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 $\theta_j \sim \text{Dir}(\boldsymbol{\alpha}), j = 1, \ldots, P$
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 $z_{ji}^{(m)} \sim \text{Mul}(\theta_j)$
   (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 $\theta_j \sim \text{Dir}(\boldsymbol{\alpha} \odot \boldsymbol{\pi}_j)$
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 $z_{ji}^{(m)} \sim \text{Mul}(\theta_j)$
   (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 $\theta_j \sim \text{Dir}(\boldsymbol{\alpha})$
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 $z_{ji}^{(m)} \sim \text{Mul}(\theta_j)$
   (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 $\theta_j \sim \text{Dir}(\boldsymbol{\alpha} \odot \boldsymbol{\pi}_j)$
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 $z_{ji}^{(m)} \sim \text{Mul}\left(\boldsymbol{\theta}_j\right)$

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

The equation (i) follows by defining

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

and

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

The equation (ii) follows by

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

and

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

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

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]:

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

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

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

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

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

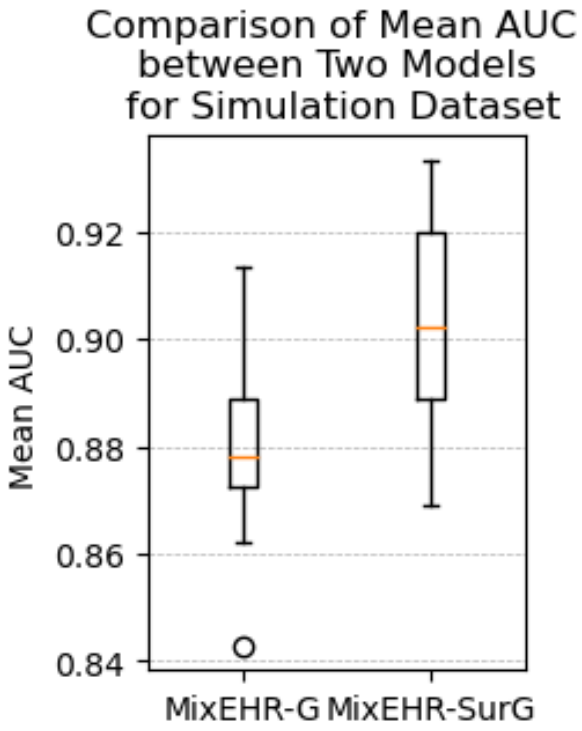

Figure S1: Comparison of the mean AUC between the pipeline MixEHR-G+Coxnet and MixEHR-SurG based on 10 simulated datasets.

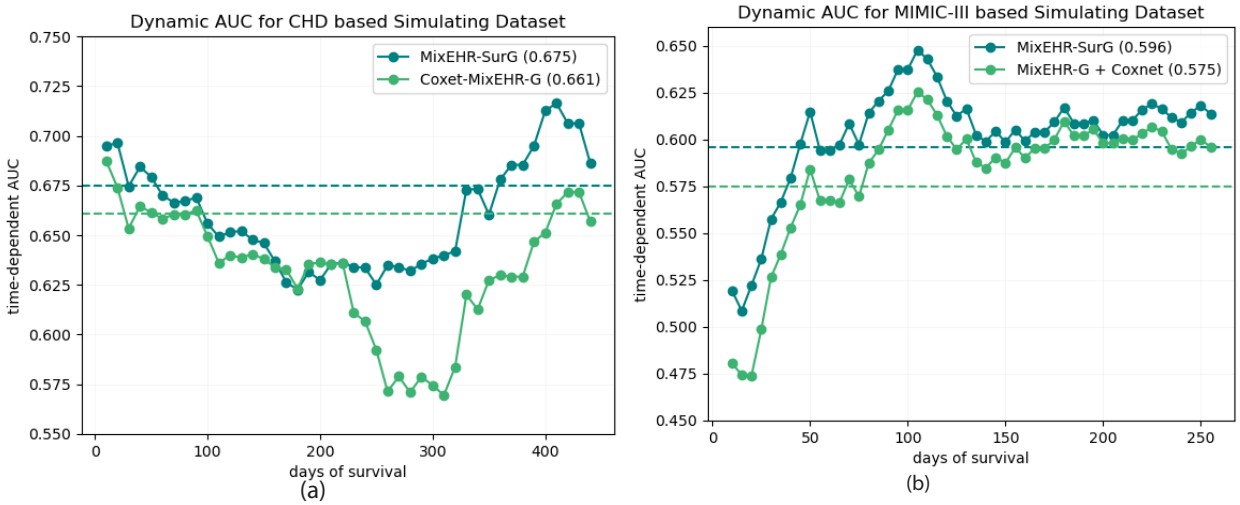

Figure S2: Dynamic AUC curves for predicting time to death in patients from the simulated data. (a) Dynamic AUC curves for predicting time to death in patients from simulating dataset based on the CHD dataset. (b) Dynamic AUC curves for predicting time to death in patients from simulating dataset based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

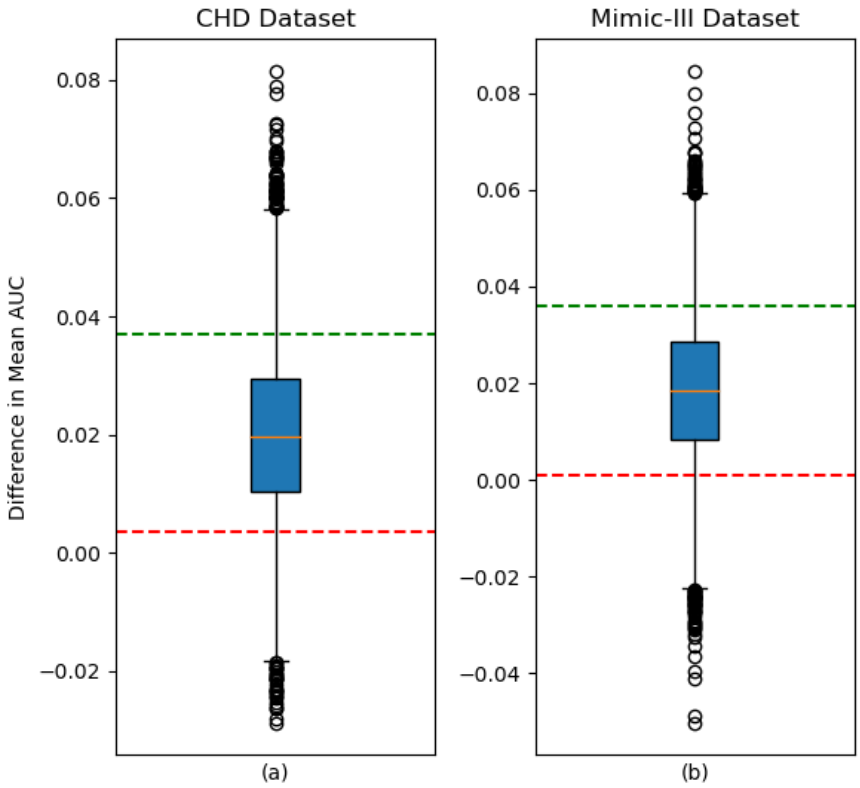

Figure S3: Comparison of mean AUC differences for mortality time prediction between MixEHR-SurG and MixEHR-G+Coxnet (ΔAUC = AUC(MixEHR-SurG) - AUC(MixEHR-G+Coxnet)), based on 10,000 bootstrap datasets for (a) CHD and (b) MIMIC-III dataset. The 75% confidence intervals are indicated by the dashed lines.

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.

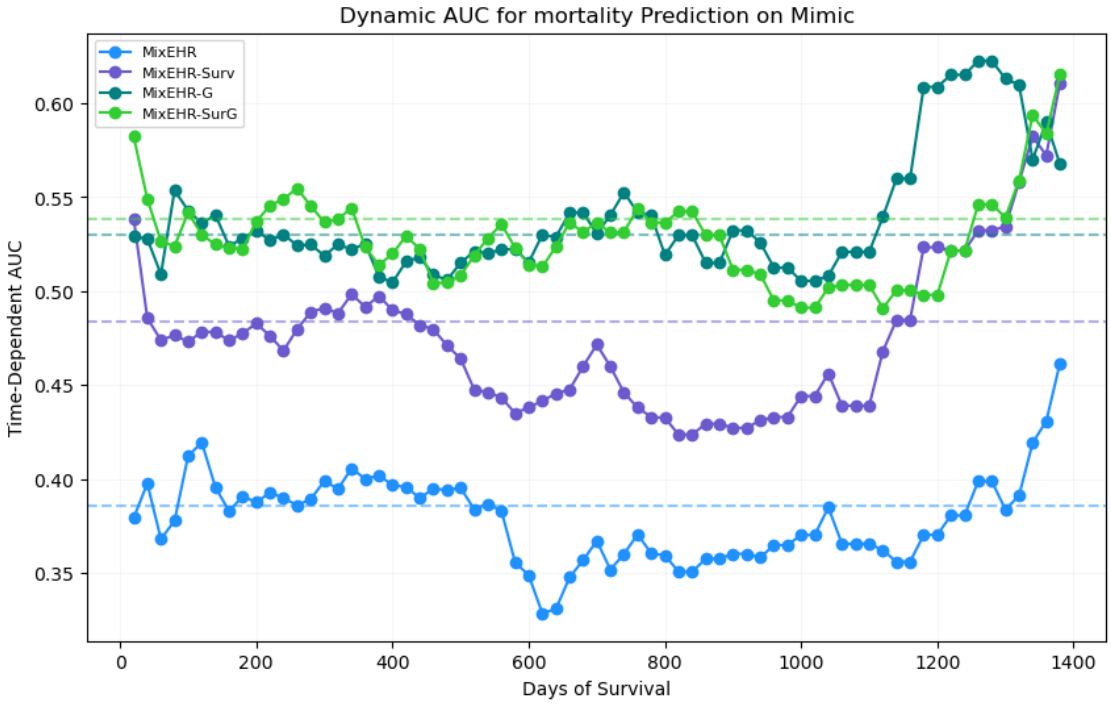

Figure S5: Dynamic AUC curves for predicting time to death in patients from the MIMIC-III dataset. We set a series of time points beginning at 20 and increasing in steps of 20, extending to 1400. At each of these intervals, we calculate the cumulative AUC, which is then used to construct the Dynamic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

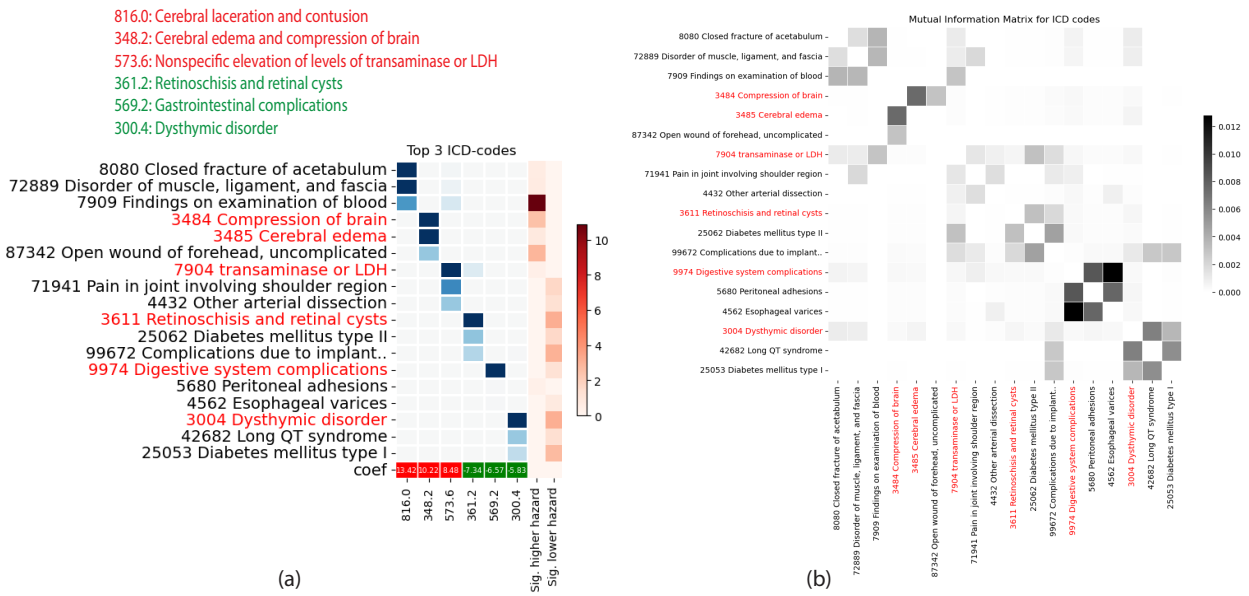

Figure S6: Comorbidity analysis of the top ICD codes for survival phenotype topics identified from the MIMIC-III data. (a) Heatmap displaying the top 3 ICD-9 codes per survival phenotype topics for the top 3 and bottom 3 phenotypes. The color gradation indicates the prevalence of each feature within 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.

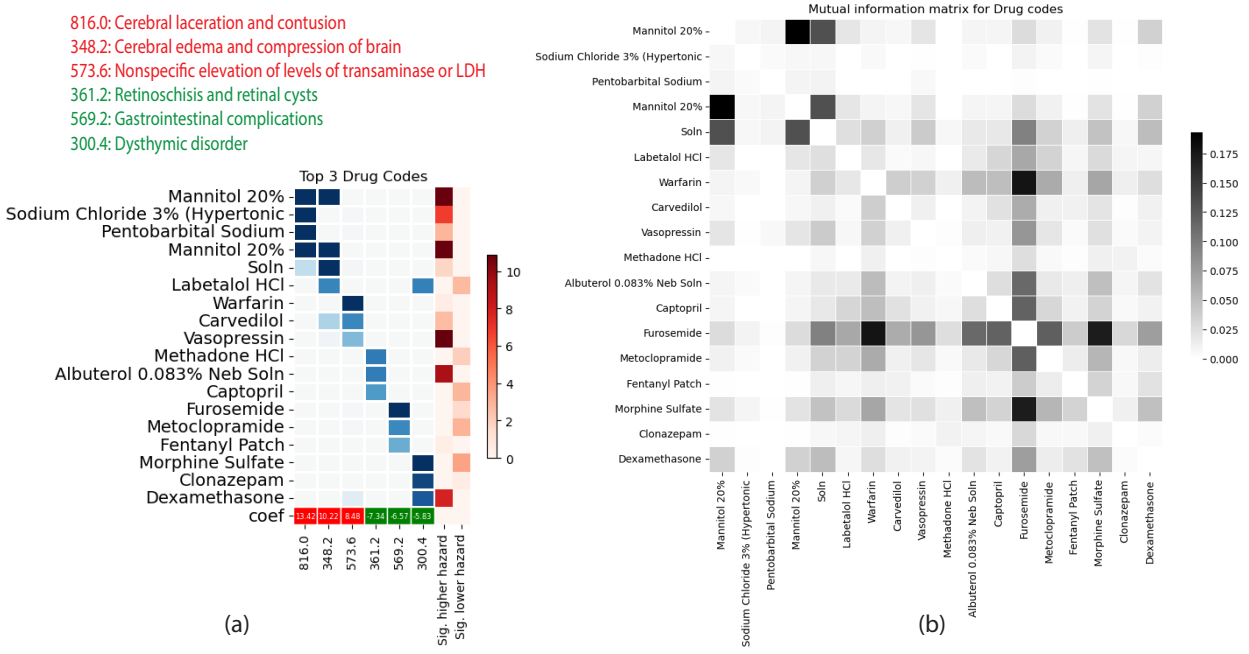

Figure S7: Comorbidity analysis of the top drug codes for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

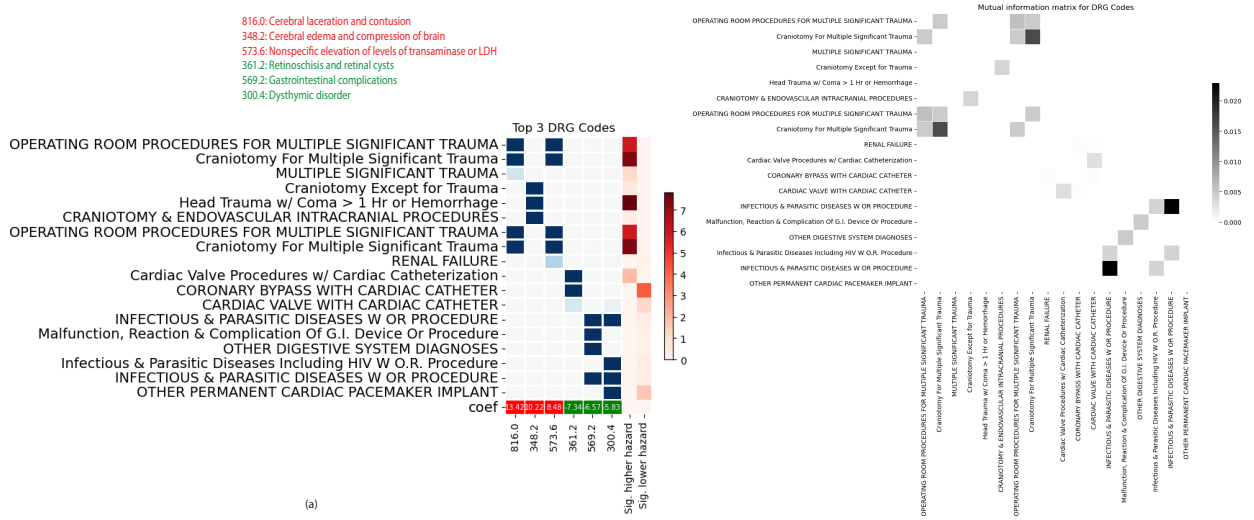

Figure S8: Comorbidity analysis of the top DRG codes for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

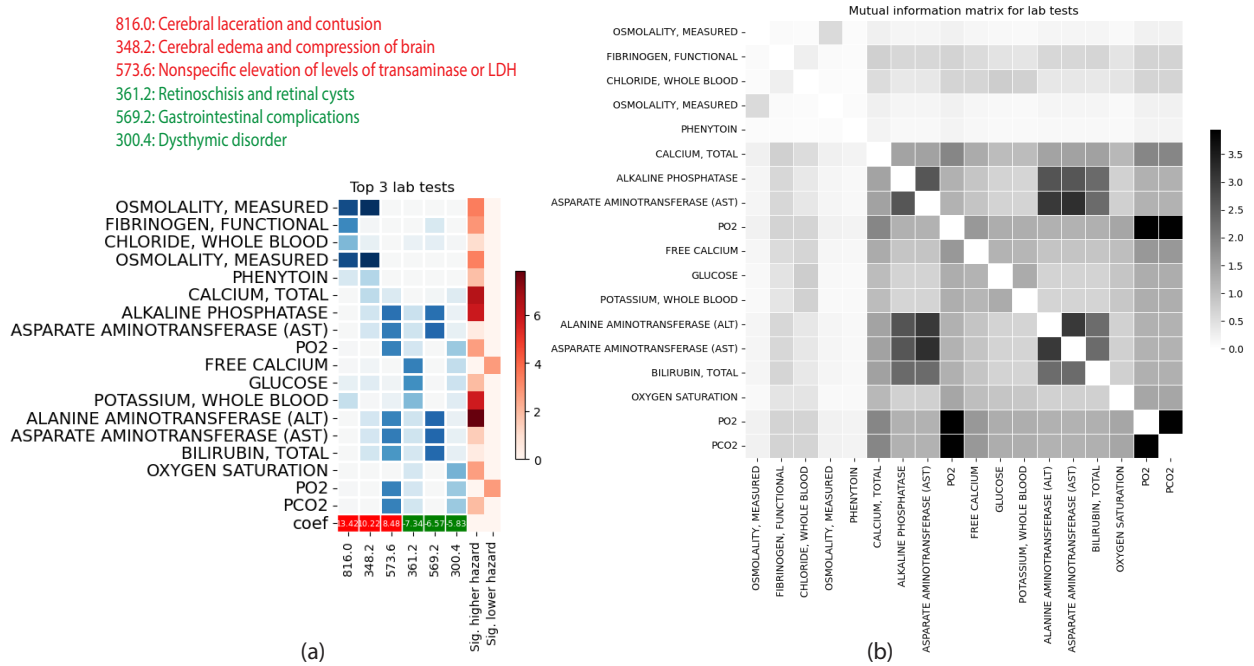

Figure S9: Comorbidity analysis of the top lab tests for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

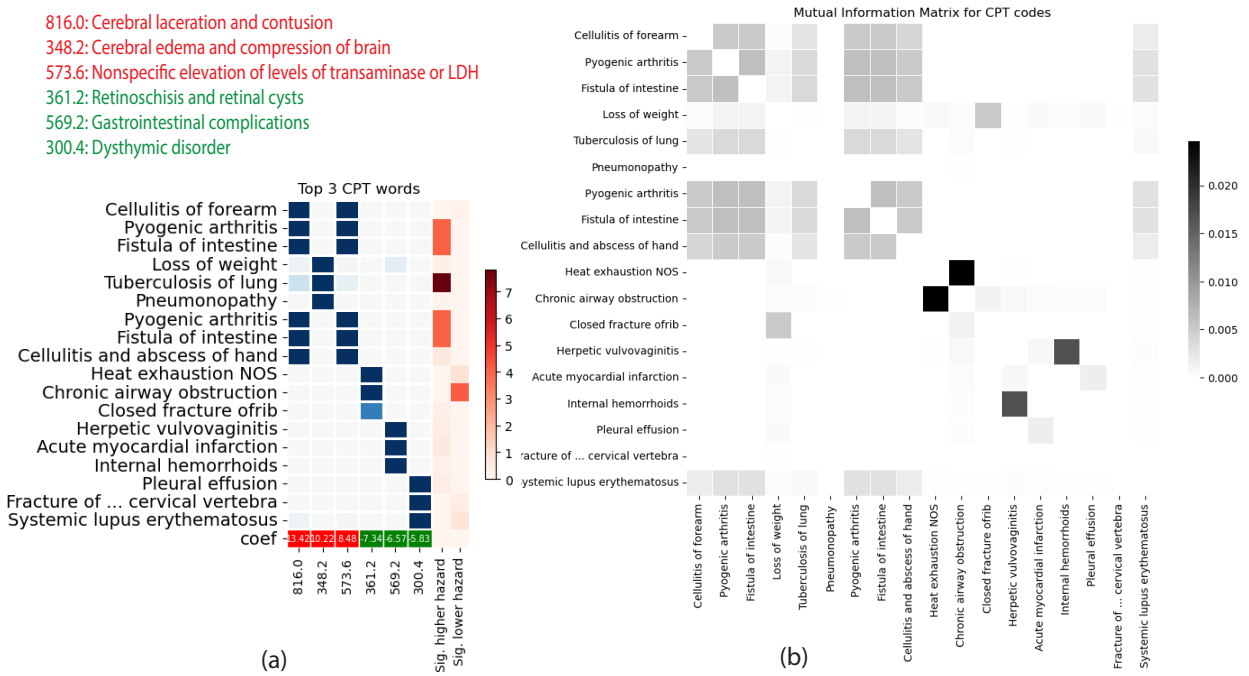

Figure S10: Comorbidity analysis of the top CPT words for survival phenotype topics identified from the MIMIC-III data. The presentation of the panels is the same as in **Supplementary Fig.** S6

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`