Supervised Multi-Specialist Topic Model With Applications on Large-Scale Electronic Health Record Data

Ziyang Song
School of Computer Science, McGill University
Montreal, QC, Canada

Xavier Sumba Toral
School of Computer Science, McGill University
Montreal, QC, Canada

Yixin Xu
School of Computer Science, McGill University
Montreal, QC, Canada

Aihua Liu
McGill Adult Unit for Congenital Heart Disease Excellence
Montreal, QC, Canada

Liming Guo
McGill Adult Unit for Congenital Heart Disease Excellence
Montreal, QC, Canada

Guido Powell
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

Aman Verma
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

David Buckeridge
School of Population and Global Health, McGill University
Montreal, Quebec, Canada
david.buckeridge@mcgill.ca

Ariane Marelli
McGill Adult Unit for Congenital Heart Disease Excellence
Montreal, QC, Canada
ariane.marelli@mcgill.ca

Yue Li
School of Computer Science, McGill
Centre for Bioinformatics, McGill University
Montreal, QC, Canada
yueli@cs.mcgill.ca

ABSTRACT

Motivation: Electronic health record (EHR) data provides a new venue to elucidate disease comorbidities and latent phenotypes for precision medicine. To fully exploit its potential, a realistic data generative process of the EHR data needs to be modelled.

Materials and Methods: We present MixEHR-S to jointly infer specialist-disease topics from the EHR data. As the key contribution, we model the specialist assignments and ICD-coded diagnoses as the latent topics based on patient's underlying disease topic mixture in a novel unified supervised hierarchical Bayesian topic model. For efficient inference, we developed a closed-form collapsed variational inference algorithm to learn the model distributions of MixEHR-S.

Results: We applied MixEHR-S to two independent large-scale EHR databases in Quebec with three targeted applications: (1) Congenital Heart Disease (CHD) diagnostic prediction among 154,775 patients; (2) Chronic obstructive pulmonary disease (COPD) diagnostic prediction among 73,791 patients; (3) future insulin treatment prediction among 78,712 patients diagnosed with diabetes as a mean to assess the disease exacerbation. In all three applications, MixEHR-S conferred clinically meaningful latent topics among the most predictive latent topics and achieved superior target prediction accuracy compared to the existing methods, providing opportunities for prioritizing high-risk patients for healthcare services.

Availability and implementation: MixEHR-S source code and scripts of the experiments are freely available at https://github.com/li-lab-mcgill/mixehrS

BCB '21, August 1–4, 2021, Gainesville, FL, USA

ACM ISBN 978-1-4503-8450-6/21/08.
https://doi.org/10.1145/3459930.3469543

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Topic model, Variational Bayesian, text mining, disease prediction, drug recommendation

ACM Reference Format:
Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. 2021. Supervised Multi-Specialist Topic Model With Applications on Large-Scale Electronic Health Record Data. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21), August 1–4, 2021, Gainesville, FL, USA.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3459930.3469543

1 INTRODUCTION

With the rapid adoption of electronic health record (EHR), there is an unprecedented opportunity to re-define medical concepts and automate disease diagnosis process. EHR include standardized

digital codes such as the International Classification of Diseases (ICD) 9 codes, which often span over tens of thousands of features. Traditional statistical methods are incapable of handling the high-dimensional EHR features and therefore often require engineering a small set of hand-crafted features. Topic models, on the other hand, show great promise in representing the entire discrete EHR data by a set of latent topics [4, 5, 14, 22]. In analogy to text categorization [4], we consider the EHR history for each patient as a document, which exhibits a mixture of memberships over a set of latent disease topics. However, existing EHR topic modeling ignores the generative process of the clinical specialists in diagnosing the patient. Also, most of EHR topic models are unsupervised and therefore require a downstream supervised classifier to perform prediction of a target disease.

In this paper, we present MixEHR-S as a novel unified supervised multi-specialist Bayesian topic model. MixEHR-S stands out from the existing methods with three key contributions. First, we explicitly model the distribution of specialists based on the patient's latent disease topic mixture. Second, we infer the specialist-specific latent disease topics, which capture the different clinical domain knowledge. Third, to predict a binary target label such as a disease diagnosis, we developed a Bayesian probit regression component to form a supervised topic model. This allows us to learn a linear classifier to predict a binary label using the topic mixture inferred for each patient. The posterior inference of the latent disease topics for each patient's diagnosis in turn takes into account the predictive likelihood of the target label. Therefore, the topic inference step and the supervised learning step can benefit from each other during the model training.

Using real-world large-scale EHR databases in Quebec, we demonstrated the utility of MixEHR-S model with three targeted applications: (1) Congenital Heart Disease (CHD) diagnostic prediction among 154,775 patients; (2) Chronic obstructive pulmonary disease (COPD) diagnostic prediction among 73,791 patients; (3) future insulin treatment prediction among 78,712 patients diagnosed with diabetes. In all three applications, we observe clinically meaningful latent topics among the most predictive latent topics and achieved superior target prediction accuracy compared to the existing methods, providing opportunities for prioritizing high-risk patients and drug recommendations.

2 RELATED METHODS

Our MixEHR-S model is related to the best-known topic model Latent Dirichlet Allocation (LDA) [4], which has been applied to raw clinical text for medical tasks in the past [5]. However, LDA is often inadequate to model complex diagnoses because it does not account for heterogeneous data categories. The mutli-view topic model UPhenome could learn diseases and patient characteristics with a fixed set of data types for heterogeneous medical records [22]. However, UPhenome is unable to infer specialist-diagnosis mechanisms as presented in our MixEHR-S.

Our Bayesian approach is also related to the frequentest non-negative matrix factorization (NMF) methods that were seen applications in the EHR domain. In particular, Joshi et al. (2016) described a NMF model to identify multiple co-occurring medical conditions using clinical notes [12]. Extending the single-view NMF framework, Gunasekar et al. (2016) proposed a collective NMF model called SiCNMF for modeling multi–source EHR Data [10]. Compared to the Bayesian topic modeling, these models are often less interpretable because they do not model the data generative processes.

In our recent work, we described a multi-modal topic model called MixEHR [14], which models different EHR data types with distinct categorical distributions. MixEHR is an unsupervised topic model, which needs training an additional classifier to predict a specific target label. One solution to this problem is to use supervised latent Dirichlet allocation (sLDA) [17], which infers latent topic distributions over documents (e.g., patients) while training a linear regression model to predict a continuous response of the documents (whereas we predict binary outcome with a Probit link). Along this direction, Zhang et al. (2017) proposes survival topic model, a supervised topic model that models jointly patients' discharge summaries and a Cox hazard ratios on patient mortality with a limited scope [30].

Although related, compared to the aforementioned work, our approach departs further from the recently popularized neural networks including our own [16] on supervised classification of target clinical outcomes [6, 7, 15, 21, 24, 25]. In these models, prediction accuracy is prioritized over model interpretability. The latter has been a challenge with the distributed representation of the neural networks. It often requires more sophisticated and computationally expensive techniques such as knowledge graph embedding with the attention mechanisms as demonstrated in GRAMS [8] to improve interpretability and mitigate over-fitting. Additionally, autoencoders [20, 26] such as Deep Patients [20] were also applied to learn low-dimensional manifold of the EHR data in an unsupervised fashion. These models often require careful fine-tuning of a large number of network hyperparameters and have similar interpretation challenges as the supervised neural networks.

3 METHODS

3.1 MixEHR-S model generative process

We model the heterogeneous medical data using a generative topic model illustrated in Fig 1. For the notation below, we use boldface to denote the vectors, capital letters for constants, and regular case for scalar variables. For each patient $j \in (1, \ldots, D)$, we index his/her ICD-9 code by $i \in \{1, \ldots, M_j\}$ for M_j total number of ICD diagnosis codes. Each ICD-9 code x_{ij} is assigned by one of the specialists b_{ij} with practitioner type $t \in \{1, \ldots, T\}$. Each patient is also associated with a binary target label y_j.

We assume that each patient j follows a disease topic mixture $\boldsymbol{\theta}_j$, which is a K-dimensional Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ with unknown hyperparameter $\boldsymbol{\alpha}$. To generate an ICD-9 code for a patient, we first sample a latent topic $z_{ij} = k$ from categorical distribution with rate set to $\boldsymbol{\theta}_j$. For vectorized notations, we represent the discrete topic assignment $z_{ij} = k$ by a binary one-hot vector $\mathbf{z}_{ij}$ such that $z_{ijk} = 1$ if $z_{ij} = k$ and $z_{ijk'} = 0 \forall k' \neq k$.

We then sample a specialist $b_{ij} = t$ from a topic distribution $\boldsymbol{\beta}_k \sim \text{Dir}(\iota)$, which is a T-dimensional Dirichlet variable with hyperparameter ι for T specialists. Given the topic assignment k and the specialist assignment t, we then sample the ICD-9 code from

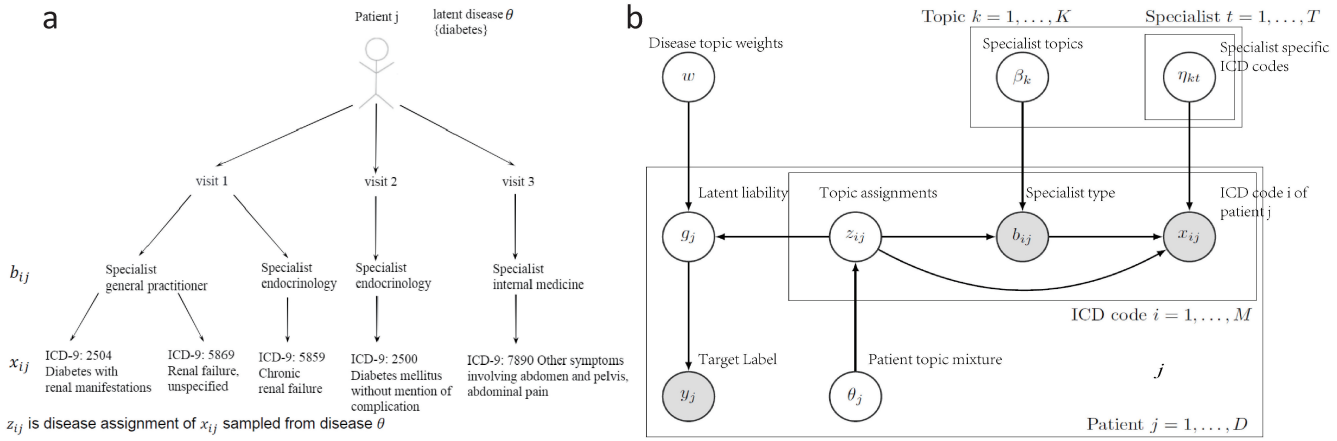

Figure 1: MixEHR-S model overview. (a) Conceptual illustration of the multi-specialist EHR data generative process. A diabetes patient j made three outpatient visits to receive a total of five ICD-9 diagnosis codes $i \in \{1, \ldots, 5\}$. Each code has its own underlying disease cause or topic z_{ij}, which was sampled from the patient disease mixture θ_j. Based on the disease cause, a specialist b_{ij} was assigned to the patient. The actual diagnosis code x_{ij} was made by the specialist based on his expertise. (b) The plate model for the multi-specialist EHR model. The probabilistic graphical model (PGM) formally characterizes the data generative process. The observed variables are the shaded nodes, including the target label y_j, specialist type b_{ij} and diagnosis code x_{ij}. The unshaded nodes are the latent variables. Hyperparameters are omitted for the purpose of clarity. The details of the variables in MixEHR-S are described in the main text and summarized in Table S1.

a categorical distribution with the rate $\boldsymbol{\eta}_{kt}$ that follows a set of V-dimensional Dirichlet distribution $Dir(\boldsymbol{\zeta}_k)$. The Dirichlet hyperparameters $\boldsymbol{\alpha}$, $\boldsymbol{\iota}$ and $\boldsymbol{\zeta}$ are given Gamma priors with fixed parameters.

To sample the target response label y_j for patient j, we first sample a Gaussian liability variable $g_j \sim \mathcal{N}(\mathbf{w}^\top \bar{\mathbf{z}}_j, 1)$, where $\mathbf{w}$ denotes the global regression coefficient and $\bar{\mathbf{z}}_j$ denotes the K-dimensional topic assignment average over the M_j codes. We then set the target response label y_j to 1 if g_j is positive otherwise $y_j = 0$. We posit a conjugate Gaussian prior over coefficient w with zero mean and uninformative constant variance τ. This is the Probit regression component of MixEHR-S model, which can be viewed as a supervised topic model.

Formally, our data generative process starts by generating the global topic distributions over the specialists and ICD-9 codes per specialist for each topic as well as the topic-specific coefficients for the response variable, respectively:

$$\boldsymbol{\beta}_k \sim \text{Dir}(\iota),\ \boldsymbol{\eta}_{kt} \sim \text{Dir}(\zeta_k),\ \mathbf{w} \sim \mathcal{N}(0, \tau^{-1}\mathbf{I}) \tag{1}$$

We then sample the local variables, namely the topic mixtures, topic assignments, specialist assignments, and ICD-9 codes for each patient j:

$$\boldsymbol{\theta}_j \sim \text{Dir}(\alpha), \quad z_{ij} \sim \text{Cat}(\boldsymbol{\theta}_j), \tag{2}$$

$$b_{ij} \sim \text{Cat}(\boldsymbol{\beta}_{z_{ij}}), \quad x_{ij} \sim \text{Cat}(\boldsymbol{\eta}_{z_{ij} b_{ij}}) \tag{3}$$

The hyperparameters are Gamma-distributed:

$$\alpha \sim \text{Gamma}(c_\alpha, d_\alpha),\ \iota \sim \text{Gamma}(c_\iota, d_\iota),\ \zeta \sim \text{Gamma}(c_\zeta, d_\zeta) \tag{4}$$

The binary label variable is sampled from a Probit distribution:

$$g_j \sim \mathcal{N}(\mathbf{w}^\top \bar{\mathbf{z}}_j, 1), \quad y_j = \begin{cases} 1, & \text{if } g_j > 0 \\ 0, & \text{if } g_j \le 0 \end{cases} \tag{5}$$

where $\bar{\mathbf{z}}_j = \frac{1}{M_j} \sum_{i=1}^{M_j} \mathbf{z}_{ij}$ and $\mathbf{z}_{ij}$ is a K-dimensional binary one-hot vector with topic indexed by z_{ij} set to 1 and the rest set to zeros.

3.2 Model Inference

Treating the latent variables as missing data, the complete joint likelihood based on the proposed model (Fig 1) is $p(\mathbf{z}, \mathbf{b}, \mathbf{x}, \mathbf{y}, \mathbf{g}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta})$. Exploiting conjugate property of the Dirichlet variables $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}$ to the respective categorical variables $\mathbf{z}$, $\mathbf{b}$, and $\mathbf{x}$, we first analytically integrated out the Dirichlet variables given the hyperparameters, resulting in the following marginal joint likelihood:

$$\begin{aligned} & p(\mathbf{z}, \mathbf{b}, \mathbf{x}, \mathbf{y}, \mathbf{g}, \mathbf{w} \mid \alpha, \iota, \zeta, \tau) \\ =& \int p(\mathbf{z}, \mathbf{b}, \mathbf{x}, \mathbf{y}, \mathbf{g}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta} \mid \alpha, \iota, \zeta, \tau) d\boldsymbol{\theta} d\boldsymbol{\beta} d\boldsymbol{\eta} \\ =& p(\mathbf{z}, \mathbf{b}, \mathbf{x} \mid d\alpha, \iota, \zeta) p(\mathbf{g} | \mathbf{z}, \mathbf{w}) p(\mathbf{y} | \mathbf{g}) p(\mathbf{w} | \tau) \end{aligned} \tag{6}$$

The full derivation is described in **Appendix** B.1. To approximate the sufficient statistics that are needed for inferring the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{b}, \mathbf{x}, \mathbf{g})$, we use variational inference [3]. In particular, we maximize the Evidence Lower Bound (ELBO):

$$\begin{aligned} \mathcal{L}(\Theta) = & \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log p(\mathbf{z}, \mathbf{g}, \mathbf{w}, \mathbf{b}, \mathbf{x}, \mathbf{y})] \\ & - \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log q(\mathbf{z}, \mathbf{g}, \mathbf{w})] \end{aligned} \tag{7}$$

Under the mean-field factorization, the proposed distribution of the latent variables have the same distributions as their priors:

$$q(\mathbf{z}, \mathbf{g}, \mathbf{w} \mid \mathbf{m}, \mathbf{S}, \boldsymbol{\gamma}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) \prod_{ijk} \gamma_{ijk}^{[z_{ij}=k]} \prod_j q(g_j \mid \lambda_j, 1) \tag{8}$$

where $\bar{\boldsymbol{\gamma}}_j = \frac{1}{M_j} \sum_i \boldsymbol{\gamma}_{ij}$ and $\boldsymbol{\gamma}_{ij}$ is a K-dimensional vector for the K topics. Maximizing ELBO with respect to the variational parameters

is equivalent to calculating the expectations [2]: $\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{g})}[\ln q(\mathbf{z}_{ij}|\boldsymbol{\gamma})]$, $\mathbb{E}_{q(\mathbf{z},\mathbf{g})}[\ln q(\mathbf{w}|\mathbf{m},\mathbf{S})]$, and $\mathbb{E}_{q(\mathbf{z},\mathbf{w})}[\ln q(\mathbf{g}|\mathbf{m},\mathbf{z})]$. Here $\mathbf{z}^{-(i,j)}$ denotes all of the latent variables except for variable $\mathbf{z}_{ij}$.

For the Bayesian regression component of the MixEHR-S model, we posit a truncated Gaussian distribution $\mathcal{TN}$ with mean $\boldsymbol{\lambda}$ and fixed variance for the liability variable $\mathbf{g}$:

$$q(g_j \mid \lambda_j) = \begin{cases} \mathcal{TN}_+(\lambda_j, 1), & \text{if } y_j = 1. \\ \mathcal{TN}_-(\lambda_j, 1), & \text{if } y_j = 0. \end{cases} \tag{9}$$

where

$$\lambda_j = \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\top]\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] \tag{10}$$

The variational distribution of the regression coefficients $\mathbf{w}$ follows a multivariate Gaussian distribution:

$$q(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) = \mathcal{N}(\mathbf{m}, \mathbf{S}) \tag{11}$$

where the variational parameters for $q(\mathbf{w} \mid \mathbf{m}, \mathbf{S})$ can be solved by completing the square (**Appendix** B.1):

$$\mathbf{S} = (\tau\mathbf{I} + \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}^\top\bar{\mathbf{z}}])^{-1}, \mathbf{m} = \mathbf{S}\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}]\mathbb{E}_{q(\mathbf{g})}[\mathbf{g}] \tag{12}$$

Importantly, the variational mean-field closed-form update for the posterior topic assignment $z_{ij} = k$ depends on both the categorical likelihood of the ICD-9 code and the predictive likelihood of the target response label: $p(\mathbf{z} \mid \mathbf{y}, \mathbf{x}) \propto p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$. Leveraging the conditional independence of these two likelihoods, we can calculate the closed-form variational inference update for topic k of ICD-9 code i in patient j:

$$\begin{aligned} \gamma_{ijk} \propto\ & (\alpha_k + E_{q(z^{-(i,j)})}[n_{jk}^{-(i,j)}]) \\ & \frac{\iota_{b_{ij}} + E_{q(z^{-(i,j)})}[m_{kb_{ij}}^{-(i,j)}]}{E_{q(z^{-(i,j)})}[m_{k.}^{-(i,j)}] + \sum_t \iota_t} \\ & \frac{\zeta_{kx_{ij}} + E_{q(z^{-(i,j)})}[p_{kb_{ij}x_{ij}}^{-(i,j)}]}{E_{q(z^{-(i,j)})}[p_{kb_{ij}.}^{-(i,j)}] + \sum_w \zeta_{kw}} \\ & \exp\left\{\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j} - \frac{1}{2M_j^2}\left[2\left(m_k\mathbf{m}^\intercal\boldsymbol{\gamma}_{j/i} + \mathbf{S}_k\boldsymbol{\gamma}_{j/i}\right) + m_k^2 + S_{kk}\right]\right\} \end{aligned} \tag{13}$$

where $\boldsymbol{\gamma}_{j/i} = \sum_{m\neq i}^{M_j}\boldsymbol{\gamma}_{ij}$ indicates the sum of all terms except for the i^{th} ICD-9 code, and $\mathbf{S}_k$ is the kth row of the covariance matrix $\mathbf{S}$. All of the variational expectations have closed-form expression conditioned on the other latent variables:

$$\begin{aligned} E_{q(\mathbf{z}^{-(i,j)})}[n_{jk}^{-(i,j)}] &= \sum_{i'\neq i}^{M_j}\gamma_{i'jk} \\ E_{q(\mathbf{z}^{-(i,j)})}[m_{b_{ij}k}^{-(i,j)}] &= \sum_{j'\neq j}^{D}\sum_{i}^{M_{j'}}[b_{ij'} = b_{ij}]\gamma_{ij'k} \\ E_{q(\mathbf{z}^{-(i,j)})}[p_{b_{ij}x_{ij}k}^{-(i,j)}] &= \sum_{j'\neq j}^{D}\sum_{i}^{M_{j'}}[b_{ij'} = b_{ij}, x_{ij'} = x_{ij}]\gamma_{ij'k} \end{aligned} \tag{14}$$

The above inference technique was originally developed only for LDA and named as collapsed variational Bayesian with zero-order Taylor expansion (CVB0) [1, 27]. Here we extended the inference to more general supervised multi-specialist topic models.

The predictive distribution of the target label y is a Bernoulli distribution when using a Gaussian response since the natural parameter $\mathbf{w}^\intercal\bar{\mathbf{z}}$ is identical to the mean parameter:

$$p(\mathbf{y}_\star \mid \bar{\mathbf{z}}_\star, \mathbf{y}, \bar{\mathbf{z}}) = \text{Bernoulli}\left(\mathbf{y}_\star \mid \Phi\left(\frac{\mathbf{m}^\intercal\bar{\mathbf{z}}_\star}{(1 + \bar{\mathbf{z}}_\star^\intercal\mathbf{S}\bar{\mathbf{z}}_\star)^{\frac{1}{2}}}\right)\right) \tag{15}$$

where $\Phi_j = \Phi(-\lambda_j)$ is the cumulative distribution function (CDF) of the standard normal distribution, and $\bar{\mathbf{z}}_\star$ and $\mathbf{y}_\star$ are the average topic counts and the target response label, respectively.

The overall inference algorithm is summarized as follows:

(1) Infer topic assignments $\mathbf{z}_{ij}$ using Eq. (13)
(2) Update sufficient statistics for the topic inference by Eq. (14)
(3) Update target prediction parameters by Eq (10) and Eq (12)
(4) Calculate the expectation of the target y (**Appendix** B.5)
(5) Update the Dirichlet hyperparameters (**Appendix** B.4)
(6) Repeat step 1-5 until little change in ELBO (default: 1e-6).

For efficient inference over large-scale EHR data, we employed a stochastic collapsed variational inference (SCVB0) [11, 14]. The full details are described in **Appendix** (Section B).

4 DATA

4.1 Quebec CHD Database (154,775 patients)

In Quebec (Canada) where universal health care is provided, every resident is assigned a unique Medicare number and all health services rendered are systematically recorded until death. In this study, three EHR databases were merged by the unique patient medicare numbers: (1) the physician's services and claims database from l983 to 2010; (2) the hospital discharge summary database from l987 to 2010; (3) the vital status database from 1983 to 2010. As a result, we have collected the EHR data for 154,775 patients. Demographic information including age and sex were also included in the databases and no specific age or sex biases were observed. The study population included patients with at least one CHD-related ICD-9 diagnosis between 1983 and 2010, whose ICD diagnoses were made by one of the 48 CV or non-CV specialists (e.g., cardiologist, thoracic surgeon, cardiac surgeon, cardio-vascular and thoracic surgeon, etc). In total, there are 9373 unique ICD-9 codes. The gold-standard labels are the binary label of CHD diagnosis made by clinicians after careful manual auditing the patient records using a clinician-developed rule-based algorithm. Among the 154,775 patients, 84,498 patients were diagnosed as true CHD patients ($y_j = 1$) and the rest are non-CHD patients ($y_j = 0$). Therefore, if we were to only use the CHD-related ICD-9 code to predict CHD labels, we could have only achieved an accuracy of 54.5%.

4.2 PopHR database

The Population Health Record (PopHR) is a semantic web application for measuring and monitoring population health and health system performance [23]. The public health insurance provider in Quebec, Canada (Régie de l'assurance maladie du Québec, RAMQ) provided the data on health service use. PopHR's current version uses an open, dynamic cohort of approximately 1.4 million people,

created by capturing a 25% random sample in the census metropolitan area of Montreal between 1998 and 2014. Follow-up ended when people died or moved out of the region of Montreal. The administrative database includes outpatient diagnoses and procedures submitted through billing claims to RAMQ, and procedures and diagnoses from hospital records. Drug dispensation data are available for people who have drug insurance through RAMQ (which includes approximately half the population and all those over 65 years of age). All data are linked through an anonymized version of the RAMQ identification number. As a proof-of-concept, we focused on a subset of the patient cohort with two distinct target diseases as described below.

Chronic Obstructive Pulmonary Disease (COPD). Our goal in this application is to infer specialist-dependent COPD topics and predict the true COPD patients. To this end, we retrieved 73,791 patients with at least one COPD related ICD-9 code (e.g. 491x, 492x, 496x) from the physician billings table of the PopHR database. The gold-standard labels were assigned to patients as of their incident event (COPD diagnostic code for hospitalization or medical billing) occurring after a minimum of two years of time at risk [29]. Among the 73,791 patients identified through ICD-9 codes, 67,380 patients were confirmed to be the true COPD patients and the rest are non-COPD patients. In total, there are 6,625 unique ICD-9 codes and 43 unique specialists.

Insulin usage among diabetes patients. In this application, we aim to infer latent topics that are predictive of whether each diabetes patient started the insulin medication 6 months after their initial diagnosis. This is useful in forecasting the disease exacerbation. We extracted 78,712 diabetes patients with the ICD-9 codes 250x and continuous public drug insurance for hospitalization or medical billing. We set the first diabetes diagnosis of a patient as the start of insurance coverage. We selected patients with a continuous public drug insurance after the first diagnosis of diabetes-related code, with continuous insurance defined as: (1) patients with at least 6 months of uninterrupted insurance, or; (2) patients with interrupted insurance records for which interruptions are less than 2 months. To avoid misclassifying the first dispensation in patients with an unrecorded history of insulin, we removed patients who used insulin within 6 months after the first diagnosis of a diabetes-related code.

For the remaining patients, we only used their ICD codes observed accumulatively up to the first diagnosis to predict their future drug usage. We treated a patient as positive if he or she started to use insulin 6 months after being diagnosed with diabetes. Among the 78,712 diabetes patients, 11,433 patients were labeled as insulin users. In order to evaluate our model on a balanced dataset, we randomly sampled 11,433 negative patients to obtain a perfectly balanced dataset (50% positive patients, 50% negative patients). There are 5,477 unique ICD-9 codes and 43 specialists in the resulting balanced diabetes dataset.

5 EXPERIMENTS

We sought to evaluate both model interpretability and prediction accuracy of our MixEHR-S model. Because the true topics of real dataset are not known, we first used simulation to assess whether MixEHR-S can recapitulate the ground-truth topics. We used the generative process of the proposed model to obtain sample data. We simulated 2,500 patients, 750 ICD diagnosis codes and 48 specialists for the evaluation. We considered 25 topics to model toy data with a 80/20 train-test split. For the purpose of comparison, we used LDA as the baseline model [4]. We ran LDA implemented in Python package scikit-learn with default variational inference algorithm. We evaluated target response prediction by the receiver operating characteristic (ROC) curve and precision-recall curve (PRC). We directly predicted the binary target labels using MixEHR-S. For the unsupervised model LDA, we trained a separate Bayesian logistic regression to predict labels given the LDA-inferred patient topic mixtures. Overall, MixEHR-S achieved an excellent topic recovery and outperformed LDA (Fig S2a). Details were described in **Appendix** A.2.

We then evaluated our MixEHR-S model based on the real EHR data from Quebec CHD, PopHR-COPD, and PopHR-diabetes datasets (Section 4). As a baseline method, we evaluated MixEHR [14], which is an unsupervised multi-modal topic model. We also evaluated LDA and the supervised LDA (sLDA), both of which operate on flatten ICD-9 codes as a single category [4, 18]. We applied the Bayesian Logistic Regression classifier that uses the topic mixtures inferred by MixEHR or LDA to predict labels. We also compared our approach with supervised models including Least Absolute Shrinkage and Selection Operator (LASSO) [28], Random Forest (RF), Gradient Boosting (GB), and two-layer feedforward neural network (NN), which directly use ICD-9 codes as raw features to predict CHD labels. LDA, LASSO, RF, and GB were implemented by the Python package scikit-learn with the default optimization algorithms. NN with two fully connected layers and 100 hidden units per layer was implemented in PyTorch.

Additionally, we evaluated several state-of-the-art EHR-focused models including MixEHR [14], Deep Patient (DP) [20] (https://github.com/natoromano/deep-patient) , Graph-based Attention Model (GRAM) [9] (https://github.com/mp2893/gram), and Sparsity-inducing Collected Non-negative Matrix Factorization (SiCNMF) [10] (https://github.com/sgunasekar/SiCNMF) using available published codes from the corresponding GitHub repositories.

For all three applications, we split each dataset into 70% training, 10% validation, and 20% testing sets. For the topic models (i.e., MixEHR, LDA, sLDA, and our MixEHR-S), we used the validation set to choose the best number of topics based on the unsupervised perplexity (i.e., the negative held-out log-likelihood) on the 10% validation patients. We evaluated on the 20% test patients in predicting the target disease or outcome based on the area under the ROC curves (AUROC) and precision-recall curves (AUPRC). To obtain robust prediction estimates, we conducted 10 repeated runs with random 70%-train/10%-validation/20%-test splits and recorded the mean values and standard deviations of AUROC and AUPRC for each model on each test set in Table 1. The detailed model specifications and hyperparameters were described in Appendix A.3.

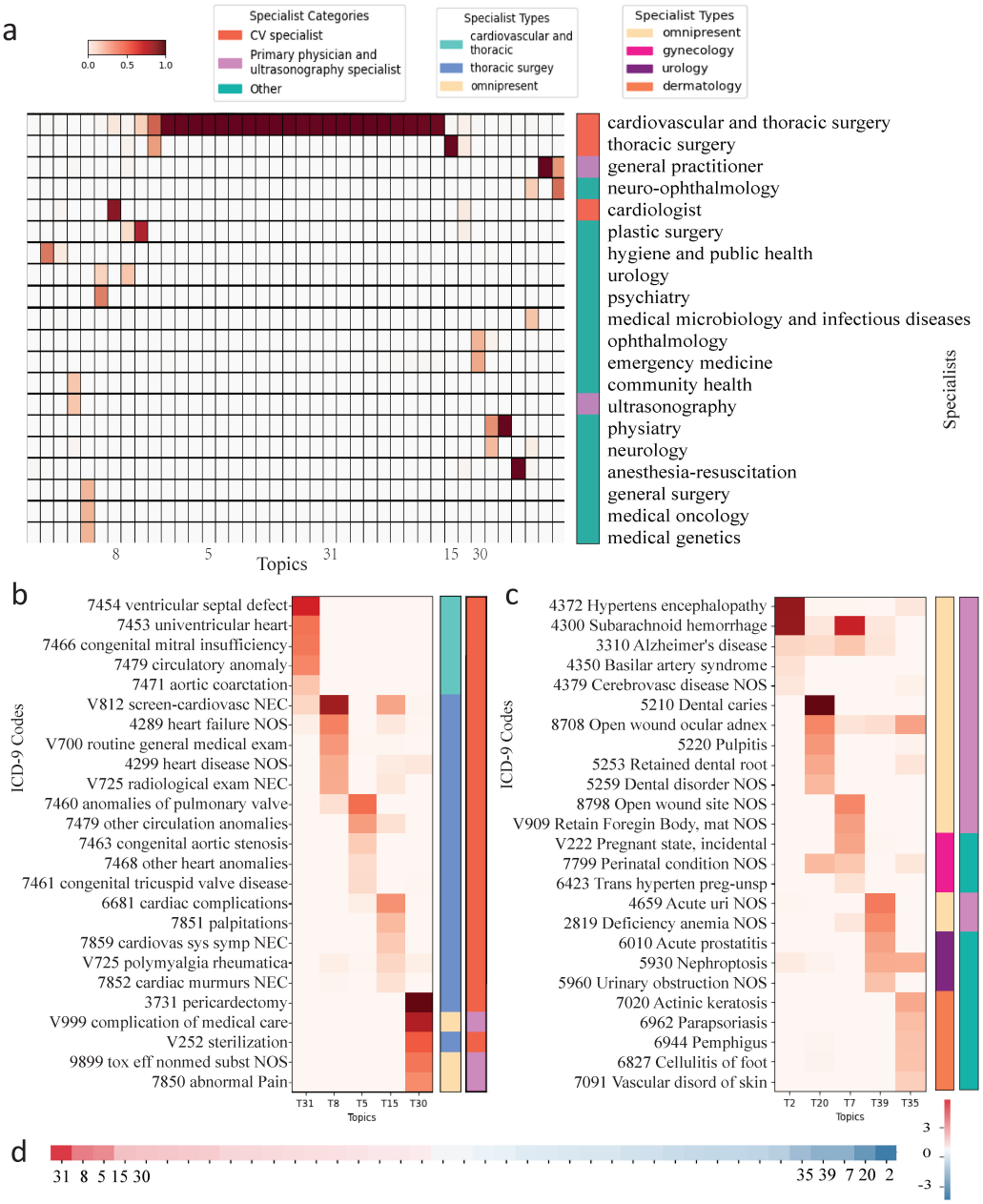

Figure 2: Disease topics inferred by MixEHR-S on the CHD dataset. (a) The inferred specialist topics. The color intensities are proportional to the inferred probabilities of specialists under each topic and the side bar indicates the specialist categories. The topics that are numbered are the most predictive topics of CHD. The full specialists are illustrated in Fig S3. (b) and (c) Top ICD-9 codes from the five most positively predictive topics and the five most negatively predictive topics. The side bars indicate the specialist types and categories. (d). Linear coefficients for the 40 topics in decreasing order.

6 RESULTS

6.1 Inferring specialist-dependent topics from Quebec CHD Dataset

We investigated the inferred 40-topic mixtures of 48 specialists that give the lowest perplexity on the validation set (Fig. 2a) (i.e., $\boldsymbol{\beta}$). As expected, we observe high probabilities for CV specialists, such as cardiovascular and thoracic surgery. Based on the learned topic coefficient $\mathbf{w}$, topics T31 and T8 are the two most predictive topics for CHD and are strongly associated with cardiovascular and thoracic surgery and cardiologist, respectively. For each topic, we chose the top five ICD-9 codes to reveal its meaning. We found a

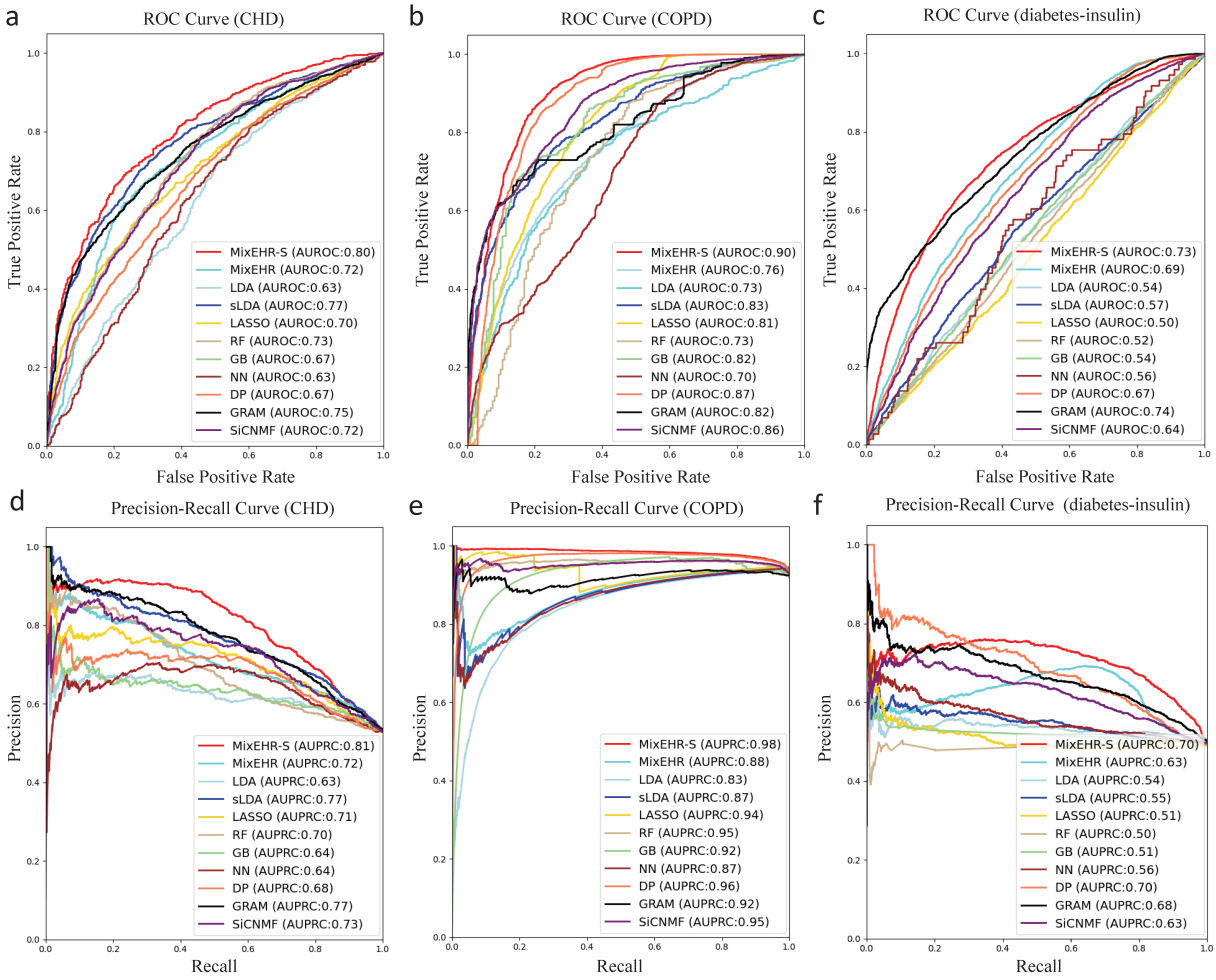

Figure 3: Prediction accuracy of responses comparing MixEHR-S model, the topic models (MixEHR, LDA, and sLDA), the models which directly apply on raw ICD-9 codes (LASSO, RF, GB, and NN), and the EHR-focused models (DP, GRAM, and SiCNMF). a-c (d-f) ROC and PR curves of all methods on CHD, COPD, and diabetes-insulin predictions with AUROC (AUPRC) of each method indicated as inset in each panel.

strong connection between the inferred CHD-positive topics and the disease pathology. In particular, topics T31, T8, and T5 contain many CHD-related diagnosis codes such as ventricular septal defect (7454), univentricular heart (7453), and anomalies of pulmonary valve (7460), acute conditions such as heart failure (4289) under T5, and procedure code such as cardiac examination under T5. Moreover, all of the top ICD diagnosis codes under topic T31 and T5 come from the CV specialists (i.e. cardiovascular and thoracic surgery or thoracic surgeon). Also, the top ICD-9 codes under topic T15 such as cardiac complications and cardiovascular symptoms also mostly make clinical sense. Topic T30, which is weakly associated with heart related surgery, gives less interpretable results since several of its top codes are diagnosed by the non-CV specialists.

Interestingly, the selected negatively predictive topics are also clinically coherent although all of them are not related to CHD as expected. For instance, we observed a clear enrichment for cerebrovascular diseases in T2. Topic T20 represents dental diseases. All of the top ICD-9 codes under the negatively predictive topics come from non-CV practitioners. Therefore, our inferred topics could be used as a clinically relevant departure point to uncover more specific therapeutic information.

We then examined the disease topic mixture memberships along the patients dimension. We selected 5 patients with the highest proportions for the top 5 most predictive topics (Fig S4b), who we considered as high-risk CHD patients. The majority of the top patients are indeed true CHD patients. In contrast, most patients under the top 5 negatively predictive topics do not have CHD labels (Fig S4c). We then examined the 5 highest-scoring ICD-9 diagnosis codes among these patients under the five most positively predictive topics and the five most negatively predictive topics. As expected, the top patients in the positively predictive topics were

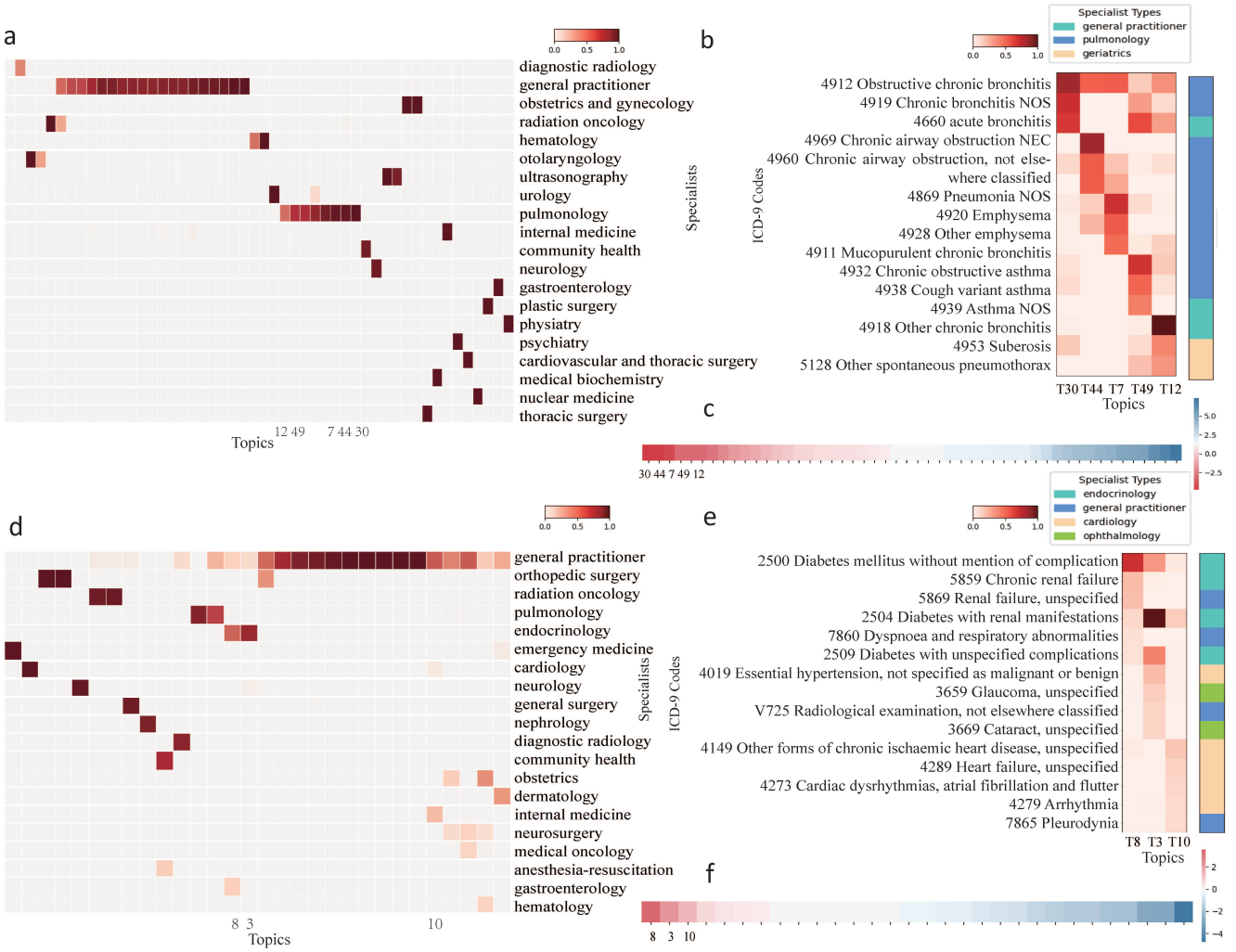

Figure 4: Disease topics inferred from the COPD and diabetes patients' EHR data extracted from the PopHR database. a. Specialist topics. Same as in Fig 2, the color intensities are proportional to the probabilities of specialists under each topic and the side bar indicates the specialist categories. b. Top ICD-9 codes from the five most positively predictive topics. The side bars indicate the specialist types. c. Linear coefficients for the 50 topics in decreasing order. d-f. specialist topics, top ICD-9 codes per top topic, and linear coefficients for the 6-month insulin-usage prediction among the diabetic patients.

mostly diagnosed with the CHD-related ICD-9 codes. In particular, most of the patients under topic T31 have highly CHD-specific ICD-9 codes 7454, 7453 and 7471, corresponding to the diagnoses of ventricular septal defect, univentricular heart, and aortic coarctation, respectively. Some of the top patients were not confirmed with CHD and may be deemed as the high-potential CHD patients. Some of the top 50 patients under topic T31 have neither the CHD label nor the top 5 CHD-related ICD codes. Nonetheless, these patients possess the other top CHD-related ICD-9 codes under the topic T31 (Fig S4d).

6.2 CHD target label prediction

Quantitatively, MixEHR-S conferred the highest AUROC (80.07%) and AUPRC (81.22%) among all methods (Fig 3a and d, Table 1). The LDA model that ignore distinct specialists performed a lot worse while the supervised sLDA model was a slightly better. MixEHR-S also outperformed the unsupervised variant MixEHR, which achieved 72.14% AUROC and 71.67% AUPRC. Additionally, the baseline discriminative models LASSO, RF, GB, and NN, which directly use the raw features, also did not predict CHD outcomes as accurately as our MixEHR-S. This is due to their inability to model the distribution of the EHR input data. As a result, they are more sensitive to the sparsity and noise intrinsic to the EHR data.

Compared to the above baseline models, the existing EHR-focused models demonstrated generally better prediction performance. DP conferred relatively low prediction accuracy possibly due to its need to operate on extensively prepossessed and filtered EHR codes. GRAM utilized the ICD-related knowledge graph to embed ICD

Table 1: Prediction performance of MixEHR-S and other methods on CHD, COPD, and diabete-insulin predictions. The mean values and standard deviations (in the brackets) of AUROC and AUPRC for each model each test set are computed on 10 randomly training and testing splits. For each application, the highest average AUROC and AUPRC among all methods are in bold.

Method	CHD		COPD		diabetes-insulin	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
MixEHR-S	**0.8007** (0.0263)	**0.8122** (0.0228)	**0.9036** (0.0290)	**0.9774** (0.0122)	**0.7341** (0.0469)	**0.7006** (0.0376)
MixEHR	0.7214 (0.0384)	0.7167 (0.0462)	0.7621 (0.0429)	0.8804 (0.0607)	0.6892 (0.0463)	0.6346 (0.0471)
LDA	0.6292 (0.0327)	0.6311 (0.0341)	0.7285 (0.0362)	0.8318 (0.0686)	0.5418 (0.0210)	0.5385 (0.0188)
sLDA	0.7684 (0.0258)	0.7708 (0.0272)	0.8317 (0.0362)	0.8665 (0.0471)	0.5622 (0.0124)	0.5509 (0.0146)
LASSO	0.6967 (0.0381)	0.7130 (0.0392)	0.8125 (0.0431)	0.9382 (0.0271)	0.5017 (0.0109)	0.5083 (0.0086)
RF	0.7291 (0.0131)	0.7012 (0.0155)	0.7312 (0.0181)	0.9453 (0.0093)	0.5238 (0.0113)	0.5044 (0.0081)
GB	0.6703 (0.0107)	0.6447 (0.0142)	0.8239 (0.0154)	0.9164 (0.0142)	0.5398 (0.0220)	0.5124 (0.0093)
NN	0.6334 (0.0439)	0.6409 (0.0504)	0.6976 (0.0360)	0.8708 (0.0557)	0.5588 (0.0591)	0.5620 (0.0475)
DP	0.6675 (0.0305)	0.6825 (0.0322)	0.8655 (0.0206)	0.9634 (0.0125)	0.6701 (0.0256)	**0.7028** (0.0296)
GRAM	0.7493 (0.0346)	0.7687 (0.0221)	0.8245 (0.0168)	0.9239 (0.0102)	**0.7387** (0.0438)	0.6815 (0.0399)
SiCNMF	0.7241 (0.0436)	0.7313 (0.0387)	0.8591 (0.0260)	0.9546 (0.0168)	0.6432 (0.0423)	0.6283 (0.0368)

code, which helped achieving more accurate prediction (74.93% AUROC and 76.87% AUPRC). Together, we attribute the highest performance of our MixEHR-S to its simultaneous inference of the specialist-specific topics and the predictive topic coefficients.

6.3 Modeling PopHR-COPD data

We then examined the 50-topic mixtures over the 43 specialists that we inferred from the PopHR-COPD data (Section 4.2). As in the CHD topic analysis above, we focused on the 5 most predictive topics of COPD by examining the specialist distribution (Fig 4a, c). Indeed, we observe that the top 5 topics (T30, T44, T7, T49, T12) are strongly associated with pulmonology specialist, which is closely related to the COPD diagnosis. Consistent to the trend in the CHD analysis, the disease relevance increases as the topic's predictive coefficient increases (Fig 4c; **Appendix** A.4). In particular, T30 is the most predictive one among all 50 topics, which exhibit the highest probability for pulmonology specialist. Topics T7 and T12 exhibit relatively weaker associations with pulmonology specialist and therefore weaker effect sizes for COPD.

The top 3 ICD-9 codes under the most predictive topic T30 are all closely related to COPD: Obstructive chronic bronchitis(4912), Chronic bronchitis NOS (4919), and acute bronchitis (4660) (Fig 4b). The top second and top third topics T44 and T7 also contain multiple COPD-related codes, which represent Chronic airway obstruction and emphysema, respectively. Overall, 10 out of the 15 top ICD-9 codes are from the pulmonology specialist type and the rest of the codes are from general practitioner and geriatrics. This makes sense since pulmonology specialist is more likely to make diagnosis for these COPD-related ICD codes. Interestingly, although topics T49 and T12 are less predictive of COPD, they are strongly associated with asthma and lung malfunction, which often manifest similar symptoms as COPD.

We then evaluated the prediction accuracy (Fig 3b and e, Table 1). MixEHR-S achieves the highest AUROC of 90.36% among all methods. The other top performing models namely DP, SiCNMF, and GRAM conferred AUROC 86.55%, 85.91%, and 82.45%, respectively. LDA alone did not work well with only 72.85% AUROC, possibly due to the its inability to capture the predictive topics of COPD and its negligence of the multi-specialist ICD-9 distributions.

Compared with LDA, the unsupervised MixEHR model [14] that learns specialist specific topics achieved higher AUROC of 76.21%. However, it still performed a lot worse on the prediction task compared to MixEHR-S. We observed much higher accuracy by the supervised LDA (sLDA) with 83.17% AUROC. Therefore, adding the supervised component into the model can improve predictive performance.

The discriminative models namely LASSO, RF, GB, NN achieved a diverse performances ranging from the lowest 69.76% (LASSO) to 82.39% (GB). LASSO failed to work on the raw ICD-9 codes with only 73.12% AUROC, implying that it is inadequate to place only sparse constraints on the otherwise high-dimensional, highly sparse, and highly noisy EHR code features. Most methods achieve high AUPRC (Fig 3e). This is partially due to the highly unbalanced nature of dataset (approximately 92% of observations are from positive class).

6.4 Predicting insulin usage from PopHR-Diabetes patients

Finally, we applied MixEHR-S to predict insulin usage among the over 78000 confirmed diabetes patients (Section 4.2). As the analysis above, we first qualitatively examined most predictive topics of insulin usage (Fig 4d,e). Here we focused on only the top 3 topics because of the rapid decrease of the predictive coefficients following them (Fig 4f). The most predictive topics T8 and T3 are asscoiated with the endocrinology and cardiology specialist types, which have strong association with diabetes diagnosis, whereas the top third topic T10 is connected with internal medicine and gastroenterology (Fig 4d; **Appendix** A.4).

The top ICD-9 codes under topics T8 and T3 are also highly enriched for three diabetes codes, namely 2500 diabetes mellitus without mention of complication, 2504 diabetes with renal manifestations, and 2509 diabetes with unspecified complications. We observed that the two renal failure ICDs under T8 with modest probability and interpreted them as renal complications of diabetes. Interestingly, topic T10 contains multiple ICD codes on cardiovascular conditions (ischemic heart disease, heart failure), which are diagnosed by cardiology specialists. Therefore, it is likely that topic T10 characterizes the cardiovascular complications of diabetes.

We then predicted the insulin usage outcome 6 months after the first diagnosis of diabetes (Fig 3c and f, Table 1). MixEHR-S achieved the highest AUROC (73.41%) and AUPRC (70.06%) or largely on-par with GRAM and DP, respectively. Other EHR-based models fell short. The discriminative models LASSO, RF, GB, and NN performed quite poorly with AUROC and AUPRC below 60%. Among the topic models, MixEHR outperforms LDA and sLDA, confirming

the benefits of learning the distribution of specialist-specific topics. Together, with MixEHR-S, we demonstrate the advantages of heterogeneous and and task-dependent topic modelling of the EHR data.

7 DISCUSSION

In this study, we presented MixEHR-S as an extension of our MixEHR [14]. Compared to the existing methods, we explicitly model the specialist assignments and ICD-coded diagnoses as latent topics. MixEHR-S can simultaneously infer the distribution of an arbitrary number of patient-specialist assignments and predict a binary disease label based on the learned disease topics. Compared to the simpler topic models, the relatively higher model complexity of our MixEHR-S does not actually incur higher computational costs as MixEHR-S is not greater than T times the complexity of LDA. This is attributable to our closed-form mean-field variational expectation-maximization updates and collapsed stochastic inference algorithm. We also estimate hyperparameters via fixed point iteration to avoid the possible impact of initial setting. We demonstrated MixEHR-S through a comprehensive set of experiments on both simulated data and two large-scale EHR databases with three targeted real-world applications. Consistent throughout our experiments, MixEHR-S not only infers meaningful specialist-dependent topics but also makes accurate prediction of target labels.

Besides the three applications, MixEHR-S can generalize to other EHR data with non-randomly distributed heterogeneous data types. For example, patients with brain disorder are more likely to have electroencephalography data, and patients with pulmonary problems will more likely to have chest radiograph. In other words, the specific data types observed among patients depend on the patient disease types.

In our future work, we will extend MixEHR-S for multi-class prediction model. This will allow us to model more heterogeneous patient cohort with related target diseases. EHR data are longitudinal. Our current MixEHR-S requires aggregating patient diagnoses over time to have a single collapsed data point to represent each patient. As one of our ongoing projects, we are exploring dynamic topic model [13] to properly account for time-dependent EHR data.

FUNDING

YL is supported by Microsoft Research. AM is supported by Canadian Institute of Health Research (CIHR) Foundation Grant.

REFERENCES

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 27–34.
[2] C Bishop. 2006. Pattern recognition and machine learning. *library.wisc.edu* (2006).
[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.
[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[5] You Chen, Joydeep Ghosh, Cosmin Adrian Bejan, Carl A Gunter, Siddharth Gupta, Abel Kho, David Liebovitz, Jimeng Sun, Joshua Denny, and Bradley Malin. 2015. Building bridges across electronic health record systems through inferred phenotypic topics. *J. Biomed. Inform.* 55, C (June 2015), 82–93.
[6] Y Cheng, F Wang, P Zhang, and J Hu. 2016. Risk prediction with electronic health records: A deep learning approach. *2016 SIAM International Conference* (2016), 432–440.
[7] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf. Proc.* 56 (Aug. 2016), 301–318.
[8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM - Graph-based Attention Model for Healthcare Representation Learning. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining* (2017).
[9] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sutter Health, Sacramento, United States, ACM Press, New York, New York, USA, 787–795.
[10] Suriya Gunasekar, Joyce C Ho, Joydeep Ghosh, Stephanie Kreml, Abel N Kho, Joshua C Denny, Bradley A Malin, and Jimeng Sun. 2016. Phenotyping using Structured Collective Matrix Factorization of Multi–source EHR Data. *arXiv* (Sept. 2016). arXiv:1609.04466 [stat.AP]
[11] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic Variational Inference. *Journal of Machine Learning Research* 14, 4 (2013), 1303–1347. http://jmlr.org/papers/v14/hoffman13a.html
[12] Shalmali Joshi, Suriya Gunasekar, David Sontag, and Joydeep Ghosh. 2016. Identifiable Phenotyping using Constrained Non-Negative Matrix Factorization. *arXiv.org* (Aug. 2016). arXiv:1608.00704v3 [stat.ML]
[13] David M Blei Lafferty and John D. 2015. Dynamic Topic Models. (May 2015), 1–9.
[14] Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M Biernacka, Euijung Ryu, Janet E Olson, Mark A Frye, Aihua Liu, Liming Guo, Ariane Marelli, Yuri Ahuja, Jose Davila-Velderrain, and Manolis Kellis. 2020. Inferring multimodal latent topics from electronic health records. *Nature Communications* 11, 1 (May 2020), 1–17.
[15] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv. org* (Nov. 2015).
[16] Xing Han Lu, Aihua Liu, Shih-Chieh Fuh, Yi Lian, Liming Guo, Yi Yang, Ariane Marelli, and Yue Li. 2021. Recurrent disease progression networks for modelling risk trajectory of heart failure. *PloS one* 16, 1 (Jan. 2021), e0245177–15.
[17] Jon D McAuliffe and David M Blei. 2008. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*, J C Platt, D Koller, Y Singer, and S T Roweis (Eds.). Curran Associates, Inc., 121–128.
[18] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. 121–128.
[19] Thomas Minka. 2000. Estimating a Dirichlet distribution.
[20] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* 6, 1 (May 2016), 1–10.
[21] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: A Convolutional Net for Medical Records. *arXiv. org* (July 2016), 22–30.
[22] Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. 2015. Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* 58, C (Dec. 2015), 156–165.
[23] Guido Antonio Powell, Yu T Luo, Aman Verma, David A Stephens, and David L Buckeridge. 2017. Multivariate and Longitudinal Health System Indicators. *Studies in health technology and informatics* 235 (2017), 266–270.
[24] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1, 1 (2018), 18.
[25] Narges Razavian and David Sontag. 2015. Temporal Convolutional Neural Networks for Diagnosis from Lab Tests. *arXiv. org* (Nov. 2015).
[26] Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. 2017. The Use of Autoencoders for Discovering Patient Phenotypes. *arXiv. org* (March 2017).
[27] Yee W Teh, David Newman, and Max Welling. 2007. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in neural information processing systems*. 1353–1360.
[28] R. Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the royal statistical society series b-methodological* 58 (1996), 267–288.
[29] Aman Verma, Guido Powell, Yu Luo, David Stephens, and David L. Buckeridge. 2018. Modeling disease progression in longitudinal EHR data using continuous-time hidden Markov models. arXiv:1812.00528 [cs.LG]
[30] Yuanyang Zhang, Richard Jiang, and Linda Petzold. 2017. Survival Topic Models for Predicting Outcomes for Trauma Patients. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 1497–1504.

A SUPPLEMENTARY INFORMATION

A.1 Description of variables of Graphical Model

Table S1: Description of Variables of MixEHR-S

Variable	Definition
y_j	Response connected to patient j
x_{ij}	ICD-9 code i of patient j
b_{ij}	Specialist type for ICD-9 code i of patient j
z_{ij}	Topic assignment for ICD code-9 i of patient j
θ_j	Topic mixture of patient j
β_k	Specialist mixture given topic k
η_{kt}	ICD-9 code mixture given topic k and specialist t
α	Dirichlet hyperparamter
ι	Dirichlet hyperparamter
ζ_k	Dirichlet hyperparamter
g_j	Latent disease liability of patient j
w	Linear coefficients
τ	Precision variable of Gaussian distribution for regression coefficient w

A.2 Simulation Results

For topic modeling, we quantitatively assessed the correlation between the inferred topics and the ground truth topics (FigS1). We computed the Pearson correlation between the inferred patient topic mixtures and the patient topic mixture matrix. MixEHR-S achieved an excellent topic recovery and outperformed LDA (Fig S2a). The improvement of MixEHR-S over LDA is attributable to explicitly modeling specialist topic distributions. In particular, LDA does not infer the specialist-patient assignments and multi-specialist topic distribution. For target label prediction, MixEHR-S achieved 93.54% AUROC and 95.26% AUPRC whereas LDA obtained only 80.17% and 66.41% AUROC and AUPRC, respectively (Fig S2b,c). The results therefore support the benefits of jointly modeling of multi-specialist topics and predicting labels compared to the LDA + Logistic Regression pipeline approach.

A.3 Baseline methods

We used the validation sets of the 3 datasets to tune the number of latent topics for topic-based models in terms of perplexity (with the chosen topic numbers in the brackets): MixEHR (K=40, 45, and 25 for CHD, COPD, diabetes-insulin, respectively); LDA (K=30, 35, and 20, respectively); sLDA (K=25, 35, and 20, respectively).

We also implemented machine learning algorithms that learn raw EHR data by using Python packages scikit-learn and PyTorch. For LASSO, we obtained best penalty parameter for the L1 term using grid search in scikit-learn package (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html). We set the number of decision trees and max depth as 300 and 3 for RF classifier. Besides, we chose the number of boosting stages and learning rate as 300 and 0.1 for GB model. We also implemented a two-layer fully connected neural network (NN) with a number of hidden units set to 100 for each layer.

We compared several existing EHR-based models. As in the original paper [20], DP was applied to reduce the dimensionality of EHR codes by training a denoising autoencoder with defaulted 500 hidden units followed by RF classifier from scikit-learn with 100 trees. For SiCNMF, we chose the number of factorization rank as 20 to avoid local minima which is suggested in the original paper.

GRAM learns the ICD-9 embedding from a 5-level taxonomy called the Clinical Classification Software (CCS) (https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixCMultiDX.txt). It then uses the ICD-9 embedding to project binary patient ICD-9 code vector onto a dense low-dimensional vector, which in turn serves as input features to a neural network for specific classification tasks. The original GRAM links the self-attention graph to a recurrent neural network (RNN) to do sequential diagnosis prediction. Since we performed only binary predictions on each target disease or outcome (i.e., CHD, COPD, and Diabetes-Insulin) by collapsing the time points, there is no recurrent unit. Therefore, the RNN (with one time point) reduced to a feedforward neural network. We use 200 attention weights and 100 embedding dimensions to represent each ICD code, and 100 one hidden layer with 200 hidden units in the feedforward network.

A.4 Supplementary Figures

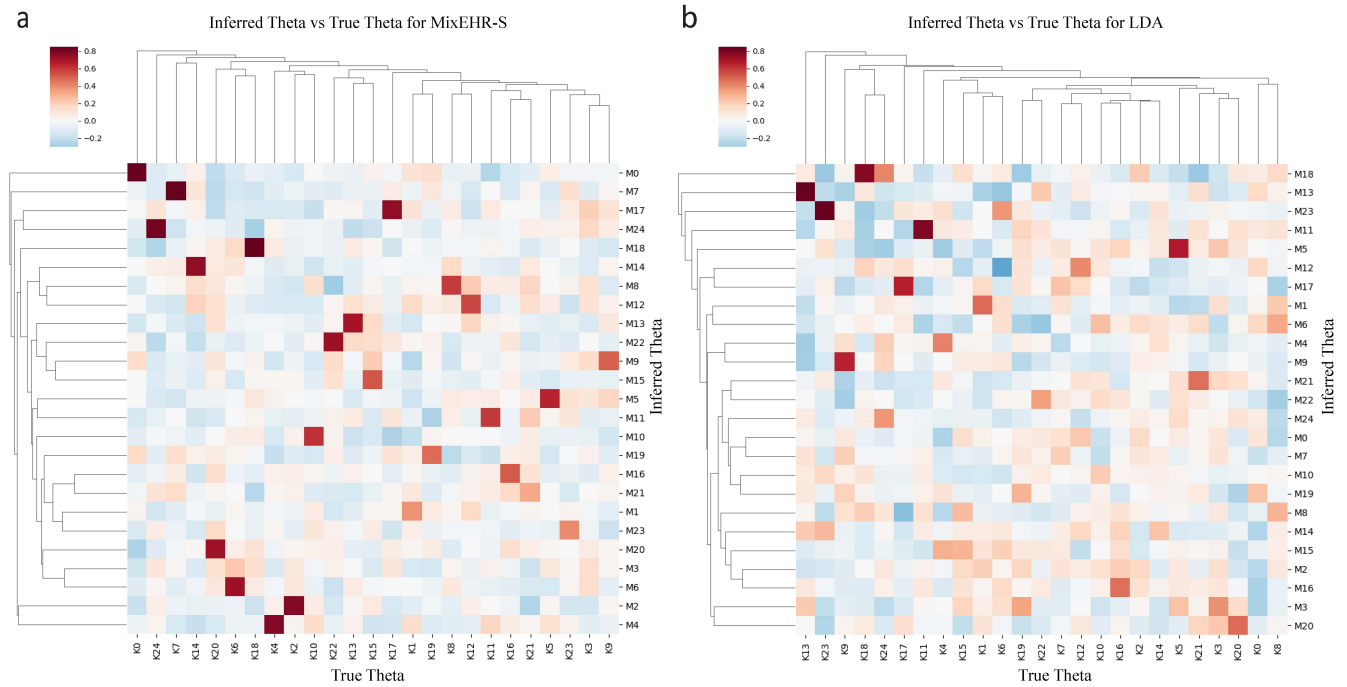

Figure S1: Topic recovery comparison from simulation dataset. To assess whether our model can recapitulate the groundtruth topics, we correlate every inferred topic with every groundtruth topic. In particular, we correlate the $D \times K$ inferred patient topic mixture $\hat{\theta}$ for D patients and K topics with the groundtruth $D \times K$ patient topic mixture matrix. The resulting Pearson correlation is therefore a $K \times K$ symmetric matrix. a Topic correlation using the inferred topic mixture by MixEHR-S model. b Topic correlation using the inferred topic mixture by LDA.

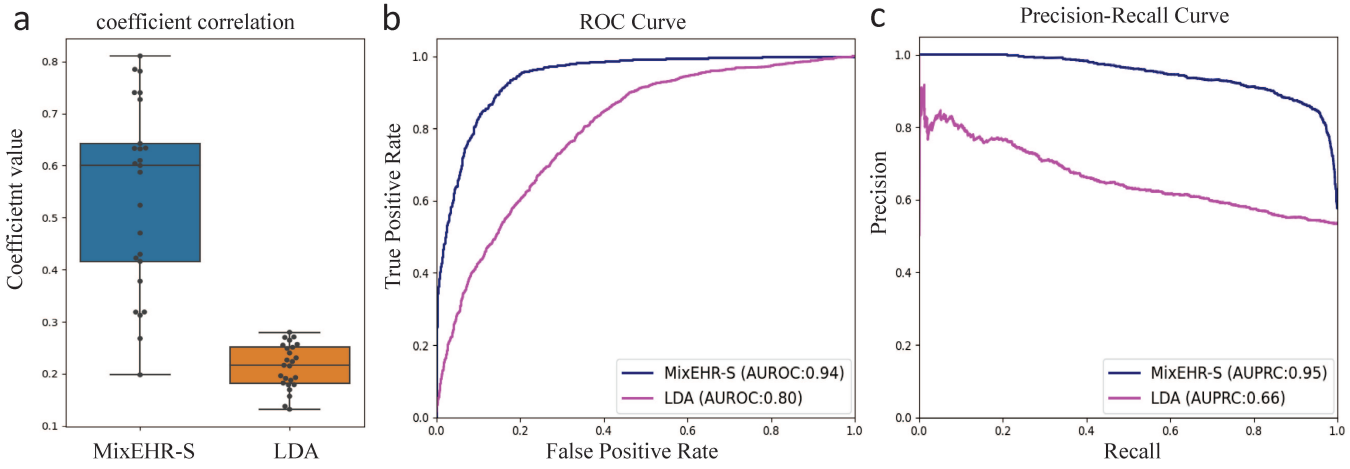

Figure S2: Evaluation of simulation data. a Correlation between true topics and inferred topics by MixEHR-S and LDA. Response prediction evaluated by b ROC and c PRC.

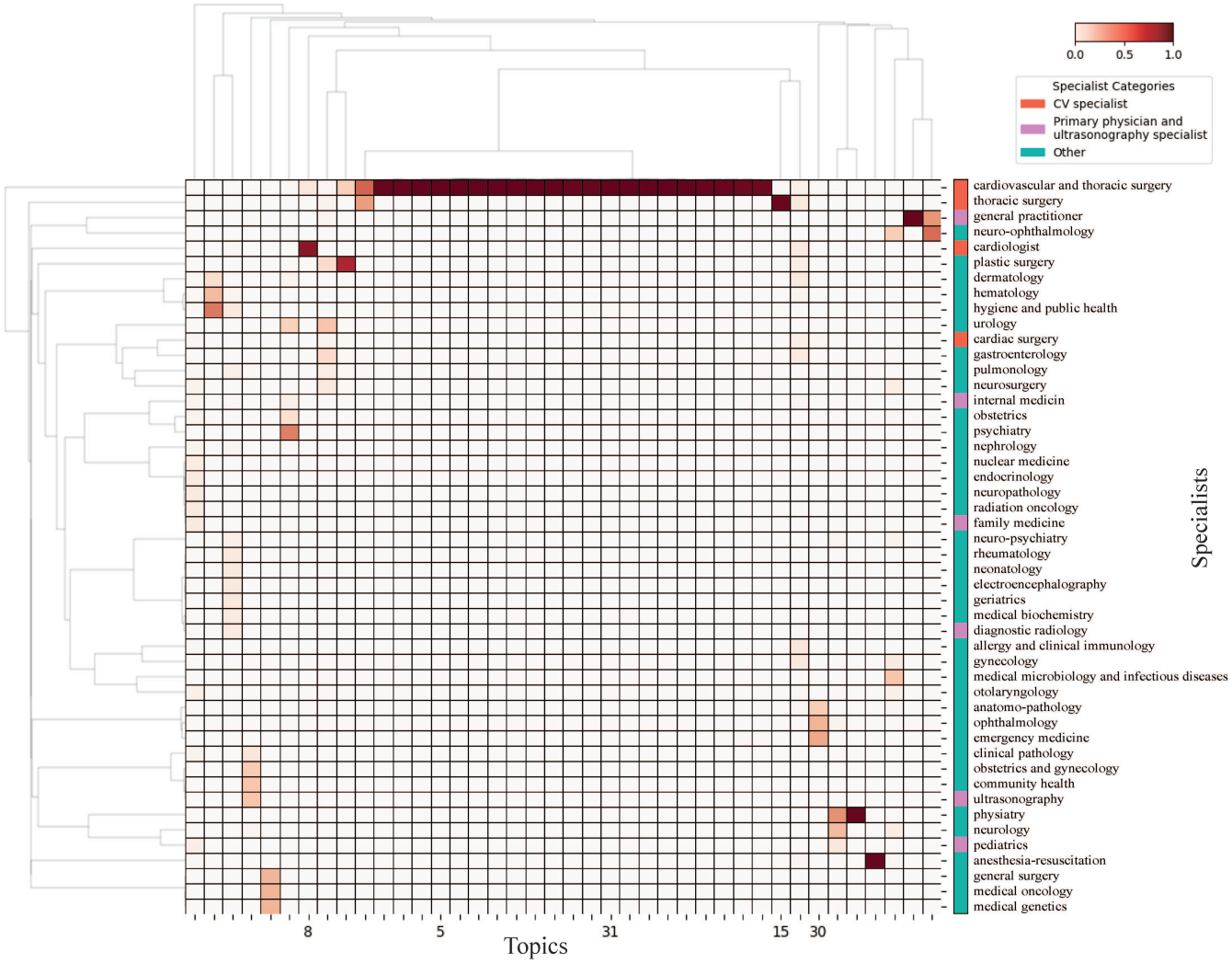

Figure S3: The original specialist topic plot for CHD dataset with with complete 48 specialists. Side bar shows the corresponding category for each medical speciality.

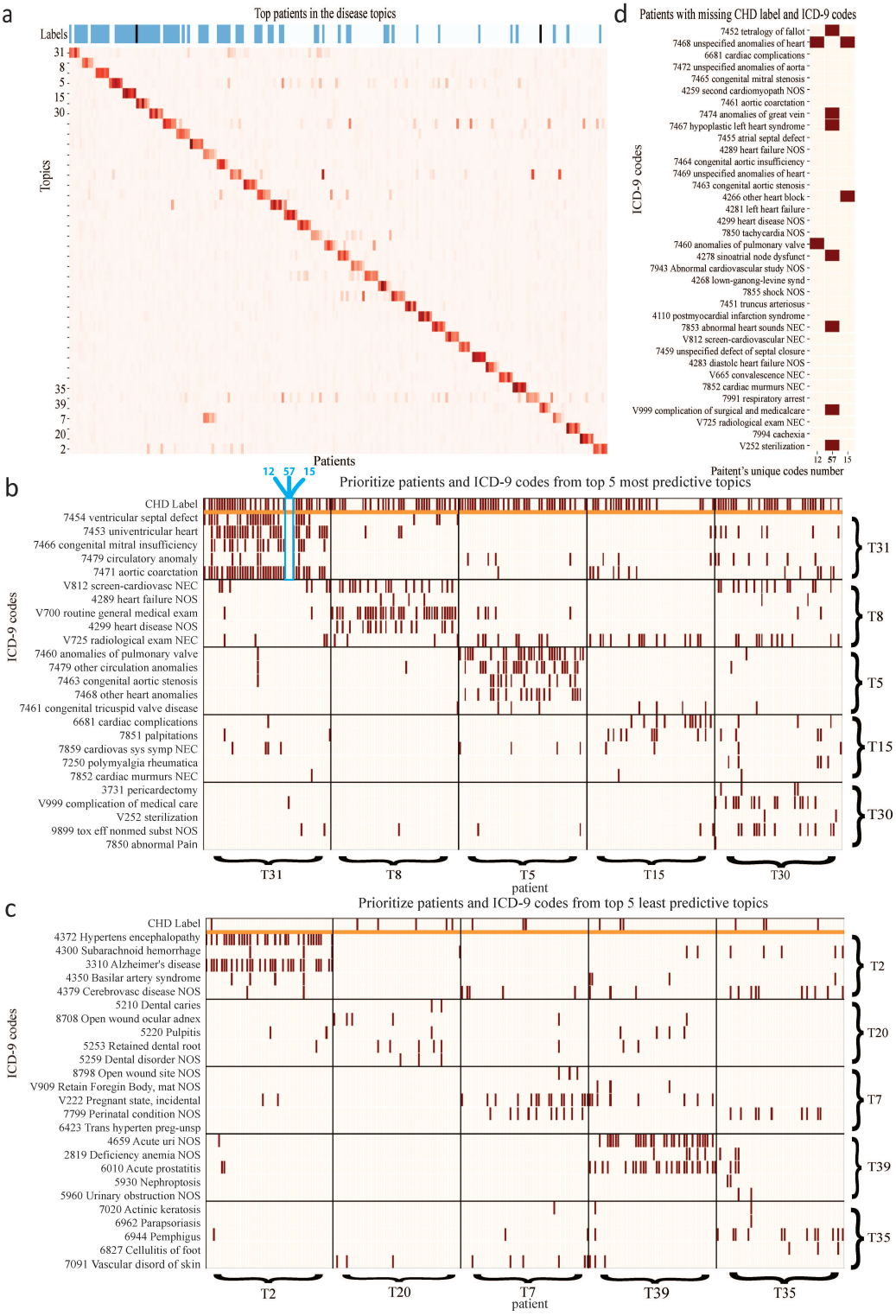

Figure S4: Prioritized patients by disease topic mixture from train data of CHD dataset. a. The top 5 patients under each topic. The patients with high proportions for the top positively and top negatively predictive topics are identified as high-risk patients and low-risk patients. b. Top ICD-9 codes of the high-risk patients. The top 5 ICD-9 codes (rows) from each topic were displayed for the top 50 patients (columns) under each topic. c. Top ICD-9 codes of the low-risk patients. d. The ICD-9 codes of the patients (12, 57, 15) highlighted in panel b.

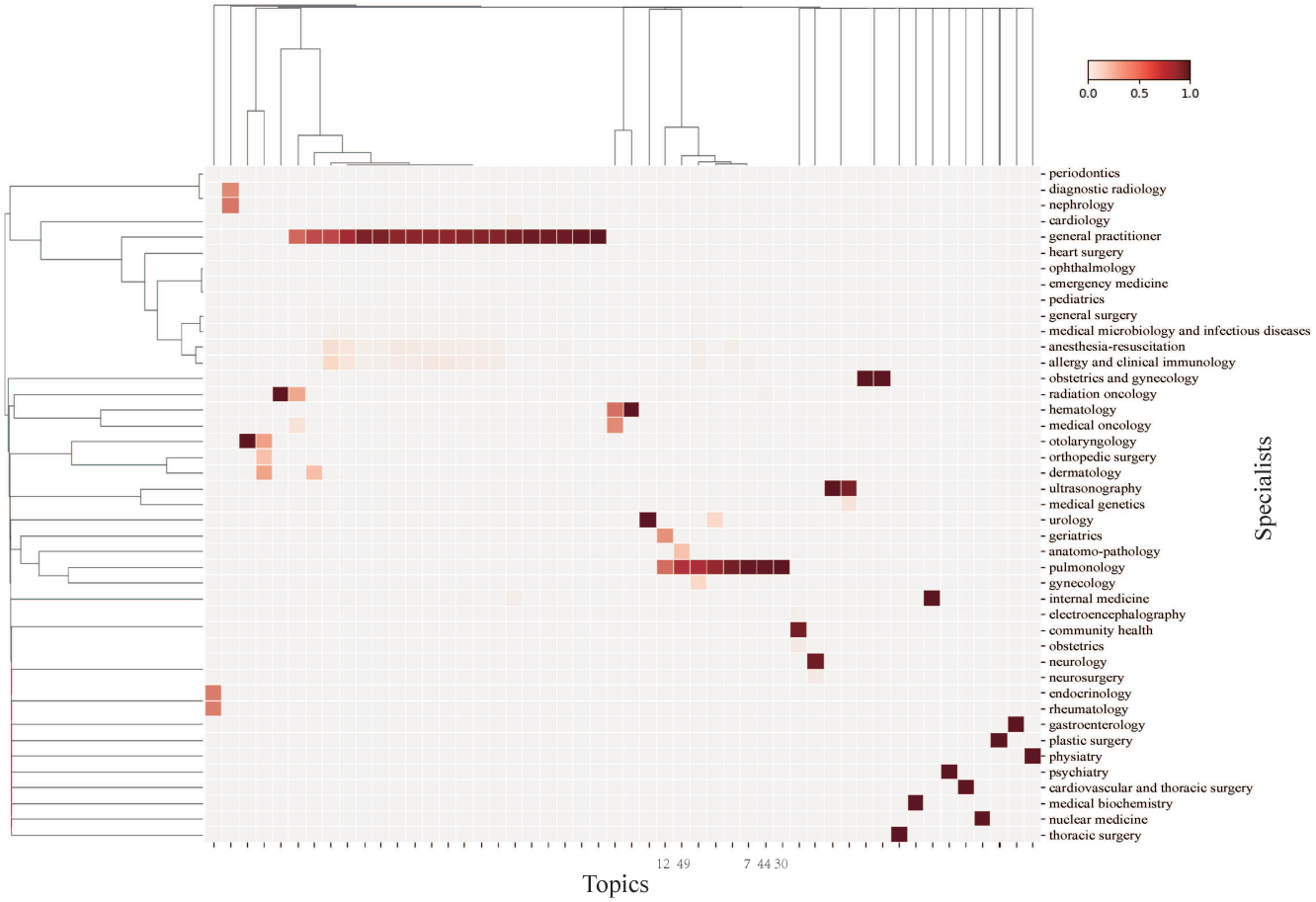

Figure S5: The original specialist topic plot for COPD patients of PopHR database with complete 43 specialists.

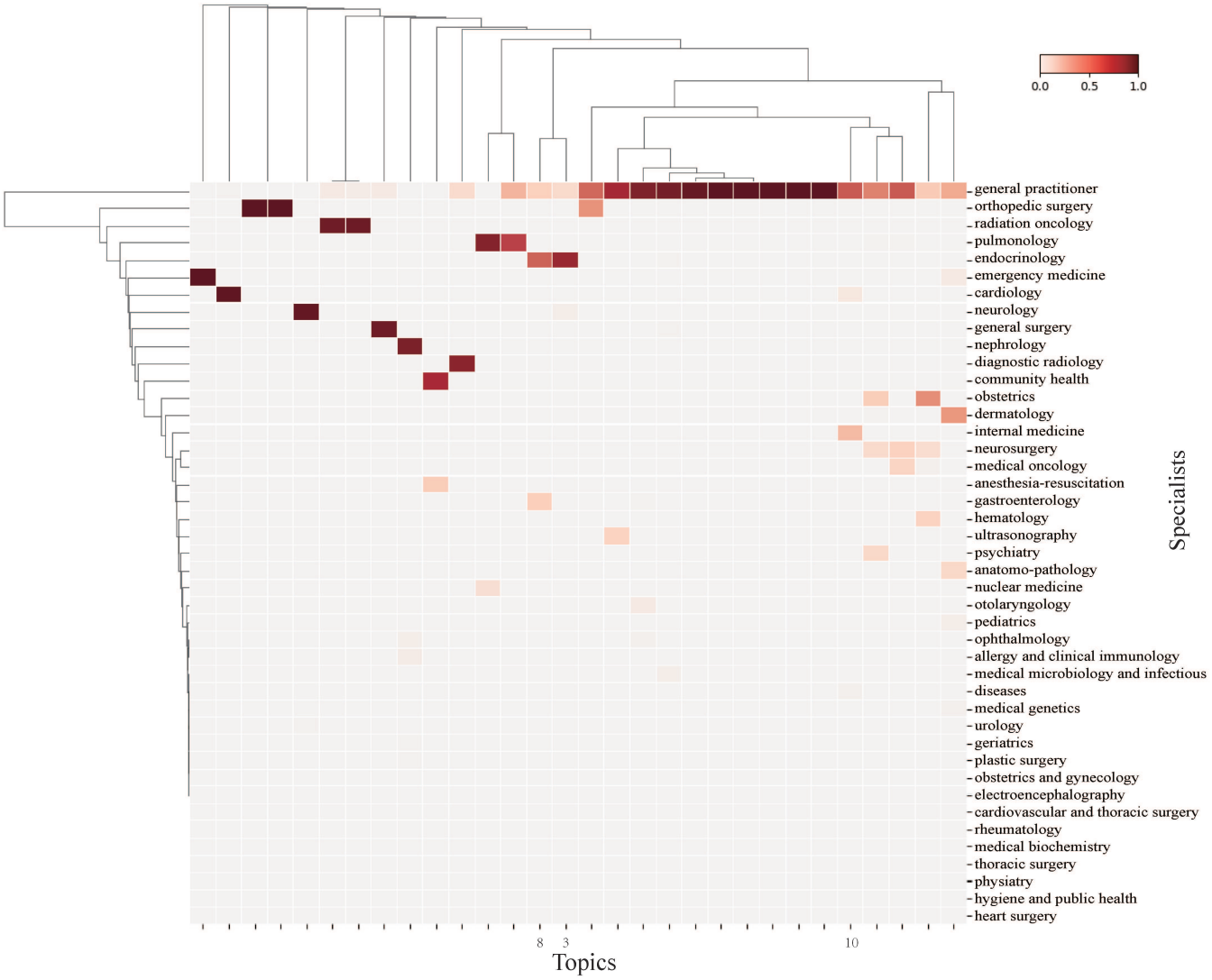

Figure S6: The original specialist topic plot for diabetes patients of PopHR database with complete 43 specialists.

B MIXEHR-S MODEL FULL DERIVATION

B.1 Derivation of MixEHR-S model likelihood

The full joint-likelihood for MixEHR-S model(Fig. 1) is:

$$\begin{aligned} p(\mathbf{z},\mathbf{b},\mathbf{x},\mathbf{y},\mathbf{g},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\beta}) =& p(\boldsymbol{\theta})p(\mathbf{z}\mid\boldsymbol{\theta})p(\mathbf{b}\mid\mathbf{z},\boldsymbol{\beta}_k)p(\mathbf{x}\mid\mathbf{b},\mathbf{z},\boldsymbol{\eta}_{kt})p(\boldsymbol{\beta}_k)p(\boldsymbol{\eta}_{kt}) \\ & p(\mathbf{y}\mid\mathbf{g})p(\mathbf{g}\mid\mathbf{z},\mathbf{w})p(\mathbf{w}) \end{aligned} \tag{16}$$

where the joint-likelihood involving the response latent variables $(\mathbf{y},\mathbf{g},\mathbf{w})$ from the supervised component of MixEHR-S model is:

$$\begin{aligned} p(\mathbf{y},\mathbf{g},\mathbf{w}\mid\mathbf{z},\tau) =& p(\mathbf{y}\mid\mathbf{g})p(\mathbf{g}\mid\mathbf{w},\mathbf{z})p(\mathbf{w}\mid\tau) \\ =& p(\mathbf{w}\mid\tau)\prod_j^D p(y_j\mid g_j)p(g_j\mid\mathbf{w},\bar{\mathbf{z}}_j) \\ =& \prod_j^D\left[\mathbb{1}(g_j>0)^{y_j}\mathbb{1}(g_j\le 0)^{(1-y_j)}(\frac{1}{2\pi})^{1/2}\exp(-\frac{(g_j-w^\intercal\bar{z}_j)^2}{2})\right] \\ & (\frac{1}{2\pi})^{K/2}(\tau)^{1/2}\exp(-\frac{\tau\mathbf{w}^\intercal\mathbf{w}}{2}) \end{aligned} \tag{17}$$

For the unsupervised learning component of MixEHR-S model, we obtain marginal likelihood by integrating out $\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\beta}$ due to the conjugacy properties of the Dirichlet and multinomial distributions:

$$\begin{aligned} p(\mathbf{z},\mathbf{b},\mathbf{x},\mathbf{g}) =& \int p(\boldsymbol{\theta})p(\mathbf{z}\mid\boldsymbol{\theta})d\boldsymbol{\theta}\times\int p(\mathbf{b}\mid\mathbf{z},\boldsymbol{\beta_k})p(\boldsymbol{\beta}_k)d\boldsymbol{\beta}_k\times\int p(\mathbf{x}\mid\mathbf{b},\mathbf{z},\boldsymbol{\eta}_{kt})p(\boldsymbol{\eta}_{kt})d\boldsymbol{\eta}_{kt} \\ =& p(\mathbf{z})p(\mathbf{b}\mid\mathbf{z})p(\mathbf{x}\mid\mathbf{b},\mathbf{z}) \\ =& \prod_j^D\frac{\Gamma(\sum_k\alpha_k)}{\prod_k\Gamma(\alpha_k)}\frac{\prod_k\Gamma(\alpha_k+\sum_i^{M_j}[z_{ij}=k])}{\Gamma(\sum_k\alpha_k+\sum_i^{M_j}[z_{ij}=k])} \\ & \times\prod_k^K\frac{\Gamma(\sum_t\iota_t)}{\prod_t\Gamma(\iota_t)}\frac{\prod_t\Gamma(\iota_t+\sum_j^D\sum_i^{M_j}[z_{ij}=k,b_{ij}=t])}{\Gamma(\sum_t\iota_t+\sum_j^D\sum_i^{M_j}[z_{ij}=k,b_{ij}=t])} \\ & \times\prod_k^K\prod_t^T\frac{\Gamma(\sum_w\zeta_{kw})}{\prod_w\Gamma(\zeta_{kw})}\frac{\prod_w\Gamma(\zeta_{kw}+\sum_j^D\sum_i^{M_j}[b_{ij}=t,z_{ij}=k,x_{ij}=w])}{\Gamma(\sum_w\zeta_{kw}+\sum_j^D\sum_i^{M_j}[b_{ij}=t,z_{ij}=k,x_{ij}=w])} \\ =& \prod_j^D\frac{\Gamma(\sum_k\alpha_k)}{\prod_k\Gamma(\alpha_k)}\frac{\prod_k\Gamma(\alpha_k+n_{jk})}{\Gamma(\sum_k\alpha_k+n_{jk})} \\ & \times\prod_k^K\frac{\Gamma(\sum_t\iota_t)}{\prod_t\Gamma(\iota_t)}\frac{\prod_t\Gamma(\iota_t+m_{tk})}{\Gamma(\sum_t\iota_t+m_{tk})} \\ & \times\prod_k^K\prod_t^T\frac{\Gamma(\sum_w\zeta_{kw})}{\prod_w\Gamma(\zeta_{kw})}\frac{\prod_w\Gamma(\zeta_{kw}+p_{ktw})}{\Gamma(\sum_w\zeta_{kw}+p_{ktw})} \end{aligned} \tag{18}$$

where (n_{jk},m_{tk},p_{twk}) are the sufficient statistics.

$$\begin{aligned} n_{jk} &= \sum_i^{M_j}[z_{ij}=k] \\ m_{tk} &= \sum_j^D\sum_i^{M_j}[z_{ij}=k,b_{ij}=t] \\ p_{twk} &= \sum_j^D\sum_i^{M_j}[z_{ij}=k,b_{ij}=t,x_{ij}=w] \end{aligned} \tag{19}$$

We then can calculate the likelihood of unsupervised component of MixEHR-S model given the following analytical expressions for each distribution of $(\mathbf{z}, \mathbf{b}, \mathbf{x})$:

$$\begin{aligned} p(\mathbf{z}) &= \prod_j^D \int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k^K \theta_{jk}^{\alpha_k - 1} \prod_i^{M_j} \theta_{jk}^{[z_{ij}=k]} d\boldsymbol{\theta}_j \\ &= \prod_j^D \int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k^K \theta_{jk}^{\alpha_k + \sum_i^{M_j} [z_{ij}=k] - 1} d\boldsymbol{\theta}_j \\ &= \prod_j^D \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + \sum_i^{M_j} [z_{ij} = k])}{\Gamma(\sum_k \alpha_k + \sum_i^{M_j} [z_{ij} = k])} \end{aligned} \tag{20}$$

$$\begin{aligned} p(\mathbf{b} \mid \mathbf{z}) &= \prod_k^K \int \frac{\Gamma(\sum_t \iota_t)}{\prod_t \Gamma(\iota_t)} \prod_t^T \beta_{kt}^{\iota_t - 1} \prod_j^D \prod_i^{M_j} \beta_{kt}^{[z_{ij}=k, b_{ij}=t]} d\boldsymbol{\beta}_k \\ &= \int \prod_k^K \frac{\Gamma(\sum_t \iota_t)}{\prod_t \Gamma(\iota_t)} \prod_t^T \beta_{kt}^{\iota_t + \sum_j^D \sum_i^{M_j} [z_{ij}=k, b_{ij}=t] - 1} d] \boldsymbol{\beta}_k \\ &= \prod_k^K \frac{\Gamma(\sum_t \iota_t)}{\prod_t \Gamma(\iota_t)} \frac{\prod_t \Gamma(\iota_t + \sum_j^D \sum_i^{M_j} [z_{ij} = k, b_{ij} = t])}{\Gamma(\sum_t \iota_t + \sum_j^D \sum_i^{M_j} [z_{ij} = k, b_{ij} = t])} \end{aligned} \tag{21}$$

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{b}, \mathbf{z}) &= \prod_k^K \prod_t^T \int \frac{\Gamma(\sum_w \zeta_{kw})}{\prod_w \Gamma(\zeta_{kw})} \prod_w \eta_{ktw}^{\zeta_{kw} - 1} \prod_j^D \prod_i^{M_j} \eta_{ktw}^{[b_{ij}=t, z_{ij}=k, x_{ij}=w]} d\boldsymbol{\eta}_{kt} \\ &= \prod_k^K \prod_t^T \int \frac{\Gamma(\sum_w \zeta_{kw})}{\prod_w \Gamma(\zeta_{kw})} \prod_w \eta_{ktw}^{\zeta_{kw} + \sum_j^D \sum_i^{M_j} [b_{ij}=t, z_{ij}=k, x_{ij}=w] - 1} d\boldsymbol{\eta}_{kt} \\ &= \prod_k^K \prod_t^T \frac{\Gamma(\sum_w \zeta_{kw})}{\prod_w \Gamma(\zeta_{kw})} \frac{\prod_w \Gamma(\zeta_{kw} + \sum_j^D \sum_i^{M_j} [b_{ij} = t, z_{ij} = k, x_{ij} = w])}{\Gamma(\sum_w \zeta_{kw} + \sum_j^D \sum_i^{M_j} [b_{ij} = t, z_{ij} = k, x_{ij} = w])} \end{aligned} \tag{22}$$

B.2 Derivation of conditional distribution $z_{ij} = k$

In order to derive the conditional distribution $p(z_{ij} = k \mid \mathbf{z}_{(-ij)}, \mathbf{b}, \mathbf{x}, \mathbf{g})$ which excludes specific z_{ij} and $(\mathbf{b}, \mathbf{x}, \mathbf{g})$, we drop the constant variables (α, ι, ζ), and also drop terms that depend on (i, j) to get the conditional distribution:

$$
\begin{aligned}
p(z_{ij}=k \mid \mathbf{z}_{(-ij)}, \mathbf{b}, \mathbf{x}, \mathbf{g}) &= \frac{p(z_{ij}, \mathbf{z}_{(-ij)}, \mathbf{b}, \mathbf{x}, \mathbf{g})}{\sum_{k'}^{K} p(z_{ij}, \mathbf{z}_{(-ij)}, \mathbf{b}, \mathbf{x}, \mathbf{g})} \\
&\propto p(z_{ij}, \mathbf{z}_{(-ij)} \mid \alpha) p(\mathbf{b} \mid z_{ij}, \mathbf{z}_{(-ij)}, \iota) p(\mathbf{x} \mid z_{ij}, \mathbf{z}_{(-ij)}, \mathbf{b}, \zeta_k) p(\mathbf{g} \mid z_{ij}, \mathbf{z}_{(-ij)}, \mathbf{w}) \\
&\propto (\alpha_k + n_{jk}^{-(i,j)}) \frac{\iota_{b_{ij}} + m_{kb_{ij}}^{-(i,j)}}{\sum_t \iota_t + m_{kt}^{-(i,j)}} \frac{\zeta_{kx_{ij}} + p_{kb_{ij}x_{ij}}^{-(i,j)}}{\sum_w \zeta_{kw} + p_{kb_{ij}w}^{-(i,j)}} \\
&\quad \times \exp\Big\{ \mathbf{w}^\intercal \Big(\frac{1}{M_j} \Big(\sum_{i' \neq i} \mathbf{z}_{i'j} + \mathbf{z}_{ij} \Big) \Big) g_j - \frac{1}{2} \mathbf{w}^\intercal \Big(\frac{1}{M_j^2} \Big(\sum_{n \neq i}^{M_j} \sum_{m \neq n}^{M_j} \mathbf{z}_{nj} \mathbf{z}_{mj}^\intercal + \sum_{n \neq i}^{M_j} \mathrm{diag}(\mathbf{z}_{nj}) \\
&\quad + \sum_{m \neq i}^{M_j} \mathbf{z}_{ij} \mathbf{z}_{mj}^\intercal + \sum_{n \neq i}^{M_j} \mathbf{z}_{nj} \mathbf{z}_{ij}^\intercal + \mathrm{diag}(\mathbf{z}_{ij}) \Big) \Big) \mathbf{w} \Big\} \\
&\propto (\alpha_k + n_{jk}^{-(i,j)}) \frac{\iota_{b_{ij}} + m_{kb_{ij}}^{-(i,j)}}{m_{k.}^{-(i,j)} + \sum_t \iota_t} \frac{\zeta_{kx_{ij}} + p_{kb_{ij}x_{ij}}^{-(i,j)}}{p_{kb_{ij}.}^{-(i,j)} + \sum_w \zeta_{kw}} \\
&\quad \exp\Big\{ \frac{\mathbf{w}^\intercal \mathbf{z}_{ij} g_j}{M_j} - \frac{1}{2M_j^2} \mathbf{w}^\intercal \Big(\sum_{m \neq i}^{M_j} \mathbf{z}_{ij} \mathbf{z}_{mj}^\intercal + \sum_{n \neq i}^{M_j} \mathbf{z}_{nj} \mathbf{z}_{ij}^\intercal + \mathrm{diag}(\mathbf{z}_{ij}) \Big) \mathbf{w} \Big\}
\end{aligned} \tag{23}
$$

where the notation $-(i,j)$ indicates that the term (i,j) is excluded and $\mathbf{z}_{ij}$ is a *one-hot* K-dimensional binary vector that sets the kth topic equal to 1 and remaining values equal to zero.

B.3 Derivation of collapsed variational Bayesian inference (JCVB0)

To approximate the posterior distributions, we perform variational inference and obtain the ELBO $\mathcal{L}(\Theta)$ in Eq. (7) by using Jensen's inequality:

$$
\begin{aligned}
\log p(\mathbf{x}, \mathbf{b}, \mathbf{y}) \geq & \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log \frac{p(\mathbf{z}, \mathbf{g}, \mathbf{w}, \mathbf{b}, \mathbf{x}, \mathbf{y})}{q(\mathbf{z}, \mathbf{g}, \mathbf{w})}] \\
= & \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log p(\mathbf{z}, \mathbf{g}, \mathbf{w}, \mathbf{b}, \mathbf{x}, \mathbf{y})] - \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log q(\mathbf{z}, \mathbf{g}, \mathbf{w})] = \mathcal{L}(\Theta)
\end{aligned} \tag{24}
$$

The ELBO can be broken down using all variables in the MixEHR-S model with their respective expectations:

$$
\begin{aligned}
\mathcal{L}(\Theta) &= \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log p(\mathbf{z}, \mathbf{g}, \mathbf{w}, \mathbf{b}, \mathbf{x}, \mathbf{y})] - \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log q(\mathbf{z}, \mathbf{g}, \mathbf{w})] \\
&= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z} \mid \alpha)] + \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{b} \mid \mathbf{z}, \iota)] + \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{x} \mid \mathbf{b}, \mathbf{z}, \zeta)] \\
&+ \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})} [\log p(\mathbf{g} \mid \mathbf{z}, \mathbf{w})] + \mathbb{E}_{q(\mathbf{w})} [\log p(\mathbf{w} \mid \tau)] + \mathbb{E}_{q(\mathbf{g})} [\log p(\mathbf{y} \mid \mathbf{g})] \\
&- \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z} \mid \boldsymbol{\gamma})] - \mathbb{E}_{q(\mathbf{g})} [\log q(\mathbf{g} \mid \boldsymbol{\lambda})] - \mathbb{E}_{q(\mathbf{w})} [\log q(\mathbf{w} \mid \mathbf{m}, \mathbf{S})]
\end{aligned} \tag{25}
$$

where the expectations of the above terms are:

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z} \mid \alpha)] &= \sum_{j}^{D} \mathbb{E}_{q(\mathbf{z})} [\log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k) \\
&+ \sum_k \log \Gamma(\alpha_k + \sum_i^{M_j} [z_{ij} = k]) - \log \Gamma(\sum_k \alpha_k + \sum_i^{M_j} [z_{ij} = k]) \\
&= \sum_j^D \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k) + \sum_k \log \Gamma(\alpha_k + E_q[n_{jk}]) - \log \Gamma(\sum_k \alpha_k + E_q[n_{jk}])
\end{aligned} \tag{26}
$$

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{b} \mid \mathbf{z}, \iota)] &= \sum_{k}^{K} \log\Gamma(\sum_{t} \iota_t) - \sum_{t} \log\Gamma(\iota_t) + \sum_{t} \log\Gamma(\iota_t + \sum_{j}^{D}\sum_{i}^{M_j}[b_{ij} = t]\gamma_{ijk}) \\
&\quad - \log\Gamma(\sum_{t} \iota_t + \sum_{j}^{D}\sum_{i}^{M_j}[b_{ij} = t]\gamma_{ijk}) \\
&= \sum_{k}^{K} \log\Gamma(\sum_{t} \iota_t) - \sum_{t} \log\Gamma(\iota_t) + \sum_{t} \log\Gamma(\iota_t + E_q[m_{tk}]) \\
&\quad - \log\Gamma(\sum_{t} \iota_t + E_q[m_{tk}])
\end{aligned} \tag{27}$$

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{b}, \mathbf{z}, \zeta)] &= \sum_{k}^{K}\sum_{t}^{T} \log\Gamma(\sum_{w} \zeta_{kw}) - \sum_{w} \log\Gamma(\zeta_{kw}) \\
&\quad + \sum_{w} \log\Gamma(\zeta_{kw} + \sum_{j}^{D}\sum_{i}^{M_j}[b_{ij} = t, x_{ij} = w]\gamma_{ijk}) \\
&\quad - \log\Gamma(\sum_{w} \zeta_{kw} + \sum_{j}^{D}\sum_{i}^{M_j}[b_{ij} = t, x_{ij} = w]\gamma_{ijk}) \\
&= \sum_{k}^{K}\sum_{t}^{T} \log\Gamma(\sum_{w} \zeta_{kw}) - \sum_{w} \log\Gamma(\zeta_{kw}) + \sum_{w} \log\Gamma(\zeta_{kw} + E_q[p_{twk}]) \\
&\quad - \log\Gamma(\sum_{w} \zeta_{kw} + E_q[p_{twk}])
\end{aligned} \tag{28}$$

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[\log p(\mathbf{g} \mid \mathbf{z}, \mathbf{w})] &= \sum_{j} \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[\log p(g_j \mid \mathbf{w}, \bar{\mathbf{z}}_j)] \\
&= \sum_{j} \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[\log \mathcal{N}(g_j \mid \mathbf{w}^\intercal \bar{\mathbf{z}}_j, 1)] \\
&= \sum_{j}^{D} -\frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[(g_j - \mathbf{w}^\intercal \bar{\mathbf{z}}_j)^2] \\
&= \sum_{j}^{D} -\frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[g_j^2 - 2\mathbf{w}^\intercal \bar{\mathbf{z}}_j g_j + \mathbf{w}^\intercal \bar{\mathbf{z}}_j \bar{\mathbf{z}}_j^\intercal \mathbf{w}] \\
&= \sum_{j}^{D} -\frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q(\mathbf{g})}[g_j^2] + \mathbb{E}_{q(\mathbf{z},\mathbf{g},\mathbf{w})}[\mathbf{w}^\intercal \bar{\mathbf{z}}_j g_j] - \frac{1}{2}\mathbb{E}_{q(\mathbf{z},\mathbf{w})}[\mathbf{w}^\intercal \bar{\mathbf{z}}_j \bar{\mathbf{z}}_j^\intercal \mathbf{w}] \\
&= \sum_{j}^{D} -\frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q(\mathbf{g})}[g_j^2] + \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}]^\intercal \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j]\mathbb{E}_{q(\mathbf{g})}[g_j] \\
&\quad - \frac{1}{2}\mathbb{E}_{q(\mathbf{w})}\left[\mathbf{w}^\intercal \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j \bar{\mathbf{z}}_j^\intercal]\mathbf{w}\right]
\end{aligned} \tag{29}$$

$$\begin{aligned}\mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{w} \mid \tau)] &= \mathbb{E}_{q(\mathbf{w})}[\log \prod_{k}^{K} (\frac{1}{2\pi})^{K/2} (\tau_k)^{1/2} \exp(-\frac{\tau_k w_k^2}{2})] \\ &= \mathbb{E}_{q(\mathbf{w})}[-\frac{K}{2}\log 2\pi + \frac{1}{2}\log \tau_k - \frac{\tau_k w_k^2}{2}] \\ &= -\frac{K}{2}\log 2\pi + \frac{1}{2}\log \tau_k - \frac{\tau_k \mathbb{E}_{q(\mathbf{w})}[w_k^2]}{2} \\ &= -\frac{K}{2}\log 2\pi + \frac{1}{2}\log \tau_k - \frac{\tau_k (m_k^2 + S_{kk})}{2} \end{aligned} \tag{30}$$

$$\mathbb{E}_{q(\mathbf{g})}[\log p(\mathbf{y} \mid \mathbf{g})] = \sum_j \mathbb{E}_{q(\mathbf{g})}[\log p(y_j \mid g_j)] = 0 \tag{31}$$

$$\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z} \mid \boldsymbol{\gamma})] = \sum_{ijk} \gamma_{ijk} \log \gamma_{ijk} \tag{32}$$

$$\begin{aligned}\mathbb{E}_{q(\mathbf{g})}[\log q(\mathbf{g} \mid \boldsymbol{\lambda})] &= \sum_j^D \mathbb{E}_{q(\mathbf{g})}[\log q(g_j \mid \lambda_j)] \\ &= \sum_j^D \mathbb{E}_{q(\mathbf{g})}[\log \{\mathcal{TN}^+(g_j; \lambda_j, 1)^{y_j} \mathcal{TN}^-(g_j; \lambda_j, 1)^{1-y_j}\}] \\ &= \sum_j^D \mathbb{E}_{q(\mathbf{g})}[\log \{\mathcal{N}(g_j; \lambda_j, 1)(\frac{1}{1-\Phi_j})^{y_j} (\frac{1}{\Phi_j})^{1-y_j}\}] \\ &= \sum_j^D \mathbb{E}_{q(\mathbf{g})}[\log \mathcal{N}(g_j; \lambda_j, 1)] - \{y_j \log(1-\Phi_j) + (1-y_j)\log \Phi_j\} \\ &= \sum_j^D -\frac{1}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{q(g)}[g_j^2] + \mathbb{E}_{q(g)}[g_j]\lambda_j - \frac{1}{2}\lambda_j^2 \\ &\quad + y_j \log(1-\Phi_j) + (1-y_j)\log \Phi_j \end{aligned} \tag{33}$$

$$\begin{aligned}\mathbb{E}_{q(\mathbf{w})}[\log q(\mathbf{w} \mid \mathbf{m}, \mathrm{S})] &= \mathbb{E}_{q(\mathbf{w})}[-\frac{K}{2}\log 2\pi - \sum_k^K (\frac{1}{2}\log S_{kk} + \frac{(w_k - m_k)^2}{2S_{kk}})] \\ &= -\frac{K}{2}\log 2\pi - \frac{K}{2} - \sum_k^K \frac{1}{2}\log S_{kk} \end{aligned} \tag{34}$$

We use mean-field factorization for variational inference. For the latent topic variable $\mathbf{z}$, we posit a multinomial distribution with variational parameter $\boldsymbol{\gamma}$:

$$q(\mathbf{z} \mid \boldsymbol{\gamma}) = \prod_{ijk} \gamma_{ijk}^{[z_{ij}=k]}, \ \log q(\mathbf{z} \mid \boldsymbol{\gamma}) = \sum_{i,j,k} [z_{ij} = k] \log \gamma_{ijk} \tag{35}$$

We can maximize the ELBO with respect to the variational parameter γ_{ijk} by calculating the expectation $\mathbb{E}_{q(\mathbf{z}^{-i,j})}[\ln q(\mathbf{z}_{ij}|\boldsymbol{\gamma})]$. The expect value of γ_{ijk} is computed over variables $(\mathbf{z}, \mathbf{b}, \mathbf{x}, \mathbf{g})$:

$$\begin{aligned}\log \gamma_{ijk} &\propto \mathbb{E}_{q(\mathbf{z}^{-(i,j)}, \mathbf{g}, \mathbf{w})}[\log p(\mathbf{z}_{ij} = k \mid \mathbf{z}_{(-ij)}, \mathbf{b}, \mathbf{x}, \mathbf{g})] \\ &= \mathbb{E}_{q(\mathbf{z}^{-(i,j)}, \mathbf{g}, \mathbf{w})}[\log p(\mathbf{z}, \mathbf{b}, \mathbf{x}, \mathbf{g})] \\ &= \mathbb{E}_{q(\mathbf{z}^{-(i,j)}, \mathbf{g}, \mathbf{w})}[\log p(\mathbf{z} \mid \alpha) + \log p(\mathbf{b} \mid \mathbf{z}, \iota) \\ &\quad + \log p(\mathbf{x} \mid \mathbf{z}, \mathbf{b}, \zeta_k) + \log p(\mathbf{g} \mid \mathbf{w}, \mathbf{z})] \end{aligned} \tag{36}$$

For the unsupervised component of MixEHR-S model, we update the first three terms whereas ignoring the last term $\log p(\mathbf{g} \mid \mathbf{w}, \mathbf{z})$ in Eq. (36). The variational update of γ_{ijk} is obtained by normalizing itself and taking the expectation of the conditional distribution in Eq. (23) (see Appendix B.2):

$$\begin{aligned}\gamma_{ijk} &= (\alpha_k + E_{q(\mathbf{z}^{-(i,j)})}[n_{jk}^{-(i,j)}]) \\ &\frac{\iota_{b_{ij}} + E_{q(\mathbf{z}^{-(i,j)})}[m_{kb_{ij}}^{-(i,j)}]}{E_{q(\mathbf{z}^{-(i,j)})}[m_{k.}^{-(i,j)}] + \sum_t \iota_t} \frac{\zeta_{kx_{ij}} + E_{q(\mathbf{z}^{-(i,j)})}[p_{kb_{ij}x_{ij}}^{-(i,j)}]}{E_{q(\mathbf{z}^{-(i,j)})}[p_{kb_{ij}.}^{-(i,j)}] + \sum_w \zeta_{kw}}\end{aligned} \tag{37}$$

where the above expected sufficient statistics are calculated by:

$$E_{\mathbf{z}^{-(i,j)}}[n_{jk}^{-(i,j)}] = \sum_{i' \neq i}^{M_j} \gamma_{i'jk} \tag{38}$$

$$E_{\mathbf{z}^{-(i,j)}}[m_{b_{ij}k}^{-(i,j)}] = \sum_{j' \neq j}^{D} \sum_{i}^{M_{j'}} [b_{ij'} = b_{ij}] \gamma_{ij'k} \tag{39}$$

$$E_{\mathbf{z}^{-(i,j)}}[p_{b_{ij}x_{ij}k}^{-(i,j)}] = \sum_{j' \neq j}^{D} \sum_{i}^{M_{j'}} [b_{ij'} = b_{ij}, x_{ij'} = x_{ij}] \gamma_{ij'k} \tag{40}$$

For the latent liability variable $\mathbf{g}$, we propose a truncated Gaussian distribution as variational distribution with parameter $\boldsymbol{\lambda}$:

$$\begin{aligned} q(g_j \mid \lambda_j) &= \begin{cases} \mathcal{TN}_+(g_j; \lambda_j, 1), & \text{if} \quad y_j = 1. \\ \mathcal{TN}_-(g_j; \lambda_j, 1), & \text{if} \quad y_j = 0. \end{cases} \\ &\propto \begin{cases} \mathbb{1}(g_j > 0) \exp\left\{ -\frac{1}{2} g_j^2 + \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\intercal] \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] g_j \right\}, & \text{if} \quad y_j = 1. \\ \mathbb{1}(g_j \leq 0) \exp\left\{ -\frac{1}{2} g_j^2 + \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\intercal] \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] g_j \right\}, & \text{if} \quad y_j = 0. \end{cases} \end{aligned} \tag{41}$$

$$\begin{aligned} \log q(g_j \mid \mathbf{z}, \mathbf{w}) &= \mathbb{E}_{q(\mathbf{z},\mathbf{w})}[\log p(g_j \mid \bar{\mathbf{z}}_j, \mathbf{w})] + \log p(y_j \mid g_j) \\ &= y_j \log \mathbb{1}(g_j > 0) + (1 - y_j) \mathbb{1}(g_j \leq 0) - \frac{1}{2} g_j^2 + \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\intercal] \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] g_j \end{aligned} \tag{42}$$

where the update of the variational parameter λ_j is:

$$\lambda_j = \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\intercal] \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] = \mathbf{m}^\intercal \bar{\boldsymbol{\gamma}}_j \tag{43}$$

For the linear coefficients $\mathbf{w}$, we propose a multivariate Gaussian distribution with mean parameter $\mathbf{m}$ and covariance parameter $\mathbf{S}$:

$$q(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) \tag{44}$$

$$\begin{aligned} \log q(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) &= \mathbb{E}_{q(\mathbf{z},\mathbf{g})}[\log p(\mathbf{g} \mid \bar{\mathbf{z}}, \mathbf{w})] + \log p(\mathbf{w} \mid \tau)] \\ &= \mathbf{w}^\intercal \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] \mathbb{E}_{q(\mathbf{g})}[\mathbf{g}] - \frac{1}{2} \mathbf{w}^\intercal (\tau \mathbf{I} + \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j^\intercal \bar{\mathbf{z}}_j]) \mathbf{w} \end{aligned} \tag{45}$$

where the expectation terms are calculated in Appendix B.5. The full-derivation of $\mathbf{w}^\intercal \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j^\intercal \bar{\mathbf{z}}_j]) \mathbf{w}$ in Eq. (53). We therefore obtain the following updates for $\mathbf{m}$ and $\mathbf{S}$:

$$\mathbf{m} = \mathbf{S} \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}] \mathbb{E}_{q(\mathbf{g})}[\mathbf{g}], \quad \mathbf{S} = (\tau \mathbf{I} + \mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}^\intercal \bar{\mathbf{z}}])^{-1} \tag{46}$$

where the expectation $\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j^\intercal \bar{\mathbf{z}}_j])$ is derived in Eq. (52).

After we assign variational distributions for $\mathbf{g}$ and $\mathbf{w}$, we can update γ_{ijk} using the supervised component of the MixEHR-S. Here, we take the expectation of the predictive likelihood $\log p(\mathbf{g} \mid \mathbf{w}, \mathbf{z}^{-(i,j)}, z_{ijk}=1, z_{ijk'}=0 \forall k' \neq k)]$ in Eq. (36):

$$
\begin{aligned}
&\mathbb{E}_{q(\mathbf{z}^{-(i,j)})}[\log p(\mathbf{g} \mid \mathbf{w}, \mathbf{z}^{-(i,j)}, z_{ijk}=1, z_{ijk'}=0 \forall k' \neq k)] \\
=&\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{g},\mathbf{w})}[\log\big(\frac{1}{\sqrt{2\pi}}\exp\big\{-\frac{(g_j-\mathbf{w}^\intercal\bar{\mathbf{z}}_j)^2}{2}\big\}\big)] \\
\propto&\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{g},\mathbf{w})}[\mathbf{w}^\intercal(\frac{1}{M_j}(\sum_{i'\neq i}\mathbf{z}_{i'j}+\mathbf{z}_{ij}))g_j-\frac{1}{2}\mathbf{w}^\intercal(\frac{1}{M_j^2}(\sum_{m\neq i}^{M_j}\mathbf{z}_{ij}\mathbf{z}_{mj}^\intercal+\sum_{n\neq i}^{M_j}\mathbf{z}_{nj}\mathbf{z}_{ij}^\intercal+\mathrm{diag}(\mathbf{z}_{ij}))\mathbf{w}] \\
\propto&\frac{\mathbf{w}^\intercal\mathbf{z}_{ij}\mathbb{E}_{q(\mathbf{g})}[g_j]}{M_j}-\frac{1}{2M_j^2}\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{w})}\left[\mathbf{w}^\intercal(\mathbf{z}_{ij}\underbrace{\sum_{m\neq i}^{M_j}\mathbf{z}_{mj}^\intercal}_{\mathbf{z}_{j/i}^\intercal}+\underbrace{\sum_{n\neq i}^{M_j}\mathbf{z}_{nj}}_{\mathbf{z}_{j/i}}\mathbf{z}_{ij}^\intercal+\mathrm{diag}(\mathbf{z}_{ij}))\mathbf{w}\right] \\
=&\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j}-\frac{1}{2M_j^2}\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{w})}\left[\mathbf{w}^\intercal(\mathbf{z}_{ij}\mathbf{z}_{j/i}^\intercal+\mathbf{z}_{j/i}\mathbf{z}_{ij}^\intercal+\mathrm{diag}(\mathbf{z}_{ij}))\mathbf{w}\right] \\
=&\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j}-\frac{1}{2M_j^2}\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{w})}\left[w_k\mathbf{z}_{j/i}^\intercal\mathbf{w}+\mathbf{w}^\intercal\mathbf{z}_{j/i}w_k+w_k^2\right] \\
=&\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j}-\frac{1}{2M_j^2}\mathbb{E}_{q(\mathbf{z}^{-(i,j)},\mathbf{w})}\left[2\sum_{m\neq i}\sum_{k'}w_kw_{k'}z_{mjk}+w_k^2\right] \\
=&\frac{m_k\mathbb{E}_{q(\mathbf{g})}[g_j]}{M_j}-\frac{1}{2M_j^2}\left[2\sum_{m\neq i}\sum_{k'}(m_km_{k'}+S_{kk'})\gamma_{mjk}+m_k^2+S_{kk}\right] \\
=&\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j}-\frac{1}{2M_j^2}\left[2(m_k\mathbf{m}^\intercal\boldsymbol{\gamma}_{j/i}+\mathrm{S}_k\boldsymbol{\gamma}_{j/i})+m_k^2+S_{kk}\right]
\end{aligned}
$$

where $\boldsymbol{\gamma}_{j/i}$ indicates the sum of all terms except for the ith ICD-9 code, i.e. $\boldsymbol{\gamma}_{j/i}=\sum_{m\neq i}^{M_j}\boldsymbol{\gamma}_{ij}$ and S_k indicates the kth row of the covariance matrix S. The expected value of $\mathbf{w}^\intercal\mathbf{w}$ is computed in Eq. (55).

We thus add the supervised component to $\boldsymbol{\gamma}$ (see Eq. (37) and Eq. (47)), obtaining full variational update for γ_{ijk}:

$$
\begin{aligned}
\gamma_{ijk}\propto&(\alpha_k+E_{q(z^{-(i,j)})}[n_{jk}^{-(i,j)}])\frac{\iota_{b_{ij}}+E_{q(z^{-(i,j)})}[m_{kb_{ij}}^{-(i,j)}]}{E_{q(z^{-(i,j)})}[m_{k.}^{-(i,j)}]+\sum_t\iota_t}\frac{\zeta_{kx_{ij}}+E_{q(z^{-(i,j)})}[p_{kb_{ij}x_{ij}}^{-(i,j)}]}{E_{q(z^{-(i,j)})}[p_{kb_{ij}.}^{-(i,j)}]+\sum_w\zeta_{kw}} \\
&\exp\left\{\frac{m_k\mathbb{E}_{q(g_j)}[g_j]}{M_j}-\frac{1}{2M_j^2}\left[2\left(m_k\mathbf{m}^\intercal\boldsymbol{\gamma}_{j/i}+\mathrm{S}_k\boldsymbol{\gamma}_{j/i}\right)+m_k^2+S_{kk}\right]\right\}
\end{aligned}
\tag{47}
$$

B.4 Hyperparameters update

We update the hyperparameters $(\alpha_k^*, \iota_t^*, \zeta_{wk}^*)$ by maximizing the marginal likelihood under the variational expectations via empirical Bayes fixed point iteration method [1, 19]:

$$
\alpha_k^*=\frac{c_\alpha-1+\alpha_k\sum_j\Psi(\alpha_k+n_{jk})-\Psi(\alpha_k)}{d_\alpha+\sum_j\Psi(\sum_k\alpha_k+n_{jk})-\Psi(\sum_k\alpha_k)} \tag{48}
$$

$$
\iota_t^*=\frac{c_\iota-1+\iota_t\sum_k\Psi(\iota_t+m_{tk})-\Psi(\iota_t)}{d_\iota+\sum_k\Psi(\sum_t\iota_t+m_{tk})-\Psi(\sum_t\iota_t)} \tag{49}
$$

$$
\zeta_{wk}^*=\frac{c_\zeta-1+\zeta_{wk}\sum_t\Psi(\zeta_{wk}+p_{twk})-\Psi(\zeta_{wk})}{d_\zeta+\sum_t\Psi(\sum_w\zeta_{wk}+p_{twk})-\Psi(\sum_w\zeta_{wk})} \tag{50}
$$

where $(c_\alpha, c_\iota, c_\zeta, d_\alpha, d_\iota d_\zeta)$ are constant values. For the experiment, we chose the following initial value setting $(1, 0.001, 2, 10, 0.01, 100)$.

B.5 Derivation of expectations of variational distributions

The expected value associated with the average of the topic assignments $\bar{\mathbf{z}}_j$ is:

$$\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j] = \bar{\boldsymbol{\gamma}}_j = \frac{1}{M_j}\sum_{i}^{M_j}\boldsymbol{\gamma}_{ij} \tag{51}$$

The expectation of $\bar{\mathbf{z}}_j\bar{\mathbf{z}}_j^\intercal$ is:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j\bar{\mathbf{z}}_j^\intercal] &= \mathbb{E}_{q(\mathbf{z})}\left[\begin{bmatrix}\frac{1}{M_j}\sum_i^{M_j} z_{ij1}\\ \vdots \\ \frac{1}{M_j}\sum_i^{M_j} z_{ijK}\end{bmatrix}\begin{bmatrix}\frac{1}{M_j}\sum_i^{M_j} z_{ij1} \cdots \frac{1}{M_j}\sum_i^{M_j} z_{ijK}\end{bmatrix}\right] \\
&= \mathbb{E}_{q(\mathbf{z})}\left[\begin{bmatrix}\frac{1}{M_j^2}\sum_i^{M_j} z_{ij1}^2 \cdots & \frac{1}{M_j^2}\sum_i^{M_j} z_{ij1}z_{ijK}\\ \vdots \ddots & \vdots \\ \frac{1}{M_j^2}\sum_i^{M_j} z_{ijK}z_{ij1} \cdots & \frac{1}{M_j^2}\sum_i^{M_j} z_{ijK}^2\end{bmatrix}\right] \\
&= \frac{1}{M_j^2}\Big(\sum_{i}^{M_j}\sum_{i'}^{M_j}\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}_{ij}]\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}_{i'j}^\intercal] + \sum_{i}^{M_j}\mathrm{diag}(\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}_{ij}])\Big) \\
&= \frac{1}{M_j^2}\Big(\sum_{i}^{M_j}\sum_{i'}^{M_j}\boldsymbol{\gamma}_{ij}\boldsymbol{\gamma}_{i'j}^\intercal + \sum_{i}^{M_j}\mathrm{diag}(\boldsymbol{\gamma}_{ij})\Big)
\end{aligned} \tag{52}$$

The expected value of $\mathbf{w}^\intercal\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j\bar{\mathbf{z}}_j^\intercal]\mathbf{w}$ given the variational distribution $q(\mathbf{w})$ is:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{w})}\left[\mathbf{w}^\intercal\mathbb{E}_{q(\mathbf{z})}[\bar{\mathbf{z}}_j\bar{\mathbf{z}}_j^\intercal]\mathbf{w}\right] &= \mathbb{E}_{q(\mathbf{w})}\left[\begin{bmatrix}w_1 \dots w_K\end{bmatrix}\begin{bmatrix}\frac{1}{M_j^2}\sum_i^{M_j}\gamma_{ij1}^2 \cdots & \frac{1}{M_j^2}\sum_i^{M_j}\gamma_{ij1}\gamma_{ijK}\\ \vdots \ddots & \vdots \\ \frac{1}{M_j^2}\sum_i^{M_j}\gamma_{ijK}\gamma_{ij1} \cdots & \frac{1}{M_j^2}\sum_i^{M_j}\gamma_{ijK}^2\end{bmatrix}\begin{bmatrix}w_1\\ \vdots \\ w_K\end{bmatrix}\right] \\
&= \mathbb{E}_{q(\mathbf{w})}\left[\begin{bmatrix}w_1 \dots w_K\end{bmatrix}\begin{bmatrix}\bar{\gamma}_{j1}^2 \cdots & \bar{\gamma}_{j1}\bar{\gamma}_{jK}\\ \vdots \ddots & \vdots \\ \bar{\gamma}_{jK}\bar{\gamma}_{j1} \cdots & \bar{\gamma}_{jK}^2\end{bmatrix}\begin{bmatrix}w_1\\ \vdots \\ w_K\end{bmatrix}\right] \\
&= \mathbb{E}_{q(\mathbf{w})}\left[\begin{bmatrix}\mathbf{w}^\intercal\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{j1} \dots \mathbf{w}^\intercal\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{jK}\end{bmatrix}\begin{bmatrix}w_1\\ \vdots \\ w_K\end{bmatrix}\right] \\
&= \mathbb{E}_{q(\mathbf{w})}\Big[\sum_{k'} w_{k'}\mathbf{w}^\intercal\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{jk'}\Big] \\
&= \sum_{k'}\mathbb{E}_{q(\mathbf{w})}\left[w_{k'}\mathbf{w}^\intercal\right]\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{jk'} \\
&= \sum_{k'}\begin{bmatrix}\mathbb{E}_{q(\mathbf{w})}[w_{k'}w_1]\\ \vdots \\ \mathbb{E}_{q(\mathbf{w})}[w_{k'}w_K]\end{bmatrix}^\intercal\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{jk'} \\
&= \sum_{k'}\begin{bmatrix}m_{k'}m_1 + S_{k'1}\\ \vdots \\ m_{k'}m_K + S_{k'K}\end{bmatrix}^\intercal\bar{\boldsymbol{\gamma}}_{j\cdot}\bar{\gamma}_{jk'}
\end{aligned} \tag{53}$$

where the $\bar{\boldsymbol{\gamma}}_{j\cdot} = \frac{1}{M_j}\sum_i\boldsymbol{\gamma}_j$ is a K-dimensional vector, each scalar value $\bar{\gamma}_{jk}$ represents the average value of γ over ICD-9 codes for the kth topic and the jth patient.

The expected value of latent liability variable $\mathbf{g}$ for each patient j is given:

$$\mathbb{E}_{q(\mathbf{g})}[g_j] = \begin{cases} \lambda_j + \phi_j/(1-\Phi_j), & \text{if} \quad y_j = 1. \\ \lambda_j - \phi_j/\Phi_j, & \text{if} \quad y_j = 0. \end{cases} \tag{54}$$

where $\phi_j = \phi(-\lambda_j)$ is the normal density and $\Phi_j = \Phi(-\lambda_j)$ is the cumulative distribution function (CDF) of the standard normal distribution. The corresponding expected value for $\mathbf{w}$ is:

$$\mathbb{E}_{q(\mathbf{w})}[\mathbf{w}] = \mathbf{m}, \quad \mathbb{E}_{q(\mathbf{w})}[\mathbf{w}^\intercal \mathbf{w}] = \mathbb{E}_{q(\mathbf{w})}[\sum_{k=1}^{K} w_k w_k] = \sum_{k=1}^{K} m_k^2 + S_{kk} \tag{55}$$

B.6 Derivation of estimates for variational expectations of mixing proportions

The approximations for mixing proportions ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$) can be calculated as follows [27]:

$$\hat{\boldsymbol{\theta}} = \frac{\alpha_k + \mathbb{E}_{q(\mathbf{z})}[n_{j.k}]}{\sum_{k'} \alpha_{k'} + \mathbb{E}_{q(\mathbf{z})}[n_{j.k'}]} \tag{56}$$

$$\hat{\boldsymbol{\beta}} = \frac{\iota_t + \mathbb{E}_{q(\mathbf{z})}[m_{.kt}]}{\sum_{t'} \iota_{t'} + \mathbb{E}_{q(\mathbf{z})}[m_{.kt'}]} \tag{57}$$

$$\hat{\boldsymbol{\eta}} = \frac{\zeta_w + \mathbb{E}_{q(\mathbf{z})}[p_{.ktw}]}{\sum_{w'} \zeta_{w'} + \mathbb{E}_{q(\mathbf{z})}[p_{.ktw'}]} \tag{58}$$

where the expected values of sufficient statistics ($n_{j.k}, m_{.kt}, p_{.ktw}$) are:

$$\mathbb{E}_{q(\mathbf{z})}[n_{j.k}] = \sum_{i}^{M_j} [z_{ij} = k] \tag{59}$$

$$\mathbb{E}_{q(\mathbf{z})}[m_{.kt}] = \sum_{j'}^{D} \sum_{i}^{M_{j'}} [z_{ij'} = k, b_{ij'} = t] \tag{60}$$

$$\mathbb{E}_{q(\mathbf{z})}[p_{.ktw}] = \sum_{j'}^{D} \sum_{i}^{M_{j'}} [z_{ij'} = k, b_{ij'} = t, x_{ij'} = w] \tag{61}$$

B.7 Derivation of predictive distribution

Here, we show that the predictive distribution is a Bernoulli distribution when using a Gaussian response since the natural parameter $\mathbf{w}^\intercal \bar{\mathbf{z}}$ is identical to the mean parameter where $\bar{\mathbf{z}}\star$ and $\mathbf{y}\star$ represent the new data points and the predicted label respectively. The full derivation of

the predictive distribution is:

$$
\begin{aligned}
p(\mathbf{y}\star \mid \bar{\mathbf{z}}\star, \mathbf{y}, \bar{\mathbf{z}}) &= \int_{-\infty}^{\infty} \int p(\mathbf{y}\star, \mathbf{g}, \mathbf{w} \mid \bar{\mathbf{z}}\star, \mathbf{y}, \bar{\mathbf{z}}) d\mathbf{w} d\mathbf{g} \\
&= \int_{-\infty}^{\infty} \int p(\mathbf{y}\star \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{w}, \bar{\mathbf{z}}\star) p(\mathbf{w} \mid \mathbf{y}, \bar{\mathbf{z}}) d\mathbf{w} d\mathbf{g} \\
&\approx \int_{-\infty}^{\infty} \int p(\mathbf{y}\star \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{w}, \bar{\mathbf{z}}\star) q(\mathbf{w}) d\mathbf{w} d\mathbf{g} \\
&= \int_{-\infty}^{\infty} \int \mathbb{1}(\mathbf{g} > 0)^{\mathbf{y}\star} \mathbb{1}_{(}(\mathbf{g} \leq 0)^{1-\mathbf{y}\star} \mathcal{N}(\mathbf{g} \mid \mathbf{w}^{\intercal} \bar{\mathbf{z}}\star, 1) \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) d\mathbf{w} d\mathbf{g} \\
&= \begin{cases} \int_{0}^{\infty} \mathcal{N}(\mathbf{g} \mid \mathbf{m}^{\intercal} \bar{\mathbf{z}}\star, 1 + \bar{\mathbf{z}} \star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star) d\mathbf{g} & \text{if} \quad \mathbf{y}\star = 1 \\ \int_{-\infty}^{0} \mathcal{N}(\mathbf{g} \mid \mathbf{m}^{\intercal} \bar{\mathbf{z}}\star, 1 + \bar{\mathbf{z}} \star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star) d\mathbf{g} & \text{if} \quad \mathbf{y}\star = 0 \end{cases} \\
&= \begin{cases} 1 - \Phi\left(\frac{-\mathbf{m}^{\intercal} \bar{\mathbf{z}}\star}{(1 + \bar{\mathbf{z}}\star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star)^{\frac{1}{2}}} \right) & \text{if} \quad \mathbf{y}\star = 1 \\ \Phi\left(\frac{-\mathbf{m}^{\intercal} \bar{\mathbf{z}}\star}{(1 + \bar{\mathbf{z}}\star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star)^{\frac{1}{2}}} \right) & \text{if} \quad \mathbf{y}\star = 0 \end{cases} \\
&= \Phi\left(\frac{\mathbf{m}^{\intercal} \bar{\mathbf{z}}\star}{(1 + \bar{\mathbf{z}} \star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star)^{\frac{1}{2}}} \right)^{\mathbf{y}\star} \left(1 - \Phi\left(\frac{\mathbf{m}^{\intercal} \bar{\mathbf{z}}\star}{(1 + \bar{\mathbf{z}} \star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star)^{\frac{1}{2}}} \right) \right)^{1-\mathbf{y}\star} \\
&= \text{Bernoulli}\left(\mathbf{y}\star \mid \Phi\left(\frac{\mathbf{m}^{\intercal} \bar{\mathbf{z}}\star}{(1 + \bar{\mathbf{z}} \star^{\intercal} \mathbf{S} \bar{\mathbf{z}}\star)^{\frac{1}{2}}} \right) \right)
\end{aligned} \tag{62}
$$