Automatic Phenotyping by a Seed-guided Topic Model

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

Yuanyi Hu
Department of Mathematics and Statistics, McGill University
Montreal, QC, Canada

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

David L. Buckeridge
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

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

ABSTRACT

Electronic health records (EHRs) provide rich clinical information and the opportunities to extract epidemiological patterns to understand and predict patient disease risks with suitable machine learning methods such as topic models. However, existing topic models do not generate identifiable topics each predicting a unique phenotype. One promising direction is to use known phenotype concepts to guide topic inference. We present a seed-guided Bayesian topic model called MixEHR-Seed with 3 contributions: (1) for each phenotype, we infer a dual-form of topic distribution: a seed-topic distribution over a small set of key EHR codes and a regular topic distribution over the entire EHR vocabulary; (2) we model age-dependent disease progression as Markovian dynamic topic priors; (3) we infer seed-guided multi-modal topics over distinct EHR data types. For inference, we developed a variational inference algorithm. Using MixEHR-Seed, we inferred 1569 PheCode-guided phenotype topics from an EHR database in Quebec, Canada covering 1.3 million patients for up to 20-year follow-up with 122 million records for 8539 and 1126 unique diagnostic and drug codes, respectively. We observed (1) accurate phenotype prediction by the guided topics, (2) clinically relevant PheCode-guided disease topics, (3) meaningful age-dependent disease prevalence. Source code is available at GitHub.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Mathematics of computing** → *Probability and statistics*; • **Information systems** → Information retrieval; • **Applied computing** → Life and medical sciences.

KEYWORDS

topic modeling, variational autoencoder, electronic health records, predictive healthcare

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542675

ACM Reference Format:
Ziyang Song, Yuanyi Hu, Aman Verma, David L. Buckeridge, and Yue Li. 2022. Automatic Phenotyping by a Seed-guided Topic Model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3542675

1 INTRODUCTION

Accelerating adoption of electronic health records (EHRs) has provided promising opportunities in medical informatics and clinical reasoning research [15, 21]. EHRs are rich, longitudinal, and heterogeneous collections of patient health information often consisting of temporally ordered visits, in which multiple types of clinical observations are made by healthcare practitioners. EHRs systematically record complementary description of patient health status, which include, among many others, demographic information, International Classification of Diseases (ICD) 9 codes, and Anatomical Therapeutic Chemical (ATC) drug prescription codes [9]. Distilling from these multi-modal and longitudinal information into interpretable clinical concepts to help summarize patients' health is an active research area in applied machine learning [27].

Machine learning approaches offer great potential in modeling EHR data with the goal of automatic phenotyping for any given patient based on their EHR data [27]. However, most applied ML methods fall into one of the two categories. The first category of methods focus on supervised learning by training a model to predict one or a small number of target diseases [19, 20]. These methods do not scale well to the inference of hundreds or thousands of phenotypes simultaneously without running into computational bottlenecks or model overfitting. The second category of methods include unsupervised approaches, which model latent lower-dimensional factors to explain the high-dimensional EHR data [7, 18, 26]. However, these unsupervised methods are often learned separately from supervised tasks such as disease diagnostic predictions.

As a well-known family of probabilistic models in the second category, topic models were shown to be an effective tool for text mining and document retrieval in natural language processing (NLP) community [4]. Topic models learn word co-occurrence patterns from high-dimensional text to discover latent topics as a set of multinomial distributions over the vocabulary. In their applications to EHR data, each patient's medical history and its associated data elements (i.e. ICD diagnostic codes) are treated as a document and word tokens, respectively. Topic models are then used to to infer a

set of latent disease topics and the topic mixture memberships of the patients from the EHR data. This class of methods exploits the sparsity of the discrete feature dimensions and are highly scalable to modeling large-scale patient EHR data as its entirety [7, 18, 26]. However, because of the lack of supervision, the inferred latent topics are not identifiable and require expensive human manual curation to interpret them especially when the number of topics is large (e.g., over 1000 topics).

In this paper, we present a hierarchical Bayesian seed-guided topic model called MixEHR-Seed. Here the "seed" is referred to as a set of diagnostic codes that are associated with a given phenotype. In our EHR applications, we make use of over 1500 expert-curated PheCodes available from the Phenome-Wide Association Studies (PheWAS) catalog [9, 10]. Each PheCode is defined by a set of ICD-9 codes, which we treat as the seeds for the corresponding phenotype. Using a large EHR database consisting of over 1.3 million individuals in Quebec, Canada for up to 20 year follow-up history, we demonstrate MixEHR-Seed as a promising approach for simultaneous automatic phenotyping and learning meaningful disease representations over a large number of diverse phenotypes.

2 OUR CONTRIBUTIONS

Our proposed model stands out from the existing methods with three key contributions.

First, our approach can leverage the PheCodes to infer over 1500 identifiable phenotype topics, each of which can be used to predict the risk of a unique disease for any given patient.

Second, we develop a unified modeling framework to model multiple types of EHR information as multi-modality so that other modalities can benefit from the seed-guidance from a single modality without the need of having the "seed" in every modality.

Third, our approach can infer age-dependent topic disease progression from the longitudinal EHR data by imposing a Markovian distribution over the age-specific topic hyperparameters. Therefore, for any given patient with a certain age, our model will learn to associate them with appropriately defined age-dependent Bayesian prior.

To efficiently approximate the posterior distribution of our proposed model, we develop a Bayesian inference algorithm combining collapsed mean-field variational inference [25] and variational autoencoder (VAE) model [17]. Specifically, our hybrid inference algorithm exploits the probabilistic dependencies between conjugated latent variables to efficiently infer the over 1500 phenotypic topic assignments for each EHR record; it also exploits the flexibility of the long short-term memory (LSTM) to infer the non-linear age-dependent topic dynamics using a recurrent neural network.

3 RELATED WORKS

Latent Dirichlet allocation (LDA) [4] models document collections to identify the underlying topics in a fully unsupervised manner. Early applications of LDA in EHR show its potential but also limited capacity to model complex EHR data [7, 26].

Supervised LDA (S-LDA) improved over LDA by incorporating generalized linear model into topic inference [6]. MixEHR-S extends S-LDA to learn multi-specialist phenotypic topics and predict a binary label simultaneously [24]. However, neither model scales well to multi-target prediction task.

Developed in the domain of public health surveillance, EpiNews is a seed-guided topic model that infers topics and temporal incidence trends from news reports [12]. However, EpiNews does not model time-series data as sequentially dependent events but rather treats the time points as tokenized features.

SureLDA [1] and MixEHR-G [2] compute initial topic probabilities from a 2-component Gaussian mixture model (GMM) on each PheCode and then used the GMM-inferred posteriors as the topic hyperparameters to guide topic inference. While both models are capable of simultaneously annotating many diverse phenotypes, their formulation differs from the proposed seed-guided topic inference and do not account for dynamic topic evolution.

UPhenome [22] and MixEHR [18] jointly learn multi-modal EHR data with distinct topic distributions. Both are trained in an unsupervised way so that the topics are not directly identifiable. Moreover, they do not model temporal patterns from the longitudinal EHR data.

Dynamic topic model (DTM) infers temporal evolution of the underlying topics using Kalman filter in the variational approximation [5]. Dynamic embedded topic model (DETM) extends DTM by learning topic embedding with VAE-based amortized inference algorithm [11]. Neither DTM nor DETM perform guided topic inference to generate identifiable topics.

4 METHODOLOGY

We first define the generative process of MixEHR-Seed in Section 4.1. We then describe an efficient hybrid Bayesian inference algorithm (Section 4.2). The main notations and complete derivations are described in **Appendix Section A.2**.

4.1 MixEHR-Seed: a seed-guided dynamic topic model

Following the topic modeling convention, we considered each patient's records as a document indexed by $d \in \{1, \ldots, D\}$. For a document of size N_d, we index its word token (i.e., EHR code) by $i \in \{1, \ldots, N_d\}$. We index the age group of a patient by $t \in \{1, \ldots, T\}$. MixEHR-Seed assumes the following data generative process (Figure 1):

(1) For age group $t = 1$, draw initial topic hyperparameter $\boldsymbol{\eta}_1 \sim \mathcal{N}(0, \delta^2 I)$
(2) For each age group $t \in \{2, \ldots, T\}$:
   - Draw age-dependent topic hyperparameter $\boldsymbol{\eta}_t \mid \boldsymbol{\eta}_{t-1} \sim \mathcal{N}(\boldsymbol{\eta}_{t-1}, \delta^2 I)$, $\boldsymbol{\alpha}_t = \text{Softplus}(\boldsymbol{\eta}_t)$
(3) For each phenotype topic $k = 1, \ldots, K$:
   (a) Draw regular topic $\boldsymbol{\phi}_k^r \sim \text{Dir}(\beta)$ over the entire vocabulary $\mathcal{V}$
   (b) Draw seed topic $\boldsymbol{\phi}_k^s \sim \text{Dir}(\mu)$ over only the seed set $\mathcal{V}_k$
   (c) Draw seed-topic rate $\pi_k \sim \text{Beta}(1, 1)$
(4) For each EHR document $d = 1, \ldots, D$:
   (a) Draw phenotype topic proportion $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}_{t_d})$, where t_d is the age group index of document d
   (b) For each EHR code $i = 1, \ldots, N_d$:
      (i) Draw topic assignment $z_{di} \sim \text{Mult}(\boldsymbol{\theta}_d)$
      (ii) Draw seed-topic indicator $x_{di} \sim \text{Bern}(\pi_{z_{di}})$.

Table 1: Notations in MixEHR-Seed

Notations	Descriptions
D	number of documents in the dataset
N_d	number of words in the document d
K	number of topics in the dataset
$\mathcal{V}$	the vocabulary in the dataset
$\mathcal{V}_k$	a set of seed words for topic k
w_{di}	word index of token i in document d
z_{di}	topic assignment for word w_{di}
x_{di}	binary seed-topic indicator of word w_{di}
$\boldsymbol{\eta} \in R^{T\times K}$	age-dependent topic hyperparameters
$\boldsymbol{\alpha} \in R^{T\times K}$	Softplus-transformed $\boldsymbol{\eta}$
$\boldsymbol{\theta} \in R^{D\times K}$	topic mixture memberships
$\boldsymbol{\phi}^s \in R^{K\times S}$	seed topic distributions
$\boldsymbol{\phi}^r \in R^{K\times V}$	regular distributions
$\boldsymbol{\pi} \in R^{K}$	seed-topic rates for a word

(iii) Draw a EHR code:

$$w_{di} \sim \begin{cases} \text{Mult}(\boldsymbol{\phi}^s_{z_{di}}), \text{if } x_{di} = 1 \\ \text{Mult}(\boldsymbol{\phi}^r_{z_{di}}), \text{otherwise} \end{cases}$$

where Dir, Multi, Bern abbreviate Dirichlet, Multinomial, and Bernoulli distributions, respectively. While the above data generative process is self-explanatory, we provide their rationales as follows.

First, each phenotype topic k is characterized by two types of distributions: the seed topic distribution $\boldsymbol{\phi}^s_k$ and the regular topic distribution $\boldsymbol{\phi}^r_k$ [16]. The regular topics govern the global phenotype distributions, and one can sample from it any EHR code in the EHR vocabulary $\mathcal{V}$; the seed topics capture the phenotype-specific information and thus are constrained to have non-zero probabilities over only the seed EHR codes in $\mathcal{V}_k$. For example, the seed codes for Asthma according to the PheCode definitions are 493[0,1,2,8,9] corresponding to extrinsic, intrinsic, chronic obstructive, other forms, and unspecified, respectively [9, 10].

Second, we express statistical uncertainty via the the Bernoulli variable x_{di} to indicate whether an EHR code w_{di} is generated from a seed topic ($x_{di} = 1$) or a regular topic ($x_{di} = 0$). The seed-topic rate π_k dictates the sampling probability that a word is drawn from the seed topic instead of the regular topic.

Third, to infer temporal disease progression in an age-dependent way, we assume that the disease topic mixture $\boldsymbol{\theta}_d$ follows a K-dimensional Dirichlet prior with age-specific topic hyperparameter $\boldsymbol{\alpha}_{t_d}$. We impose a Markovian distribution over the dynamic Gaussian variables $\boldsymbol{\eta}_{1:T}$, which is then transformed to non-negative values $\boldsymbol{\alpha}$ via a softplus activation as the Dirichlet hyperparameters.

After the training (Section 4.2), MixEHR-Seed generates the following useful information for downstream analysis: (1) identifiable phenotype topic mixture $\boldsymbol{\theta}_d$ for each document d that can be used to simultaneously infer the risk of K phenotypes of each patient for automatic phenotyping; (2) regular phenotype topic distribution $\boldsymbol{\phi}^r_k$ for discovering disease co-morbidities w.r.t. each phenotype k over the entire EHR vocabulary; (3) temporal disease trend $\boldsymbol{\alpha}_k$ to investigate age-dependent phenotype prevalence.

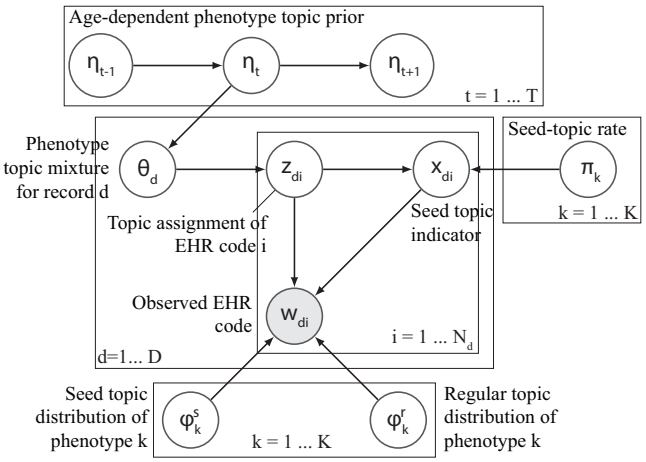

Figure 1: The probabilistic graphical model of MixEHR-Seed. The shaded node represents the observed variable and the unshaded nodes represent the latent variables. For each EHR document d, its topic mixture θ_d is governed by an age-dependent topic prior η_{t_d}. The topic assignment z_{di} of each EHR code follows a multinomial distribution with rate set to be the topic mixture θ_d. Each EHR code w_{di} for code i in EHR document d is sampled from a mixture distribution: $x_{di}\boldsymbol{\phi}^s_{z_{di}} + (1 - x_{di})\boldsymbol{\phi}^r_{z_{di}}$, where $\boldsymbol{\phi}^r_k$ endows the global regular topic distribution and $\boldsymbol{\phi}^s_k$ the seed topic distribution for phenotype k. The notations are listed in Table 1.

4.2 Inference

4.2.1 Collapsed variational inference. To reduce probabilistic dependencies among latent variables, we integrated out the latent variables $\boldsymbol{\theta}$ by exploiting the conditional independency in the PGM (Figure 1) and the conjugate relationship between Dirichlet variables $\boldsymbol{\theta}$ and the multinomial topic assignment variables $\mathbf{z}$. Similarly, we integrated out Dirichlet variables $\boldsymbol{\phi}^r$ and $\boldsymbol{\phi}^s$ due to its conjugacy to the multinomial EHR code variables $\mathbf{w}$ and the PGM structure. Note that we can always compute the sufficient statistics of these Dirichlet variables given the expectations of the multinomial variables and the hyperparameters (see Eq. (15)). Therefore, we focus on inferring the distributions of the seed-guided topic assignments $\mathbf{z}$ and the age-dependent topic hyperparameters $\boldsymbol{\eta}$. For the K-dimensional hyperparameters $\boldsymbol{\pi}$ over the K topics, we used empirical Bayes to optimize their fixed point estimates.

To infer the remaining latent variables $\mathbf{z}$ and $\boldsymbol{\eta}$, we used variational inference by turning our inference problem into an optimization problem over an evidence lower bound (ELBO) of marginal likelihood:

$$L_{ELBO} = \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta} \mid \beta, \mu, \boldsymbol{\pi})] - \mathbb{E}_q[\log q(\mathbf{z}, \boldsymbol{\eta})] \quad (1)$$

Under the independent mean-field assumption, we propose a fully factorized variational family for the latent variables $\mathbf{z}$ and $\boldsymbol{\eta}$:

$$q(\mathbf{z}, \boldsymbol{\eta}) = \prod_d^D \prod_i^{N_d} q(z_{di} \mid \boldsymbol{\gamma}_{di}) \prod_t^T q(\boldsymbol{\eta}_t \mid \boldsymbol{\eta}_{1:t-1}) \quad (2)$$

where the variational densities of $\mathbf{z}$ and $\boldsymbol{\eta}$ are multinomial and Gaussian, respectively. We employ collapsed variational inference and amortized variational inference to optimize the variational parameters of $q(\mathbf{z})$ and $q(\boldsymbol{\eta})$, respectively.

4.2.2 Inferring seed-guided topic assignments. Without seed words, the variational parameter of the topic assignment γ_{dik} for code i in EHR document d under topic k is:

$$\gamma_{dik} \propto \frac{\exp(\mathbb{E}_{q(\mathbf{z}^{-di})}[\log p(\mathbf{w}, \mathbf{z})])}{\sum_{k=1}^{K} \exp(\mathbb{E}_{q(\mathbf{z}^{-di})}[\log p(\mathbf{w}, \mathbf{z})])} \tag{3}$$

where the notation $-di$ means the exclusion of token i in record d.

With the set of seed words for phenotype topic k, we consider four scenarios when inferring the topic assignment of token i in document d:

(1) If w_{di} is a seed word under topic k, the posterior probability it is sampled from the *seed* topic distribution of topic k is:

$$\gamma^{ss}_{dik} \propto (\mathbb{E}_q[m^{-di}_{dk}] + \mathbb{E}_q[\alpha_{t_d k}]) \cdot \frac{\mathbb{E}_q[s^{-di}_{wk}] + \mu}{\mathbb{E}_q[s^{-di}_{.k}] + V_k \mu} \cdot \pi_k \tag{4}$$

(2) If w_{di} is a seed word under topic k, the posterior probability it is sampled from the *regular* topic distribution of topic k is:

$$\gamma^{sr}_{dik} \propto (\mathbb{E}_q[m^{-di}_{dk}] + \mathbb{E}_q[\alpha_{t_d k}]) \cdot \frac{\mathbb{E}_q[n^{-di}_{wk}] + \beta}{\mathbb{E}_q[n^{-di}_{.k}] + V\beta} \cdot (1 - \pi_k) \tag{5}$$

(3) If w_{di} is a regular word w.r.t. topic k, the posterior probability it is sampled from the *seed* topic distribution of topic k is 0.

(4) If w_{di} is a regular word w.r.t. topic k, the posterior probability it is sampled from the *regular* topic distribution of topic k is:

$$\gamma^{rr}_{dik} \propto (\mathbb{E}_q[m^{-di}_{dk}] + \mathbb{E}_q[\alpha_{t_d k}]) \cdot \frac{\mathbb{E}_q[n^{-di}_{wk}] + \beta}{\mathbb{E}_q[n^{-di}_{.k}] + V\beta} \tag{6}$$

Here n_{wk} denotes the number of times word w is assigned to regular topic k, $n_{.k}$ is the total number of words assigned to topic k across all documents, s_{wk} is the number of times the seed word w is assigned to seed topic k, and $s_{.k}$ is the total number of times of all seed words assigned to seed topic k. Their expected values of sufficient statistics are computed as follows:

$$\begin{aligned}
\mathbb{E}_q[n^{-di}_{wk}] &= \sum_{d' \neq d}^{D} \sum_{i}^{N_{d'}} [w_{d'i} = w](\gamma^{rr}_{d'ik} + \gamma^{sr}_{d'ik}) \\
\mathbb{E}_q[n^{-di}_{.k}] &= \sum_{d' \neq d}^{D} \sum_{i}^{N_{d'}} \gamma^{rr}_{d'ik} + \gamma^{sr}_{d'ik} \\
\mathbb{E}_q[s^{-di}_{wk}] &= \sum_{d' \neq d}^{D} \sum_{i}^{N_{d'}} [w_{d'i} = w]\gamma^{ss}_{d'ik} \\
\mathbb{E}_q[s^{-di}_{.k}] &= \sum_{d' \neq d}^{D} \sum_{i}^{N_{d'}} \gamma^{ss}_{d'ik}
\end{aligned} \tag{7}$$

And the expected sufficient statistics with respect to topic mixture membership of document d are computed as follows:

$$\mathbb{E}_q[m^{-di}_{dk}] = \sum_{i' \neq i}^{N_d} \pi_k \gamma^{ss}_{di'k} + (1 - \pi_k)(\gamma^{sr}_{di'k} + \gamma^{rr}_{di'k}) \tag{8}$$

Note that the inference of regular topics by Eq. (6) benefits from guided information through $\mathbb{E}_q[m_{dk}]$, whose calculation in Eq. (8) relies on both seed topics and regular topics. This inter-dependent relationship is the main driver of our guided mechanism. Finally, We normalize the topic assignments as proper probabilities such that $\sum_k^K \gamma^{ss}_{dik} + \gamma^{sr}_{dik} = 1$ and $\sum_k^K \gamma^{rr}_{dik} = 1$ for normalization: $\gamma^{ss}_{dik} = \frac{\gamma^{ss}_{dik}}{\sum_k^K \gamma^{ss}_{dik} + \gamma^{sr}_{dik}}$, $\gamma^{sr}_{dik} = \frac{\gamma^{sr}_{dik}}{\sum_k^K \gamma^{ss}_{dik} + \gamma^{sr}_{dik}}$, $\gamma^{rr}_{dik} = \frac{\gamma^{rr}_{dik}}{\sum_k^K \gamma^{rr}_{dik}}$.

The seed-topic rates $\boldsymbol{\pi}$ are estimated by maximizing the marginal likelihood function under the variational expectations:

$$\boldsymbol{\pi} = \frac{\sum_d^D \sum_i^{N_d} \boldsymbol{\gamma}^{ss}_{di}}{\sum_d^D \sum_i^{N_d} \boldsymbol{\gamma}^{ss}_{di} + \boldsymbol{\gamma}^{sr}_{di}} \tag{9}$$

4.2.3 Multi-modal extension. To extend to multi-modality, we incorporate the diverse types of EHR codes to form a unified modeling framework. Given the modality m without seed codes, we update modality-specific variational parameter $\gamma^{rr(m)}_{dik}$ as below:

$$\gamma^{rr(m)}_{dik} \propto (\mathbb{E}_q[m^{-di}_{dk}] + \mathbb{E}_q[\alpha_{t_d k}]) \cdot \frac{\mathbb{E}_q[n^{(m)-di}_{wk}] + \beta}{\mathbb{E}_q[n^{(m)-di}_{.k}] + \mathcal{V}^{(m)}\beta} \tag{10}$$

Therefore, the variational update relies on *document-level* sufficient statistics $\mathbb{E}_q[m_{dk}]$ and the *modality-specific* sufficient statistics $\mathbb{E}_q[n^{(m)}_{wk}] = \sum_d^D \sum_i^{N_d} [w_{di} = w]\gamma^{rr(m)}_{dik}$ and $\mathbb{E}_q[n^{(m)}_{.k}] = \sum_w^V \mathbb{E}_q[n^{(m)}_{wk}]$. Therefore, the inference of regular topics for the modality without the seed words in Eq. (10) also benefits from seed-guided information through the document-level sufficient statistics computed over each modality m that does not contain seed codes:

$$\mathbb{E}_{q(z)}[m^{(m)}_{dk}] = \sum_{i}^{N_d} \gamma^{rr(m)}_{dik} \quad \mathbb{E}_{q(z)}[m_{dk}] = \sum_{m}^{M} E_{q(z)}[m^{(m)}_{dk}] \tag{11}$$

4.2.4 Inferring age-dependent topic hyperparameters. To infer age-dependent dynamic topic hyperparameters $\boldsymbol{\eta}$, we employed an amortized variational inference framework that makes use of LSTM [13] to capture the temporal dependency from younger age groups to older age groups [11]. The LSTM takes as input all preceding variables $\boldsymbol{\eta}_{1:t-1}$ and the normalized bag-of-words representation $\tilde{\mathbf{w}}_t$ averaged over all documents from patients at age group t. It then outputs the mean and variance of the proposed Gaussian distribution $q(\boldsymbol{\eta}_t)$:

$$q(\boldsymbol{\eta}_t \mid \boldsymbol{\eta}_{1:t-1}, \tilde{\mathbf{w}}_t) : \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2 I)$$

$$[\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t] = \text{LSTM}(\boldsymbol{\eta}_{1:t-1}, \tilde{\mathbf{w}}_t) \tag{12}$$

$$\hat{\boldsymbol{\eta}}_t \sim \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \epsilon; \quad \epsilon \sim \mathcal{N}(0, I) \tag{13}$$

Here we apply the reparameterization trick to sample $\boldsymbol{\eta}_t$ by first sampling a Gaussian noise $\epsilon \sim N(0, I)$, and then computing a valid sampled Gaussian value $\hat{\boldsymbol{\eta}}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \epsilon$ [17]. We can then approximate the variational expectations of the temporal topic hyperparameters $\mathbb{E}_q[\boldsymbol{\alpha}]$ using the sampled Gaussian value $\hat{\boldsymbol{\eta}}_t$ via a Softplus activation:

$$\mathbb{E}[\hat{\boldsymbol{\alpha}}_t] \approx \text{Softplus}(\hat{\boldsymbol{\eta}}_t) = \ln(1 + \exp(\hat{\boldsymbol{\eta}}_t)) \tag{14}$$

We used a LSTM with 3 connected layers and 200 hidden units each layer. To optimize the LSTM parameters, we carried out gradient descent by Adam optimizer with a set learning rate 0.0005.

Algorithm 1 Inference algorithm of MixEHR-Seed

Input: a collection of medical documents $\mathbf{w}$
Initialize the expected sufficient statistics $\mathbb{E}_q[n]$, $\mathbb{E}_q[s]$, $\mathbb{E}_q[m]$ by Eq. (23) (**Appendix**).
repeat
  for each mini-batch of 1000 EHR docs **do**
    % *E-step*:
    Infer age-dependent topic dynamics $\boldsymbol{\eta}$ by Eq. (12)
    Compute variational expectations of $\boldsymbol{\alpha}$ by Eq. (14)
    Infer topic assignments $\mathbf{z}$ by Eq. (4), (5), (6), (10)
    Compute topic sufficient statistics by Eq. (7), (8)
    % *M-step*:
    Estimate hyperparameters $\boldsymbol{\pi}$ by Eq. (9)
    Maximize $L_{ELBO}(\boldsymbol{\eta})$ w.r.t. the LSTM parameters
  end for
  Evaluate L_{ELBO}
until L_{ELBO} has converged

4.2.5 Summary of the inference algorithm. The complete inference algorithm is summarized in Algorithm 1. To handle large-scale data, for the inference of topic assignments $\mathbf{z}$ and age-dependent topic hyperparameters $\boldsymbol{\eta}$, we perform stochastic variational inference using mini-batches of 1,000 EHR documents per batch [11, 14, 17]. We used the validation set to fine-tune the topic hyperparameters μ and β to minimize the held-out negative log-likelihood. Upon convergence of the ELBO, we can compute the collapsed variables $(\boldsymbol{\theta}, \boldsymbol{\phi}^r, \boldsymbol{\phi}^s)$ with the respective variational expectations:

$$\mathbb{E}_q[\theta_{dk}] = \frac{\mathbb{E}_q[m_{dk}] + \mathbb{E}_q[\alpha_{t_d k}]}{\mathbb{E}_q[m_{d.}] + \sum_{k=1}^{K} \mathbb{E}_q[\alpha_{t_d k}]}$$

$$\mathbb{E}_q[\phi^r_{wk}] = \frac{\mathbb{E}_q[n_{wk}] + \beta}{\mathbb{E}_q[n_{.k}] + \beta V}, \quad \mathbb{E}_q[\phi^s_{wk}] = \frac{\mathbb{E}_q[s_{wk}] + \mu}{\mathbb{E}_q[s_{.k}] + \mu V_k} \tag{15}$$

5 EXPERIMENTS

5.1 PopHR data

The Population Health Record (PopHR) was created for monitoring population health in Quebec, Canada [23, 28]. The database hosts a massive amount of longitudinal heterogeneous claim data from the provincial government health insurer in Quebec, Canada (Régie de l'assurance maladie du Québec, RAMQ) on health service use. In total, there are approximately 1.3 million participants in the PopHR database, which represent a randomly sampled 25% of the population in the metropolitan area of Montreal between 1998 and 2014. Cohort membership is maintained dynamically by removing deceased residents and actively enrolling newborns and immigrants.

For evaluating our model, we used the entire cohort of over 1.3 million individuals along with more than 122 million healthcare records. For the clinical features, we used the ICD-9 codes and the ATC drug codes from the PopHR database. Among these EHR documents, there are 8,539 unique ICD-9 codes and 1,126 unique ATC codes. To perform age-dependent topic analysis, we binned each patient's medical history into a set of temporally organized documents spanning five years of age; the derived dataset consists of approximately 3.9 million clinical documents across 19 age groups.

The PopHR database includes rule-based labels for 11 phenotypes (PheCodes): Attention deficit hyperactivity disorder (ADHD) (313.1), Asthma (495), Autism (313.3), Congestive heart failure (CHF) (428), Chronic Obstructive Pulmonary Disease (COPD) (496), Diabetes (250), Epilepsy (345), HIV (071), Hypertension (401), Ischemic Heart Disease (IHD) (411), and Schizophrenia (295.1). We used these labels as the gold-standard to evaluate the accuracy of our automatic phenotyping algorithm (i.e., using the corresponding topic mixture $\boldsymbol{\theta}_k$ for our prediction score for phenotype k).

5.2 PheCodes

As the phenotype topic seed-guide, we extracted seed ICD-9 codes for phenotypes based on the expert-maintained PheWAS catalog (https://phewascatalog.org/phecodes) [9, 10]. The PheWAS catalog contains a reference table that groups ICD-9 codes into about 1,800 phenotypes. To represent broader parent phenotype, PheWAS also provides a set of hierarchically roll-up phenotype codes from the subphenotypes. Specifically, the compiled PheCodes are divided into two levels: (1) integer-level codes (i.e. 495) refer to parent phenotypes (e.g., Asthma); (2) 1-decimal-level codes (i.e. 313.1) refer to subphenotypes (e.g., ADHD). In our experiment, we used the subphenotypes with 1-decimal PheCodes (1,569 in total) as the phenotype topics. To compute 570 parent-level phenotypes at the integer-level PheCodes, we aggregated the inferred topic distributions $\boldsymbol{\Phi}$ and phenotype mixtures $\boldsymbol{\Theta}$ for the corresponding 1-decimal-level PheCodes.

5.3 Evaluation on phenotype predictions

Here we sought to evaluate the prediction accuracy of the 11 phenotypes, where we had rule-based labels (Section 5.1). Because our longitudinal EHR data can have up to 20-year follow-up for an individual, we may have several 5-year bins of EHR documents for the same individual. To predict the 11 phenotypes of each individual j, we aggregated the topic mixture memberships at the document-level onto the individual-level to obtain the 11 phenotype risk scores: $\boldsymbol{\Theta}_j = \sum_{d=1}^{D_j} \boldsymbol{\Theta}_{dj}$.

To obtain robust quantitative evaluation, we experimented with 5 repeated runs on random split of 70% training, 10% validation, and 20% test sets. We assessed the prediction accuracy of 11 target phenotypes on the test set using the area under precision-recall curves (AUPRC). We summarized the average AUPRC in Table. 2.

5.4 Baseline methods

As a simple baseline, we used the normalized counts of ICD codes under each PheCode for each patient as its predicted probabilities and called this method MaxPred [16]. As another baseline method called SeedGMM, we ran a two-component Gaussian mixture model (GMM) on each of the normalized PheCode counts to obtain predicted probabilities [1].

We also compared the predictive power of using two unsupervised topic models: the LDA that models individuals directly and DTM that models age-dependent topic dynamics with a Kalman filter rather than LSTM in our MixEHR-Seed [4, 5]. Neither of the two methods is able to use seed-guide. For these two methods, we used the validation set to choose the best topic number based on the

Table 2: Prediction performance of 11 target phenotypes by the proposed MixEHR-Seed model variants and 5 baseline methods on the test set. The mean value of AUPRC for each model on the test set are computed over 5 random splits of train-validation-test set. The best AUPRC scores for each phenotype were boldfaced.

Phenotype	MixEHR-Seed (multi-modal)	MixEHR-Seed (multi-modal) + SVM	MixEHR-Seed (uni-modal)	MixEHR-Seed (uni-modal) + SVM	MixEHR-Seed (uni-modal) Kalman	MaxPred	SeedGMM	LDA + SVM	DTM + SVM
ADHD	91.12	**92.86**	89.21	90.58	86.41	85.02	83.33	82.23	82.86
Asthma	70.36	**72.21**	64.96	69.12	61.56	50.73	49.98	52.38	59.67
Autism	92.71	**93.35**	89.66	91.35	89.91	86.90	85.20	85.86	89.62
CHF	74.95	**76.24**	69.81	72.25	60.18	52.28	51.92	44.50	58.37
COPD	76.84	**77.18**	78.51	80.21	77.52	74.46	75.60	64.81	77.41
Diabetes	87.76	**90.05**	86.54	88.93	83.96	70.53	71.97	78.62	82.38
Epilepsy	52.75	55.34	52.21	**56.38**	46.10	42.16	40.31	39.14	45.30
HIV	70.22	**74.26**	67.19	75.28	60.19	53.11	48.37	46.92	46.67
Hypertension	80.67	**85.84**	80.40	84.14	79.96	72.08	74.26	72.47	79.22
IHD	76.81	**77.89**	73.31	76.43	73.86	65.51	66.48	68.43	72.78
Schizophrenia	94.53	**96.04**	89.24	90.31	88.25	88.08	87.53	85.21	86.37

unsupervised perplexity and implemented a downstream support vector machine (SVM) classifier to perform prediction.

For the purpose of comparison, we also applied our MixEHR-Seed on the ICD codes to infer age-dependent topic hyperparameters via Kalman filter variational approximation same as in DTM [5]. The details of the variational inference based on Kalman filter is described in **Appendix** A.4. Moreover, we also applied SVM classifier on the inferred θ from our proposed MixEHR-Seed to assess additional gain of prediction accuracy from the strong classifier. We implemented LDA and DTM using Python packages scikit-learn and Gensim (https://radimrehurek.com/gensim/) with default inference algorithm. The supervised SVM classifier was also trained with scikit-learn package.

6 RESULTS

6.1 Automatic phenotyping prediction accuracy

We examined our model on disease prediction task by comparing predicted probabilities against the rule-based phenotype labels, which were established by heuristic observations over the medical history for each patient [3]. Table 2 displays the average AUPRC of compared methods in predicting 12 target phenotypes, where the rule-based labels are available. We observed that MixEHR-Seed outperformed all of the baseline models. This is possibly attributable to the more precise topic discovery by effectively leveraging the PheCode information for each phenotype.

Moreover, MixEHR-Seed using both ICD codes and ATC codes conferred higher AUPRC than MixEHR-Seed using only ICD. In fact, it is the best predictive model for most diseases (10 of the 11 selected phenotypes). This demonstrated the benefit of modeling multi-modal topic distributions on the EHR data (Section 4.2.3). Our MixEHR-Seed model achieved accurate prediction (over 80% AUPRC) in predicting ADHD, Autism, Diabetes, Hypertension, Schizophrenia. Nevertheless, our model performed worse in predicting Epilepsy possibly due to lesser accurate ICD seed codes and ambiguous rule-based labels because of their similarities with other diseases.

DTM that infers age-dependent topic priors confers higher prediction accuracy compared with LDA that uses a flat topic prior. In particular, the incorporation of temporal information yielded substantial prediction improvement over LDA for chronic diseases including CHF, COPD, Hypertension. This may be attributable to DTM's ability to model the distinct age-dependent disease evolution of these diseases. Moreover, the MixEHR-Seed with Kalman filter approximation generally outperformed DTM+SVM, highlighting the advantage of the seed-guided approach. Lastly, the default MixEHR-Seed using the VAE framework with amortized LSTM model conferred higher AUPRC compared to MixEHR-Seed using Kalman filter. This is possibly attributable to the flexibility of the LSTM model that led to more closely approximated posterior distribution.

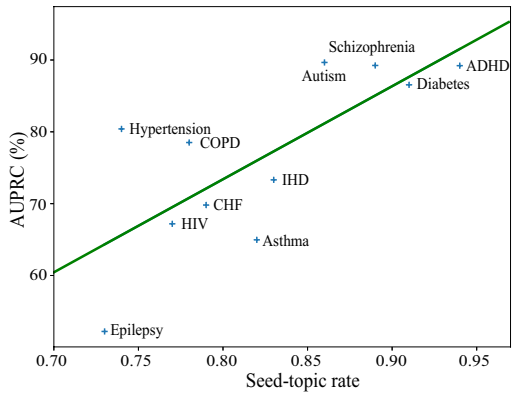

Figure 2: Scatterplot of seed-topic rate π_k and AUPRC of 11 target phenotypes. The inferred π_k for topic k indicates the rate that a seed EHR code is assigned to the seed topic distribution rather than the regular topic distribution.

Figure 3: Regular disease topics ϕ^r inferred by MixEHR-Seed on the PopHR database. Most of top ICD diagnostic codes from the regular topics belong to seed sets, therefore the inference of regular topics benefits from the guided information. Moreover, the topic inference in the ATC modality without seed codes also benefits from the seed-guided information since the identified ATC codes exhibit meaningful clinical information. a. Top 4 ICD codes along with major categories from the regular topics for 11 target phenotypes. The annotated red color indicates the seed ICD codes for the corresponding phenotype topics. b. Top 4 ATC codes along with major categories from the regular topics for 11 target phenotypes.

Interestingly, we observed a strong positive linear relationship between the inferred seed topic rate π_k and the AUPRC by MixEHR-Seed (Figure 2; Pearson's correlation coefficient is 0.74). Therefore, perhaps the seed topic rate implies the reliability of the seed codes and the resulting model's ability to predict the phenotypes in the current dataset.

6.2 MixEHR-Seed confers meaningful phenotype topics

To qualitatively evaluate topic interpretability, we examined the top 4 ICD codes and the top 4 ATC codes from the regular topics ϕ^r for the 11 selected disease phenotypes. We visualized the topic distributions for the ICD and ATC modalities (Figure 3). Indeed, MixEHR-Seed captured coherent and diverse types of clinical features for the 11 target phenotypes. In particular, the top 4 highest-scoring ICD codes under each topic exhibit strong clinical relevance to the target phenotypes (Figure 3a).

We observed that the inferred regular topics also tend to confer high probabilities to the seed ICD codes that were used to guide the seed topic inference during the training. For instance, the regular topic for Schizophrenia (PheCode 295.1) assigned high probabilities to ICD codes 2953, 2950, 2957, 2951 corresponding to the diagnoses of paranoid type Schizophrenia, simple type Schizophrenia, schizoaffective disorder, and disorganized type Schizophrenia, respectively. All of these ICD-9 codes are the seed codes used to define

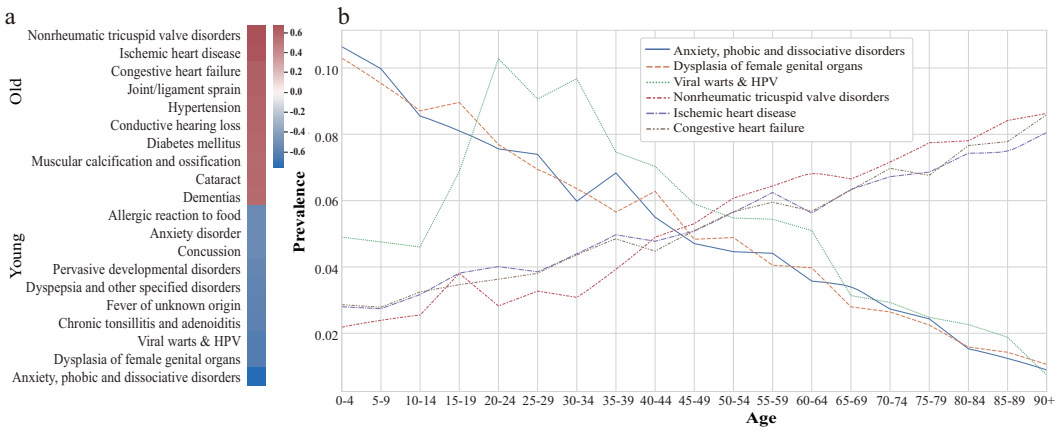

Figure 4: Inferred highly age-dependent phenotypes in Quebec population. a Top 10 age-correlated phenotypes by computing Pearson's correlations between MixEHR-Seed's age-dependent topic hyperparameters α and an increasing age vector. b Temporal progression of top 3 age-correlated phenotypes across age groups in the Quebec population.

Schizophrenia by PheWAS. The top ICD-9 codes of phenotype Diabetes including 2506, 2504, 2500, 2509 are also Diabetes-specific diagnostic codes. All of these ICD diagnostic codes are the seed codes of its associated subphenotype Diabetes mellitus (PheCode 250.0). Conversely, some of the top ICD codes are not seed codes (as colored in black in Figure 3a) but they also exhibit clinical relevance to the phenotype, showing our model's ability to discover unknown disease co-morbidity features. For example, alcoholic psychoses (2918) and compulsive conduct disorder (3122) are not a respective seed code for ADHD and Autism but they are associated with psychiatric behaviours.

Also, The top ICD codes under each phenotype often fall within the same disease categories as defined by the ICD taxonomy (as indicated by the left-side color bar in Figure 3a), which is not used by our model during the training. For instance, most of the top ICD codes for the ADHD, Autism, and Schizophrenia topics belong to the category of mental disorders; the top codes of CHF, Hypertension, and IHD all belong to the category of the circulatory system.

Moreover, the top 4 ATC codes under each phenotype topic also convey meaningful clinical information despite the fact that none of the ATC codes is in the seed codes (i.e., all of the seed codes are ICD-9 codes.) (Figure 3b). For instance, the ATC drug codes for ADHD, Autism, Epilepsy, and Schizophrenia are commonly used for treatment of nervous system (i.e., the general drug category shown in the left side-bar). On the other hand, the common drugs from Diabetes, CHF, and Hypertension belong to the general drug category for cardiovascular system. These results therefore highlight the effectiveness of multi-modal topic inference via the simple and elegant closed-form update solution (Eq. (10)).

6.3 Age-dependent disease topic prevalence

We explored temporal progression of phenotypes of interest in the population by examining the $T \times K$ age-dependent topic prior hyperparameters $\boldsymbol{\alpha}$. To evaluate relative prevalence of disease topics, we normalized each phenotype's dynamic topic parameter $\boldsymbol{\alpha}_k$ over T age groups. Figure 4 illustrates the relative prevalence of the top 3 disease topics that positively or negatively correlated with age. We observed distinct temporal evolution patterns of diseases such as viral warts and HPV, whose prevalence peaks around the most sexually active age between 20 and 24 and then decreases dramatically. In contrast, anxiety and dysplasia of female genital organs are at-birth conditions. The risk of heart diseases such as nonrheumatic tricuspid valve disorders, IHD, CHF, Hypertension rise up as age increases. This is also expected since senior people are often at the higher risk of having the circulatory and cardiovascular diseases.

7 DISCUSSION

Existing machine learning methods are limited in their abilities to predict thousands of target labels while learning their connections with the input features. This hinders their applications on important problems such as understanding the network connections among diverse human disease phenotypes. To address this challenge, we developed a seed-guided Bayesian topic model MixEHR-Seed. The crux of our approach is to infer a large number of topics (i.e., over 1500 topics in our application) each corresponding to exactly one target label. To accomplish that, we compute each topic with two distributions: the seed-topic distribution over only the known key words (i.e., seed EHR codes) for the target label and the regular-topic distribution over the entire feature vocabulary. While the former provides anchors to the inference making each topic identifiable, the latter compensates the incompleteness of the keyword vocabularies by providing the full-spectrum of distribution.

To account for demographic and multi-modality information in the EHR data, we extended our model in two ways: (1) inferring age-dependent temporal topic distributions by exploiting the longitudinal aspect of the EHR data; (2) a unified multi-modal framework to incorporate diverse types of clinical observations including not only the ICD-9 codes but also data types where the seed codes are not available (e.g., drug codes). To efficiently approximate the unknown parameters, we developed a hybrid Bayesian learning algorithm combining variational mean-field inference and gradient-based amortized inference using deep learning techniques.

We demonstrated MixEHR-Seed on a large-scale administrative claims database in Quebec, Canada, consisting of an entire cohort of over 1.3 million individuals along with more than 122 million health records for 8000 unique ICD-9 and 1200 unique ATC codes. MixEHR-Seed simultaneously performed automatic phenotyping (i.e., diagnostic predictions) over all patients and learning disease comorbidities from the heterogeneous longitudinal EHR data. To guide phenotypic topic inference, we leveraged 1569 expert-curated PheCodes from PheWAS database [10] to infer the same number of phenotype topics that comprehensively characterize a broad spectrum of diverse diseases. We observed that the inferred phenotype topic mixture memberships delivered accurate prediction of phenotype risks over 11 phenotypes, for which we had gold-standard labels. By associating each patient's phenotype topic mixture with age-dependent topic prior, we learned interesting temporal disease prevalence across age groups in the Quebec population. Our topic analysis found that the seed-guided inference mechanism does not only restrict seed topics to concentrate on the seed ICD-9 diagnostic codes but also helps the regular topics to capture clinically relevant ICD-9 codes and ATC drug codes to define each of 1569 phenotypes.

Together, we envision broad applications of MixEHR-Seed in healthcare and other research domains such as genomics in predicting high-dimensional labels using a set of associated features as guide while teasing apart topical patterns.

REFERENCES

[1] Yuri Ahuja, Doudou Zhou, Zeling He, Jiehuan Sun, Victor Castro, Vivian Gainer, Shawn Murphy, Chuan Hong, and Tianxi Cai. 2020. sureLDA: A multidisease automated phenotyping method for the electronic health record. *Journal of the American Medical Informatics Association : JAMIA* 27 (06 2020). https://doi.org/10.1093/jamia/ocaa079

[2] Yuri Ahuja, Yuesong Zou, Aman Verma, David Buckeridge, and Yue Li. 2021. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. *bioRxiv* (2021). https://doi.org/10.1101/2021.12.17.473215

[3] M T Betancourt, K C Roberts, T-L Bennett, E R Driscoll, G Jayaraman, and L Pelletier. 2014. Monitoring chronic diseases in Canada: the Chronic Disease Indicator Framework. *Chronic diseases and injuries in Canada* 34 Suppl 1 (2014), 1–30.

[4] DM Blei, AY Ng, and MI Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[5] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 113–120. https://doi.org/10.1145/1143844.1143859

[6] David M. Blei and Jon D. McAuliffe. 2007. Supervised Topic Models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 121–128.

[7] You Chen, Joydeep Ghosh, Cosmin Bejan, Carl Gunter, Siddharth Gupta, Abel Kho, David Liebovitz, J. Sun, Joshua Denny, and Bradley Malin. 2015. Building Bridges Across Electronic Health Record Systems Through Inferred Phenotypic Topics. *Journal of biomedical informatics* 55 (04 2015). https://doi.org/10.1016/j.jbi.2015.03.011

[8] Eliezer de Souza da Silva, Helge Langseth, and Heri Ramampiaro. 2017. Content-Based Social Recommendation with Poisson Matrix Factorization. In *ECML/PKDD*.

[9] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology* 31, 12 (2013), 1102–1111.

[10] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 26, 9 (2010), 1205–1210.

[11] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The Dynamic Embedded Topic Model. arXiv:1907.05545 [cs.CL]

[12] Saurav Ghosh, Prithwish Chakraborty, Elaine Nsoesie, Emily Cohn, Sumiko Mekaru, John Brownstein, and Naren Ramakrishnan. 2016. Temporal Topic Modeling to Assess Associations between News Trends and Infectious Disease Outbreaks. *arXiv:1606.00411* 7 (06 2016). https://doi.org/10.1038/srep40841

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] 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

[15] George Hripcsak and DJ Albers. 2012. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA* 20 (09 2012). https://doi.org/10.1136/amiajnl-2012-001145

[16] Jagadeesh Jagarlamudi, Hal Daumé, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France) *(EACL '12)*. Association for Computational Linguistics, USA, 204–213.

[17] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[18] Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna Biernacka, Euijung Ryu, Janet Olson, Mark 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 (05 2020), 2536. https://doi.org/10.1038/s41467-020-16378-3

[19] Katherine Liao, Tianxi Cai, Guergana Savova, Shawn Murphy, Elizabeth Karlson, Ashwin Ananthakrishnan, Vivian Gainer, Stanley Shaw, Zongqi Xia, Peter Szolovits, Susanne Churchill, and Isaac Kohane. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 350 (04 2015), h1885–h1885. https://doi.org/10.1136/bmj.h1885

[20] Katherine Newton, Peggy Peissig, Abel Kho, Suzette Bielinski, Richard Berg, Vidhu Choudhary, Melissa Basford, Christopher Chute, Iftikhar Kullo, Rongling Li, Jennifer Pacheco, Luke Rasmussen, Leslie Spangler, and Joshua Denny. 2013. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association : JAMIA* 20 (03 2013). https://doi.org/10.1136/amiajnl-2012-000896

[21] Sonal Parasrampuria and Jawanna Henry. 2019. Hospitals' Use of Electronic Health Records Data, 2015-2017.

[22] Rimma Pivovarov, Adler Perotte, Edouard Grave, John Angiolillo, Chris Wiggins, and Noémie Elhadad. 2015. Learning Probabilistic Phenotypes from Heterogeneous EHR Data. *Journal of biomedical informatics* 58 (10 2015). https://doi.org/10.1016/j.jbi.2015.10.001

[23] Arash Shaban-Nejad, Maxime Lavigne, Anya Okhmatovskaia, and David Buckeridge. 2016. PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data: The Population Health Record (PopHR). *Annals of the New York Academy of Sciences* 1387 (10 2016). https://doi.org/10.1111/nyas.13271

[24] 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 *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (Gainesville, Florida) *(BCB '21)*. Association for Computing Machinery, New York, NY, USA, Article 6, 26 pages. https://doi.org/10.1145/3459930.3469543

[25] Yee Whye Teh, David Newman, and Max Welling. 2006. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems* (Canada) *(NIPS'06)*. MIT Press, Cambridge, MA, USA, 1353–1360.

[26] Yanshan Wang, Yiqing Zhao, Terry Therneau, Elizabeth Atkinson, Ahmad P. Tafti, Nan Zhang, Shreyasee Amin, Andrew Limper, Sundeep Khosla, and Hongfang Liu. 2019. Unsupervised Machine Learning for the Discovery of Latent Disease Clusters and Patient Subgroups Using Electronic Health Records. *Journal of Biomedical Informatics* 102 (12 2019), 103364. https://doi.org/10.1016/j.jbi.2019.103364

[27] Wei-Qi Wei and Joshua Denny. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine* 7 (04 2015). https://doi.org/10.1186/s13073-015-0166-y

[28] Mengru Yuan, Guido Powell, Maxime Lavigne, Anya Okhmatovskaia, and David Buckeridge. 2018. Initial Usability Evaluation of a Knowledge-Based Population Health Information System: The Population Health Record (PopHR). *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2017 (04 2018), 1878–1884.

A APPENDIX

A.1 Simulation

As ground-truth topics are unobserved from the real data, we conducted a simulation to validate how well our method can recover the ground-truth topics. In the data generative process, we simulated 1,000 documents, 10 time points, and 20 topics for evaluation, where each document is associated with a uniformly sampled time point. To incorporate prior knowledge, we randomly chose 20 PheCodes from PheWAS with 388 associated ICD codes to generate topics with dynamic prior $\boldsymbol{\alpha}$ across 10 time points. We then simulated the documents by following the generative process described in Section 4.1. We also experimented a static simulation with the same generative process except that the topic prior *alpha* is fixed to a small constant value.

We assessed topic recovery by calculating the Pearson's correlation coefficient between the true and inferred topic mixture memberships θ for each document.We evaluated an ablated Static MixEHR-Seed with a flat topic prior $\boldsymbol{\alpha}$ and our proposed Dynamic MixEHR-Seed with dynamic topic prior $\boldsymbol{\alpha}$. We also compared the two models with a baseline, where we computed the normalized occurrence of seed words under each topic for each document as its predicted probabilities. In the static simulation, our static MixEHR-Seed outperformed the raw count baseline highlighting the benefits of the probabilistic inference algorithm (87.14% vs 76.84%). In the dynamic simulation, Dynamic MixEHR-Seed conferred the highest correlation because of its ability to capture the dynamic topic prior information (Figure 5). Therefore, both simulations validated the correctness of our implementation.

A.2 Additional details of MixEHR-Seed

The complete joint likelihood of MixEHR-Seed model (Table 1) is:

$$p(\mathbf{w}, \mathbf{z}, \mathbf{x}, \boldsymbol{\phi}^s, \boldsymbol{\phi}^r, \theta, \boldsymbol{\alpha}, \boldsymbol{\eta}) = p(\boldsymbol{\eta})p(\boldsymbol{\alpha} \mid \boldsymbol{\eta})p(\theta \mid \boldsymbol{\alpha})$$
$$p(\mathbf{z} \mid \theta)p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\pi})p(\boldsymbol{\phi}^r \mid \beta)p(\boldsymbol{\phi}^s \mid \mu)p(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \boldsymbol{\phi}^s, \boldsymbol{\phi}^r) \quad (16)$$

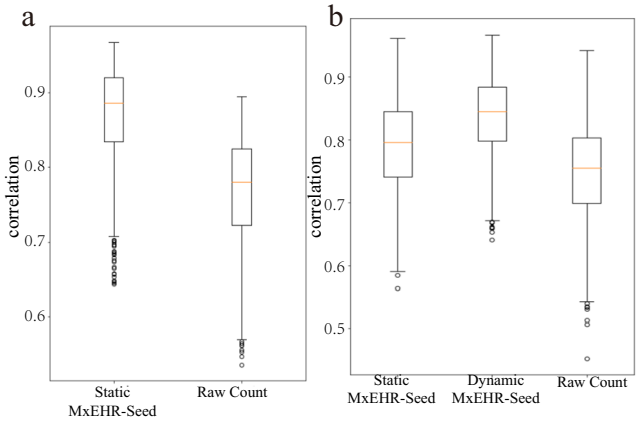

Figure 5: Model comparison by simulation. Pearson correlation between inferred and groundtruth topic mixtures for 1000 simulated documents with a static topic prior and b dynamic topic prior (Section A.1).

Due to conjugate property of Dirichlet and multinomial distributions we can integrate out the Dirichlet variables:

$$p(\mathbf{z} \mid \boldsymbol{\alpha}) = \int p(\theta, \mathbf{z} \mid \theta)d\theta = \prod_d^D \frac{\Gamma(\sum_k^K \alpha_{t_d k})}{\prod_k^K \Gamma(\alpha_{t_d k})} \frac{\prod_k^K \Gamma(m_{dk} + \alpha_{t_d k})}{\Gamma(m_{d.} + \sum_k^K \alpha_{t_d k})}$$

$$p(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \beta, \mu) = \iint p(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \boldsymbol{\phi}^r, \boldsymbol{\phi}^s)p(\boldsymbol{\phi}^r \mid \beta)p(\boldsymbol{\phi}^s \mid \mu)d\boldsymbol{\phi}^r d\boldsymbol{\phi}^s$$
$$= \prod_k^K \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \frac{\prod_w^V \Gamma(n_{wk} + \beta)}{\Gamma(n_{.k} + V\beta)} \frac{\Gamma(V_k\mu)}{\Gamma(\mu)^{V_k}} \frac{\prod_w^{V_k} \Gamma(s_{wk} + \mu)}{\Gamma(s_{.k} + V_k\mu)}$$

The conditional distribution $p(z_{di} = k \mid z_{-di}, \mathbf{w}, \mathbf{x})$ is:

$$p(z_{di} = k \mid z_{-di}, \mathbf{w}, \mathbf{x}) = \frac{p(z_{di} = k, z_{-di}, \mathbf{w}, \mathbf{x})}{\sum_k^K p(z_{di} = k, z_{-di}, \mathbf{w}, \mathbf{x})}$$
$$\propto (m_{dk}^{-di} + \alpha_{t_d k}) \left(\frac{n_{wk}^{-di} + \beta}{n_{.k}^{-di} + V\beta}(1 - \pi_k) \right)^{[x_{di}=0]} \left(\frac{s_{wk}^{-di} + \mu}{s_{.k}^{-di} + V_k\mu}\pi_k \right)^{[x_{di}=1]}$$

The resulting marginal likelihood can be approximated by evidence lower bound (ELBO):

$$\log p(\mathbf{w} \mid \beta, \mu, \boldsymbol{\pi}) = \log \int \sum_{\mathbf{z}} \sum_{\mathbf{x}} p(\mathbf{w}, \mathbf{x}, \mathbf{z}, \boldsymbol{\eta} \mid \beta, \mu, \boldsymbol{\pi})d\boldsymbol{\eta}$$
$$\geq \int \sum_{\mathbf{z}} q(\mathbf{z}, \boldsymbol{\eta}) \log(\frac{\sum_{\mathbf{x}} p(\mathbf{w}, \mathbf{x}, \mathbf{z}, \boldsymbol{\eta} \mid \beta, \mu, \boldsymbol{\pi})}{q(\mathbf{z}, \boldsymbol{\eta})})d\boldsymbol{\eta}$$
$$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\eta})}[\log \sum_{\mathbf{x}} p(\mathbf{w}, \mathbf{x}, \mathbf{z}, \boldsymbol{\eta}, \mid \beta, \mu, \boldsymbol{\pi})] - \mathbb{E}_{q(\mathbf{z},\boldsymbol{\eta})}[\log q(\mathbf{z}, \boldsymbol{\eta})]$$
$$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\eta})}[\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta} \mid \beta, \mu, \boldsymbol{\pi})] - \mathbb{E}_{q(\mathbf{z},\boldsymbol{\eta})}[\log q(\mathbf{z}, \boldsymbol{\eta})] = L_{ELBO}$$

Here we also marginalize out the indicators $\mathbf{x}$ to simplify the target ELBO. For the respective latent variables $\mathbf{z}$ and $\boldsymbol{\eta}$, we propose independent multinomial and Gaussian variational distributions such that the resulting ELBO can be approximated with closed-form:

$$L_{ELBO} = \mathbb{E}_{q(\boldsymbol{\eta})}[\log p(\boldsymbol{\eta})] + \mathbb{E}_{q(\mathbf{z},\boldsymbol{\eta})}[\log p(\mathbf{z} \mid \boldsymbol{\alpha})] + \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{w} \mid \mathbf{z}, \beta, \mu, \boldsymbol{\pi})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z} \mid \boldsymbol{\gamma})] - \mathbb{E}_{q(\boldsymbol{\eta})}[\log q(\boldsymbol{\eta})] \quad (17)$$

$$\mathbb{E}_{q(\boldsymbol{\eta})}[\log p(\boldsymbol{\eta})] = -\sum_t^T \sum_k^K \frac{1}{2}(\log 2\pi + \log \sigma_p^2 + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\sigma_p^2}) \quad (18)$$

$$\mathbb{E}_{q(\boldsymbol{\eta})}[\log q(\boldsymbol{\eta})] = -\sum_t^T \sum_k^K \frac{1}{2}(\log 2\pi + 1 + \log \sigma_q^2) \quad (19)$$

$$\mathbb{E}_{q(\mathbf{z},\boldsymbol{\alpha})}[\log p(\mathbf{z} \mid \boldsymbol{\alpha})] = \sum_d^D \log \Gamma(\sum_k^K \mathbb{E}_{q(\boldsymbol{\alpha})}[\alpha_{t_d k}])$$
$$- \sum_k^K \log \Gamma(\mathbb{E}_{q(\boldsymbol{\alpha})}[\alpha_{t_d k}]) + \sum_k^K \log \Gamma(\mathbb{E}_{q(\mathbf{z})}[m_{dk}] + \mathbb{E}_{q(\boldsymbol{\alpha})}[\alpha_{t_d k}])$$
$$- \log \Gamma(\mathbb{E}_{q(\mathbf{z})}[m_{d.}] + \sum_k^K \mathbb{E}_{q(\boldsymbol{\alpha})}[\alpha_{t_d k}]) \quad (20)$$

$$\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z} \mid \boldsymbol{\gamma})] = \sum_k^K \sum_d^D \sum_i^{N_i} \gamma_{dik} \log \gamma_{dik} \quad (21)$$

For the likelihood term in ELBO, we approximate the expectation using a lower bound for the expected value of log-sum $E_q[\log \sum_k x_k] \geq \log \sum_k \exp\{E_q[\log x_k]\}$ [8]:

$$\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{w} \mid \mathbf{z}, \beta, \mu, \boldsymbol{\pi})] = \mathbb{E}_{q(\mathbf{z})}[\log \sum_{\mathbf{x}} p(\mathbf{w}, \mathbf{x} \mid \mathbf{z}, \beta, \mu, \boldsymbol{\pi})]$$

$$= \mathbb{E}_{q(\mathbf{z})}\Big[\log\Big(\prod_k^K \frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\frac{\prod_w^V \Gamma(n_{wk}+\beta)}{\Gamma(n_{\cdot k}+V\beta)}(1-\pi_k)^{n_{wk}}$$

$$+ \prod_k^K \frac{\Gamma(V_k\mu)}{\Gamma(\mu)^{V_k}}\frac{\prod_w^{V_k}\Gamma(s_{wk}+\mu)}{\Gamma(s_{\cdot k}+V_k\mu)}\pi_k^{s_{wk}}\Big)\Big]$$

$$\geq \log\Bigg(\prod_k^K \frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\frac{\prod_w^V \Gamma(\mathbb{E}_{q(\mathbf{z})}[n_{wk}]+\beta)}{\Gamma(\mathbb{E}_{q(\mathbf{z})}[n_{\cdot k}]+V\beta)}(1-\pi_k)^{\mathbb{E}_{q(\mathbf{z})}[n_{wk}]}$$

$$+ \prod_k^K \frac{\Gamma(V_k\mu)}{\Gamma(\mu)^{V_k}}\frac{\prod_w^{V_k}\Gamma(E_{q(\mathbf{z})}[s_{wk}]+\mu)}{\Gamma(\mathbb{E}_{q(\mathbf{z})}[s_{\cdot k}]+V_k\mu)}\pi_k^{\mathbb{E}_{q(\mathbf{z})}[s_{wk}]}\Bigg) \tag{22}$$

The update of variational parameter $\boldsymbol{\gamma}_{di}$ is described in Section 4.2.

A.3 Initialization of variational expectations

We initialized sufficient statistics for $\mathbb{E}_{q(\mathbf{z})}[n_{wk}]$ and $\mathbb{E}_{q(\mathbf{z})}[s_{wk}]$ as follows. If a word w is a seed word for topic k, it was assigned to seed topic count n_{wk} instead of its regular topic count s_{wk} according to a seed-topic rate π_k; if a word w is a regular word w.r.t. topic k, it is assigned to the regular topic count s_{wk} with uninformative uniform prior $\frac{1}{K}$:

$$\mathbb{E}_{q(\mathbf{z})}[n_{wk}] = \sum_d^D \sum_i^{N_d} [w_{di} = w, w \in \mathcal{V}_k](1-\pi_k)$$
$$+ [w_{di} = w, w \notin \mathcal{V}_k]\frac{1}{K}$$
$$\mathbb{E}_{q(\mathbf{z})}[s_{wk}] = \sum_d^D \sum_i^{N_d} [w_{di} = w, w \in \mathcal{V}_k]\pi_k \tag{23}$$

We set the initial value of $\boldsymbol{\pi}$ as 0.7 across all k topics. Finally, we initialized $\mathbb{E}_{q(\mathbf{z})}[m_{dk}]$ with uniform rate: $\mathbb{E}_{q(\mathbf{z})}[m_{dk}] = \frac{N_d}{K}$.

A.4 Kalman Filter variational approximation

An alternative variational approach to infer age-dependent topic hyperparameters $\boldsymbol{\eta}$ is via the Kalman filter algorithm as suggested by [5]. The state space model is defined as follows:

$$\boldsymbol{\eta}_t \mid \boldsymbol{\eta}_{t-1} \sim N(\boldsymbol{\eta}_{t-1}, \delta^2 I), \quad \hat{\boldsymbol{\eta}}_t \mid \boldsymbol{\eta}_t \sim N(\boldsymbol{\eta}_t, \hat{v}_t^2 I) \tag{24}$$

where $\hat{\boldsymbol{\eta}}_t$ and $\hat{v}_t$ are the variational parameters. For each topic k, we first compute the forward mean $\boldsymbol{\mu}_t = \mathbb{E}[\boldsymbol{\eta}_t \mid \hat{\boldsymbol{\eta}}_{1:t}]$ and variance $\mathbf{V}_t = \mathbb{E}[(\boldsymbol{\eta}_t - \boldsymbol{\mu}_t)^2 \mid \hat{\boldsymbol{\eta}}_{1:t}]$ with initial values $\boldsymbol{\mu}_0$ and $\mathbf{V}_0$:

$$\mu_{tk} = \frac{\hat{v}_t^2}{V_{t-1,k}+\delta^2+\hat{v}_t^2}\mu_{t-1,k} + (1 - \frac{\hat{v}_t^2}{V_{t-1,k}+\delta^2+\hat{v}_t^2})\hat{\eta}_{tk}$$
$$V_{tk} = \frac{\hat{v}_t^2}{V_{t-1,k}+\delta^2+\hat{v}_t^2}(V_{t-1,k}+\delta^2) \tag{25}$$

We then calculate the mean $\tilde{\boldsymbol{\mu}}_{t-1} = \mathbb{E}[\boldsymbol{\eta}_{t-1} \mid \hat{\boldsymbol{\eta}}_{1:T}]$ and variance $\tilde{\mathbf{V}}_{t-1} = \mathbb{E}[(\boldsymbol{\eta}_{t-1} - \tilde{\boldsymbol{\mu}}_{t-1})^2 \mid \hat{\boldsymbol{\eta}}_{1:T}]$:

$$\tilde{\mu}_{t-1,k} = \frac{\delta^2}{V_{t-1,k}+\delta^2}\mu_{t-1,k} + (1 - \frac{\delta^2}{V_{t-1,k}+\delta^2})\tilde{\mu}_{tk}$$
$$\tilde{V}_{t-1,k} = V_{t-1,k} + (\frac{V_{t-1,k}}{V_{t-1,k}+\delta^2})^2(\tilde{V}_{tk} - (V_{t-1,k}+\delta^2)) \tag{26}$$

where we have initial values $\tilde{\boldsymbol{\mu}}_T = \boldsymbol{\mu}_T$ and $\tilde{\mathbf{V}}_T = \mathbf{V}_T$ during the backward steps.

We rewrite the variational expectations and the related terms in ELBO w.r.t. $\boldsymbol{\eta}$ in Eq. (18) and Eq. (19):

$$\mathbb{E}_{q(\boldsymbol{\eta})}[\log p(\boldsymbol{\eta})] = -\frac{TK}{2}(\log\delta^2 + \log 2\pi) - \frac{1}{2\delta^2}\sum_{t=1}^T ||\tilde{\boldsymbol{\mu}}_t - \tilde{\boldsymbol{\mu}}_{t-1}||^2 \tag{27}$$
$$- \frac{1}{2\delta^2}\sum_{t=1}^T\sum_{k=1}^K\Big(\tilde{V}_{tk} + \tilde{V}_{t-1,k}\Big)$$

$$\mathbb{E}_{q(\boldsymbol{\eta})}[\log q(\boldsymbol{\eta})] = -\frac{TK}{2}(\log 2\pi + 1) - \frac{1}{2}\sum_{t=1}^T\sum_{k=1}^K \log\tilde{V}_{tk} \tag{28}$$

Finally, we maximize the ELBO w.r.t. $\hat{\boldsymbol{\eta}}_t$ via gradient descent:

$$\frac{\partial L_{ELBO}(\hat{\boldsymbol{\eta}})}{\partial \eta_{tk}} = -\frac{1}{\delta^2}(\tilde{\mu}_{tk} - \tilde{\mu}_{t-1,k})(\frac{\partial\tilde{\mu}_{tk}}{\partial\hat{\eta}_{tk}} - \frac{\partial\tilde{\mu}_{t-1,k}}{\partial\hat{\eta}_{t-1,k}})$$
$$+ \Bigg(\sum_{d:t_d=t}^D \Psi(\sum_k^K\tilde{\mu}_{tk}) - \sum_k^K\Psi(\tilde{\mu}_{tk}) + \sum_k^K\Psi(\mathbb{E}_{q(\mathbf{z})}[m_{dk}] + \tilde{\mu}_{tk})$$
$$- \Psi(\mathbb{E}_{q(\mathbf{z})}[m_{d\cdot}] + \sum_k^K\tilde{\mu}_{tk})\Bigg)\frac{\partial\tilde{\mu}_{tk}}{\partial\hat{\eta}_{tk}} \tag{29}$$