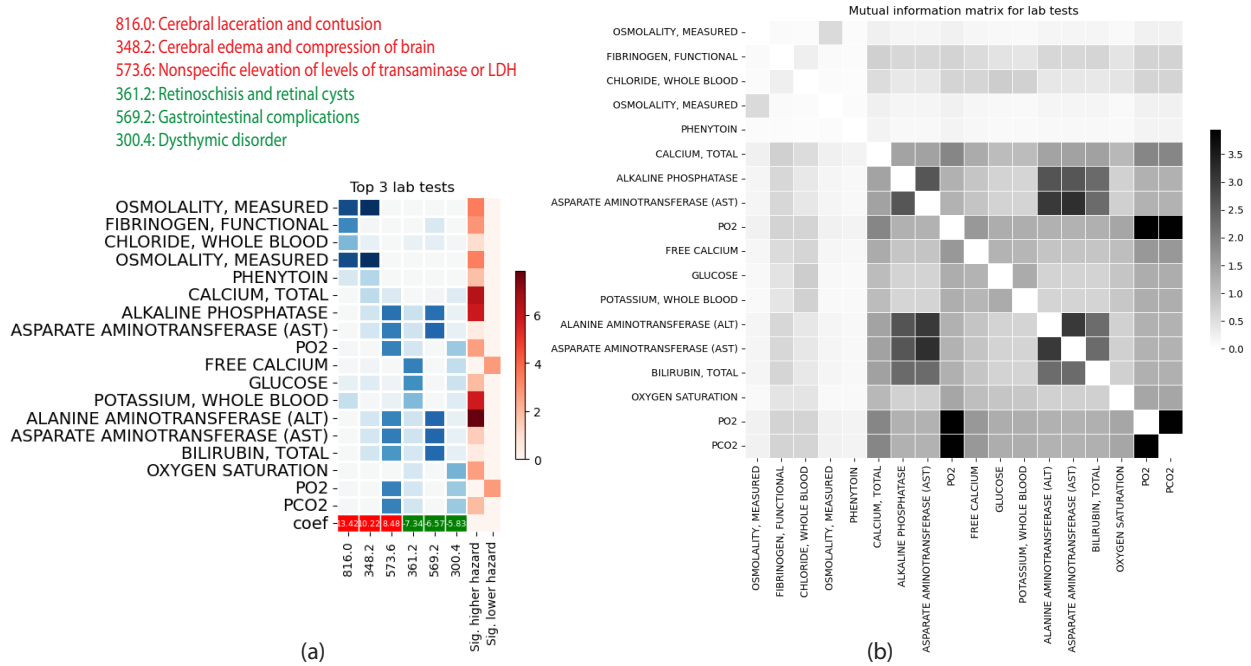


Figure S9: Comorbidity analysis of the top lab tests for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

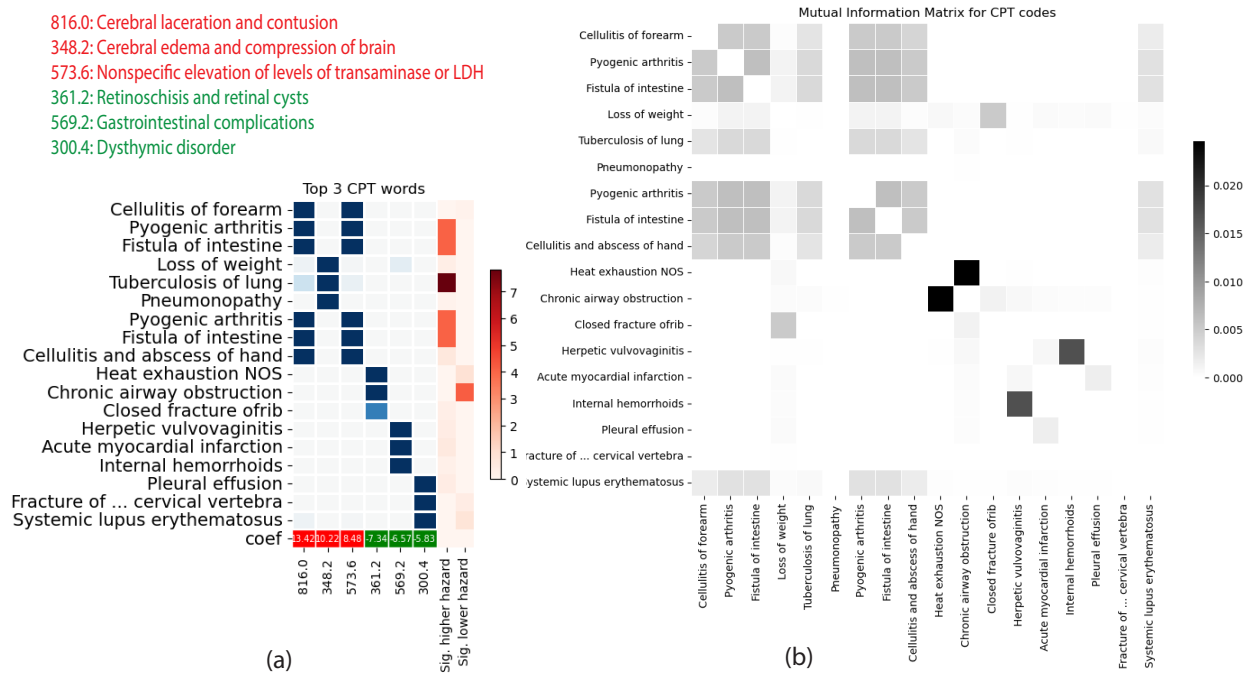


Figure S10: Comorbidity analysis of the top CPT words for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

References

[1] K. P. Liao, J. Sun, T. A. Cai, N. Link, C. Hong, J. Huang, J. E. Huffman, J. Gronsbell, Y. Zhang, Y.-L. Ho, et al., High-throughput multimodal automated phenotyping (map) with application to phewas, Journal of the American Medical Informatics Association 26 (11) (2019) 1255–1262.

[2] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (suppl_1) (2004) 5228–5235.

[3] Y. Teh, D. Newman, M. Welling, A collapsed variational bayesian inference algorithm for latent dirichlet allocation, Advances in neural information processing systems 19 (2006).

[4] I. Sato, H. Nakagawa, Rethinking collapsed variational bayes inference for lda, arXiv preprint arXiv:1206.6435 (2012).

[5] T. Minka, Estimating a dirichlet distribution (2000).

[6] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for cox's proportional hazards model via coordinate descent, Journal of statistical software 39 (5) (2011) 1.

[7] S. Pölsterl, scikit-survival: A library for time-to-event analysis built on top of scikit-learn, Journal of Machine Learning Research 21 (212) (2020) 1–6.
URL http://jmlr.org/papers/v21/20-729.html

[8] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society: Series B (Methodological) 34 (2) (1972) 187–202.

[9] T. M. Therneau, A Package for Survival Analysis in R, r package version 3.5-7 (2023).
URL https://CRAN.R-project.org/package=survival