MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record

Yuri Ahuja[a,b,*], Yuesong Zou[c], Aman Verma[d], David Buckeridge[d,*], Yue Li[c,*]

[a]*Department of Biostatistics, Harvard TH Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA*
[b]*Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA*
[c]*School of Computer Science, McGill University, 3480 Rue University, Montreal, QC H3A 2A7, Canada*
[d]*School of Population and Global Health, McGill University, 2001 McGill College Avenue, Montreal, Québec H3A 1G1, Canada*

Abstract

Electronic Health Records (EHRs) contain rich clinical data collected at the point of the care, and their increasing adoption offers exciting opportunities for clinical informatics, disease risk prediction, and personalized treatment recommendation. However, effective use of EHR data for research and clinical decision support is often hampered by a lack of reliable disease labels. To compile gold-standard labels, researchers often rely on clinical experts to develop rule-based phenotyping algorithms from billing codes and other surrogate features. This process is tedious and error-prone due to recall and observer biases in how codes and measures are selected, and some phenotypes are incompletely captured by a handful of surrogate features. To address this challenge, we present a novel automatic phenotyping model called MixEHR-Guided (MixEHR-G), a multimodal hierarchical Bayesian topic model that efficiently models the EHR generative process by identifying latent phenotype structure in the data. Unlike existing topic modeling algorithms wherein the inferred topics are not identifiable, MixEHR-G uses prior information from informative surrogate features to align topics with known phenotypes. We applied MixEHR-G to an openly-available EHR dataset of 38,597 intensive care patients (MIMIC-III) in Boston, USA and to administrative claims data for a population-based cohort (PopHR) of 1.3 million people in Quebec, Canada. Qualitatively, we demonstrate that MixEHR-G learns

*Correspondence to yuri_ahuja@hms.harvard.edu, david.buckeridge@mcgill.ca, yueli@cs.mcgill.ca

interpretable phenotypes and yields meaningful insights about phenotype similarities, comorbidities, and epidemiological associations. Quantitatively, MixEHR-G outperforms existing unsupervised phenotyping methods on a phenotype label annotation task, and it can accurately estimate relative phenotype prevalence functions without gold-standard phenotype information. Altogether, MixEHR-G is an important step towards building an interpretable and automated phenotyping system using EHR data.
Keywords: electronic health records, deep phenotyping, topic modeling

---

1. Introduction

Electronic Health Records (EHRs) offer high volume comprehensive observational patient health data for clinical and translational research [1, 2]. The past 15 years have seen an explosion in EHR adoption, with the proportion of acute care hospitals and primary care practices in the United States using EHRs increasing from 9% and 17% respectively in 2008 to 96% and 86% by 2017 [3, 4]. Comprised of multiple types of data, such as lab tests, prescriptions, free-text clinical notes, and codified billing information including International Classification of Diseases (ICD), Current Procedural Terminology (CPT), and Diagnosis Related Group (DRG) codes, EHR data provide a comprehensive description of patients' interactions with the healthcare system over time. Such rich, large-scale data have myriad potential applications including personalized disease risk estimation and treatment recommendation, pillars of precision medicine [5]. However, effective use of EHR data for such applications is often hampered by a lack of reliable disease phenotype labels.

Many studies use billing codes (i.e. ICD codes) as surrogates for phenotype labels, a practice that works well for some phenotypes but notoriously poorly for others, including rheumatoid arthritis and chronic kidney disease [6–9]. Other studies have physicians or other health experts manually review patient records or devise rule-based phenotyping algorithms based on surrogate features such as ICD codes [10–17]. An example of this approach is the PheKB community phenotyping knowledge base, which has shown considerable success in generating accurate, portable rules [17]. While generally reliable, both chart review and rule production are tedious, time-intensive processes applicable

to labeling a handful of phenotypes for discovery research but not for phenotyping at scale, as might be needed for a multi-phenotype hypothesis-generating study such as a Phenome-Wide Association Study (PheWAS). Moreover, rule-based methods are subject to recall and observer biases in how features are selected.

To address this challenge, researchers have proposed a variety of automated computational phenotyping algorithms that infer phenotypes with little to no expert input or annotation. Harutyunyan et al. (2019) [18] used 17 handcrafted clinical variables to classify 25 Clinical Classifications Software (CCS) codes using logistic regression or recurrent neural networks, with each CCS defining a group of related ICD-9 codes as a disease category. Some notable unsupervised learning approaches include work in non-negative matrix/tensor factorization [19–21], mixture models [22–30], and neural network-based embedding learning [18, 31, 32]. One strategy that has shown considerable success is to use a handful of quickly identifiable surrogate features, including ICD codes and natural language processing (NLP)-derived mentions of a target phenotype in clinical notes, to generate “silver-standard" phenotype labels - or “anchors" - that can be subsequently used to guide phenotype estimation in a “weakly supervised" manner [25, 26, 29, 30, 33]. An ongoing challenge remains how best to leverage the thousands of additional features in the EHR to improve upon these silver-standard labels. Several methods employ regression with regularization or dropout to de-noise silver-standard labels using other features [25, 33], but this strategy tends to be sensitive to high feature dimensions and sparsity typical for EHR data.

Topic modeling algorithms are well-conditioned to this setting. Designed to model highly-sparse text data with vocabularies often consisting of tens of thousands of words, topic models have shown considerable aptitude at discovering latent structure (i.e. topics) in such datasets [30, 34–36]. Indeed, several studies have demonstrated success applying derivatives of the widely used topic model Latent Dirichlet Allocation (LDA) to EHR data, considering patients as documents, disease phenotypes as topics, and EHR features as words [30, 34]. One notable model, MixEHR, builds upon LDA to 1) simultaneously model an arbitrary number of data modalities (e.g., ICD codes, lab results) with modality-specific distributions, and 2) treat observations as not missing at random (NMAR), reflecting the

fact that clinicians typically collect data such as lab tests with a diagnosis in mind [34]. However, MixEHR is fully unsupervised and thus prone to learning topics that lack clinical meaning. On the other hand, the recent sureLDA method constrains topics to align with well-defined phenotypes using a weakly supervised surrogate-based approach, thereby learning meaningful, interpretable topics [30]. However, sureLDA treats all features as draws from topic-specific multinomial distributions, rendering it unable to properly model other data modalities such as lab tests and vital signs. Moreover, its Gibbs Sampling inference procedure does not scale well to datasets with more than thousands of features or a few hundred thousand patients.

In this study we present MixEHR-Guided (MixEHR-G), which is built upon two previously established methods: sureLDA [30] and MixEHR [34]. Our key contribution is the elegant use of guided topic priors originally introduced in sureLDA to efficiently infer multi-modal topic distributions over heterogeneous EHR data types using a variational inference algorithm, which was introduced in MixEHR. This powerful combination allows us to infer from large-scale EHR datasets over 1500 multi-modal phenotype topics each corresponding to a unique well-defined disease phenotype. We apply MixEHR-G to the Population Health Record (PopHR) dataset [37], which contains longitudinal administrative claims data for a random sample of 1.3 million residents of Quebec, Canada (**Methods**), and to the Medical Information Mart for Intensive Care (MIMIC-III) dataset [38], a publicly available dataset of Intensive Care Unit (ICU) observations at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. In both applications, MixEHR-G is able to effectively infer meaningful phenotype distributions over high-dimensional multimodal categorical EHR data including ICD codes, prescriptions, and discretized lab results. We use MixEHR-G's patient-topic mixture distributions to predict patient phenotypes and to analyze disease prevalence as a function of age and sex.

2. Methods

2.1. MixEHR-G model description

Building upon MixEHR, MixEHR-G uses a probabilistic joint topic modeling approach that projects a patient's high-dimensional and heterogeneous clinical record onto a low-

dimensional latent topic mixture membership over disease phenotypes. In contrast to MixEHR, MixEHR-G incorporates noisy phenotype labels obtainable *a priori* as topic priors for each phenotype to guide posterior topic inference. In this way, MixEHR-G harnesses the unsupervised topic modeling prowess of MixEHR to learn interpretable and well-defined phenotypes in a manner similar to the anchor-and-learn framework [39] or sureLDA [30]. In this study, we use Phenotype Codes (PheCodes), expert-defined groupings of ICD-9 codes developed for Phenome-Wide Association Studies (PheWAS), as topic priors. PheCodes have been shown to better align with clinically meaningful disease phenotypes and better replicate known genetic associations in genome-wide association studies (GWAS) than individual ICD-9 codes or Clinical Classification Software (CCS) groupings of ICD codes [40]. That said, MixEHR-G will work with any phenotype prior, including the ICD taxonomy or CCS codes; PheCodes are not the only phenotypic anchor schema compatible with our method. Crucially, the only change to MixEHR's computational framework that MixEHR-G introduces is the PheCode-driven initialization and re-optimization of the Dirichlet hyperparameter α, so no functionality featured by MixEHR is lost.

The MixEHR-G algorithm consists of 3 key steps: (1) assemble a comprehensive and heterogeneous EHR dataset (i.e. ICD codes, prescriptions, lab data, etc.); (2) use the Multimodal Automated Phenotyping (MAP) method to obtain initial probabilities for each reference phenotype based on key surrogate features within these data; and (3) train the MixEHR model [34] on the entire EHR dataset weighting the K asymmetrical patient-topic Dirichlet hyperparameters by the probabilities obtained in (2). We assume that there are in total K reference phenotypes corresponding to K true binary phenotype states that are unobserved *a priori*. Given a patient's EHR data, we predict the K phenotypes by inferring the posterior distributions of K guided topics, each corresponding to exactly one reference phenotype due to the use of constrained topic priors. **Fig.** 1 depicts this procedure, and we now describe each of the 3 steps in detail.

Step 1: preprocess EHR data. Following a topic modeling approach [41], we treat each patient record as a document and the various EHR codes throughout that patient's record as tokens. Because we do not consider the sequential order of observed codes, we essentially model the patient record as a bag of words (or rather a bag of EHR codes).

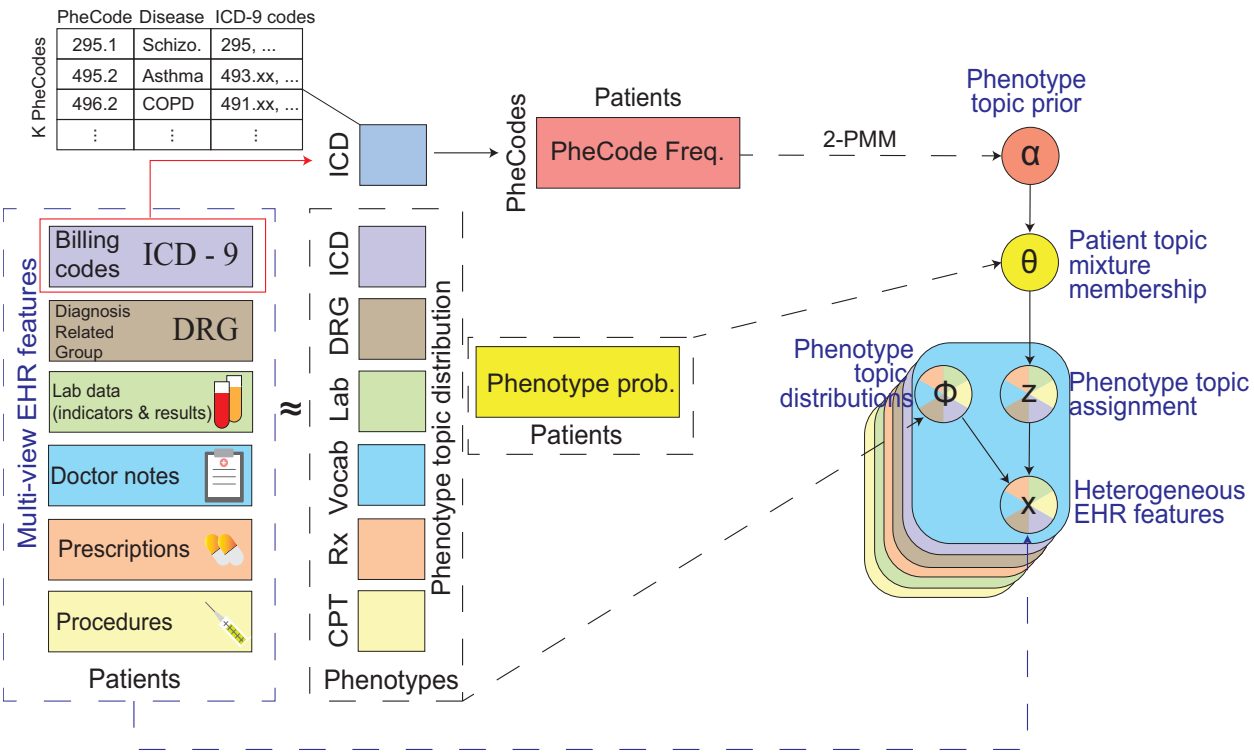


Figure 1: **Schematic of the MixEHR-G algorithm.** We factorize the heterogeneous EHR data matrices into phenotype topic distributions (i.e. phenotype topic distribution matrices) and 1 common patient topic mixture (i.e. patient phenotype probability matrix). To guide patient topic inference, we first infer the PheCode-guided topic prior $\boldsymbol{\alpha}_d$ for each patient d. This is done by training two-component Lognormal and Poisson Mixture models on the normalized frequency of the ICD codes that define the PheCode, and taking a weighted average of the estimated posterior probabilities for the higher components of the two models. We then incorporate this weighted posterior as the Dirichlet hyperparameters $\boldsymbol{\alpha}_d$ for the patient topic mixture $\boldsymbol{\theta}_d$. Finally, we infer the posterior distributions of the phenotype topic assignment for each EHR token j of patient d ($z_{d,j}$) and phenotype topic distributions ($\boldsymbol{\phi}$) using a collapsed mean-field variational inference algorithm.

EHR codes are often filtered by feature selection methods such as the Surrogate-Assisted Feature Extraction (SAFE) method [28], which may filter out useful information. Instead MixEHR-G has the capacity to assign probability weights to each EHR code under each phenotype topic, making it robust to a large number of uninformative codes. We included all available EHR codes without any filtering when assembling features. **Fig.** 1 depicts how multimodal count data are combined into multiple discrete-count matrices - one per data type - for input into the MixEHR-G algorithm.

Step 2: Initializing prior probabilities using MAP. We obtain prior probabilities for the K reference phenotypes, which we denote as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, using a modified MAP algorithm [26]. Standard MAP estimates π_k by fitting Poisson and Lognormal mixture

models to the counts of (1) phenotype k's core ICD code and (2) NLP-curated mentions of phenotype k in clinical notes, normalizing using a healthcare utilization feature H, and taking a weighted average of these various mixture models. Instead of ICD and NLP features, we use Phenotype Codes (PheCodes), which are coarser than ICD codes and have been found to better align with diseases described in clinical practice [40]. We mapped ICD codes to PheCodes using the established PheWAS mapping [42]. We then compiled PheCode counts at the level of (1) integer codes (i.e. 296.11 $\rightarrow$ 296), which we refer to as parent phenotypes (i.e. COPD), and (2) 1-decimal codes (i.e. 296.11 $\rightarrow$ 296.1), which we refer to as subphenotypes (i.e. emphysema). We then ran MAP (including both Poisson and Lognormal mixture models) on these phenotype and subphenotype PheCode counts to estimate prior probabilities for each corresponding (sub)phenotype. For most diseases, patients with a corresponding PheCode count of 0 (i.e. filter-negative) rarely have the disease, so we set the corresponding topic prior $\pi_k = 0$ for these patients. Note that we do not include additional 'empty' topics to model information not contained in the PheCodes as we assume that including a topic for every PheCode sufficiently covers all phenotype information in each dataset.

Step 3: training MixEHR-G. MixEHR models EHR data using a generative latent topic model inspired by Latent Dirichlet Allocation [34, 41] (**Fig.** 1). We assume that each phenotype topic $k \in \{1, ..., K\}$ under EHR data type $t \in \{1, ..., T\}$ is characterized by a latent probability vector over $W^{(t)}$ EHR features, $\boldsymbol{\Phi}_k^{(t)} = [\phi_{w,k}^{(t)}]_{W^{(t)}}$, which follows a Dirichlet distribution with unknown hyperparameter β_{wt}. For each patient $d \in \{1, ..., D\}$, we assume that the patient's phenotype mixture membership θ_d is generated from a K-dimensional Dirichlet distribution with unknown asymmetric hyperparameters $\boldsymbol{\alpha} = [\alpha_k]_K$. Thus, to generate EHR observation j of datatype t for patient d, we first draw a latent topic $z_{dj}^{(t)}$ from a multinomial distribution with probability vector θ_d. Given $z_{dj}^{(t)}$, we then draw an EHR code $x_{d,j}^{(t)}$ from a multinomial distribution with probability vector $\boldsymbol{\Phi}_{z_{d,j}^{(t)}}^{(t)}$.

MixEHR-G extends MixEHR by setting the Dirichlet topic hyperparameters for patient d, $\boldsymbol{\alpha}_d$, to a scalar multiplied by the prior probability vector over the K phenotypes: $\boldsymbol{\alpha}_d = \alpha \boldsymbol{\pi}_d$, where $\boldsymbol{\pi}_d$ is obtained from Step 2. Consequently, patient d's expected mixture membership for phenotype k, $\alpha_{d,k}$ is scaled by patient d's prior probability of phenotype k, reflecting the

intuition that a higher prior likelihood of phenotype presence should beget a proportionally higher probability of feature attribution to that phenotype. Therefore, patient d will have zero mixture membership for phenotype k if the topic prior is equal to zero ($\pi_{d,k} = 0$).

We train MixEHR-G using the joint collapsed variational Bayesian inference protocol detailed in Li et al. (2020) [34]. The key variational topic inference step is:

$$\gamma_{d,j,k}^{(t)} \propto \left(\alpha_{d,k}^{*} + \tilde{n}_{d,.,k}^{-(d,j)}\right) \left(\frac{\beta_{x_{d,j}^{(t)}}^{(t)} + [\tilde{n}_{.,x_{d,j}^{(t)},k}^{(t)}]^{-(d,j)}}{\sum_w \beta_w^{(t)} + [\tilde{n}_{.,w,k}^{(t)}]^{-(d,j)}}\right) \tag{1}$$

where $\gamma_{d,j,k}^{(t)}$ denotes the variational probabilities of the kth topic assigned to EHR code j of patient d: $q(z_{d,j}^{(t)} = k) \equiv \gamma_{d,j,k}^{(t)}$ and the notation $n^{-(d,j)}$ indicates the exclusion of the contribution from patient d and EHR code j. The sufficient statistics are simply the summation of the inferred topic probabilities over all patients and all EHR codes except for patient d and EHR code j:

$$\tilde{n}_{d,.,k}^{-(d,j)} = \sum_{t=1}^{T} \sum_{j' \neq j}^{M_d^{(t)}} \gamma_{d,j',k}^{(t)}, \quad [\tilde{n}_{.,x_{d,j}^{(t)},k}^{(t)}]^{-(d,j)} = \sum_{d' \neq d}^{D} \sum_{j'=1}^{M_{d'}^{(t)}} [x_{d',j'}^{(t)} = x_{d,j}^{(t)}] \gamma_{d',j',k}^{(t)} \tag{2}$$

Unlike MixEHR, instead of iteratively optimizing the asymmetrical hyperparameters $\boldsymbol{\alpha}$, we now just optimize the scalar α. During the M-step of the EM learning algorithm, we update α by maximizing the marginal likelihood under the variational expectation via an empirical Bayes fixed-point update:

$$\alpha'_{d,k} \leftarrow \frac{a_\alpha - 1 + \alpha_{d,k} \sum_d \Psi(\alpha_{d,k} + \tilde{n}_{dk} + \tilde{m}_{dk}) - \Psi(\alpha_{d,k})}{b_\alpha + \sum_d \Psi(\tilde{n}_{dk} + \sum_k \alpha_{d,k}) - \Psi(\sum_k \alpha_{d,k})} \tag{3}$$

$$\alpha_{d,k}^{*} \leftarrow \pi_k \alpha'_{d,k} \frac{\sum_k \alpha'_{d,k}}{\sum_k \pi_k \alpha'_{d,k}} \tag{4}$$

The remainder of the algorithm is identical to that detailed in Li et al. (2020) [34] and thus omitted here. All functionality featured by MixEHR-G, including imputation of missing lab results, is preserved. When MixEHR-G converges, we obtain the count of total EHR features assigned to phenotype topic k for subject d as the patient-level phenotype mixture score: $\tilde{\theta}_{d,k} = \sum_j \gamma_{d,j,k}$. We use $\tilde{\boldsymbol{\Theta}} = [\tilde{\theta}_{d,k}]_{D \times K}$ for D patients over K topics for downstream

analyses.

2.2. Implementation of existing methods

We implemented sureLDA [30], MAP [26], and PheNorm [25] using their respective R packages: sureLDA: `https://cran.r-project.org/web/packages/sureLDA/index.html`; MAP: `https://search.r-project.org/CRAN/refmans/MAP/html/MAP.html`; PheNorm: `https://cran.r-project.org/web/packages/PheNorm/index.html`. We implemented MixEHR [34] using the source code from the github page: `https://github.com/li-lab-mcgill/mixehr`.

3. EHR datasets

3.1. EHR data overview

We applied MixEHR-G with phenotype topics defined by the set of K unique PheCodes (including both integer-level and 1-decimal codes in the PheCode hierarchy) to the PopHR and MIMIC-III datasets [37, 38]. Briefly, PopHR contains both chronic and acute disease information from outpatient visits as recorded in ICD codes, administrative treatment (ACT) codes, and RxNorm medication codes. MIMIC-III contains highly acute disease information from ICU admissions, often with comorbid or causative chronic diseases (e.g., flash pulmonary edema secondary to heart failure in a patient with chronic hypertension), as recorded in clinical notes, ICD-9 billing codes, prescriptions, DRG billing codes, CPT procedural codes, lab tests, and lab results. We identified $K = 1387$ PheCodes in PopHR (512 3-integer codes and 875 3-integer+1-decimal codes) and $K = 1515$ in MIMIC-III (499 3-integer codes and 1016 3-integer+1-decimal codes), reflecting the different clinical settings of the two datasets. Reassuringly, we found that patients tend to have multiple PheCode anchors in both datasets (a median of 14 integer codes and 16 1-decimal codes per patient in PopHR, and a median of 6 integer codes and 7 1-decimal codes per patient in MIMIC-III) **Supplementary Fig.** S1. This suggests that the PheCodes are sensitive but not overly specific, providing strong guidance for MixEHR-G's topics while allowing for learning and divergence from the anchor. We used these two datasets to demonstrate the generalizability of MixEHR-G and gain contrasting insights between the datasets regarding

the characterization of a common set of diseases. Besides the distinct population coverage of the two datasets, they also involve different feature modalities and levels of disease acuity (**Methods**).

Using MIMIC-III data as a demonstration (**Fig.** 1), we learned seven sets of topic distributions in the form of the basis matrices corresponding to each of the data modalities detailed above. Notably, lab data contribute two modalities corresponding to the binary indicators (i.e, whether the lab test was ordered) and lab test results coded as categorical variables with discrete values of normal and abnormal (more details below). We link these seven basis matrices using a common patient-topic mixture matrix, which is proportional to the EHR data multinomial likelihood and the prior inferred from the ICD-9 signatures under each PheCode.

3.2. Applications to the MIMIC-III dataset

MIMIC-III is a large, single-center database comprising multimodal EHR data associated with 53,423 distinct hospital admissions for 38,597 adult patients and 7,870 neonates admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012 [38]. The dataset was downloaded from the PhysioNet database (`mimic.physionet.org`) and used in accordance with the PhysioNet user agreement. For all data types, we removed nonspecific ICD, DRG, and CPT codes that were frequently observed among patients based on their inflection points. Clinical notes from `NOTEEVENTS.csv` were pre-processed and converted to bag-of-words format using the R library `tm`. We filtered out common stop words, punctuations, numbers, and whitespace, and we converted all remaining words to lowercase.

For lab data from `LABEVENTS.csv`, we utilized the FLAG column to ascertain whether a test result was normal or abnormal. For a given patient, we recorded the frequencies of both normal and abnormal lab results for each test.

For prescription data from `PRESCRIPTIONS.csv`, we concatenated the `DRUG_NAME_GENERIC`, `GSN`, and `NDC` columns to create a compound ID for each prescription. We combined entries with the same compound ID to eliminate feature redundancy by drug formulation. Likewise, for diagnosis-related group (DRG) codes from `DRGCODES.csv`, we created a

compound ID as the concatenation of `DRG_TYPE`, `DRG_CODE`, `DESCRIPTION`, `DRG_SEVERITY`, and `DRG_MORTALITY`. We kept the original IDs for ICD-9 diagnosis codes (`ICD-9-CM`) from `DIAGNOSES_ICD.csv` and procedure codes (`ICD-9-CPT`) from `PROCEDURES_ICD.csv`.

Our pre-processed MIMIC-III dataset consisted of 33,388 unique tokens from clinical notes, 6,215 ICD-9 codes, 1,770 CPT codes, 564 lab tests, 8,409 prescriptions, and 3,086 DRG codes, altogether totaling 53,432 unique features.

The MIMIC-III data are not longitudinal. Most patients in the MIMIC-III dataset only have one ICU admission, wherein each ICD-9 code is observed at most once upon discharge. As a result, each PheCode is observed at most once for each patient. In lieu of running MAP on PheCode counts to obtain prior phenotype probabilities as detailed in *Step 2: Initializing prior probabilities using MAP*, we set the hyperparameters $\alpha_{d,k}$ for phenotype k to one or zero depending on whether the associated PheCode was observed for patient d or not, respectively.

3.3. Applications to the PopHR dataset

PopHR is a multimodal database that integrates massive amounts of longitudinal heterogeneous data from multiple distributed sources (e.g. inpatient and outpatient physician claims, hospital discharge abstracts, outpatient drug claims) for 1.3 million patients in the Quebec province of Canada between 1998 and 2014 [37]. We used three data modalities from this dataset: ICD-9 diagnostic codes, administrative procedure codes (ACT), and prescriptions all from outpatient physician encounters. For prescriptions, we used drug identification numbers (DINs) as unique IDs; for the other two we used ICD-9 and ACT codes respectively. We removed all features with document frequencies above 25% to mitigate the influence of common, imprecise features. This left 5,935 unique ICD-9 codes, 4,037 ACT codes, and 1,311 DIN codes for a total of 11,283 unique count features. Some DIN codes indicate different dosages or formulations of the same active drug ingredient. We found this to be useful information for modeling phenotypes as the same drug is often prescribed differently for different conditions. However, for qualitative topic analyses such as the heatmap in **Fig.** 2a, we aggregated topic probabilities for DIN codes by active drug ingredient to simplify interpretation.

Unlike MIMIC-III, PopHR is a longitudinal dataset in which a subject can encounter multiple occurrences of any given ICD-9 code (and thus PheCode) over time. Thus, we computed prior phenotype probabilities using the MAP procedure described in *Step 2: Initializing prior probabilities using MAP*.

4. Experimental design

4.1. Phenotyping accuracy evaluation

For the PopHR data [37], we applied expert-defined rule-based phenotyping algorithms for 12 chronic diseases. We selected these 12 diseases - schizophrenia, ischemic heart diseases (IHD), hypertension, HIV, epilepsy, diabetes, COPD, congestive heart failure (CHF), autism, asthma, acute myocardial infarction (AMI), and ADHD - as they were the only phenotypes with clear PheCode alignments for which we had gold-standard rule-based algorithms. For 9 of these phenotypes - acute myocardial infarction (AMI), asthma, congestive heart failure (CHF), COPD, diabetes, hypertension, ischemic heart disease (IHD), epilepsy, and schizophrenia - rules were defined by the Chronic Disease Surveillance Division of the Public Health Agency of Canada [43]. For the remaining 3 phenotypes, for ADHD we used the definition described in Vasiliadis et al. (2017) [44], for autism we followed the 2017 Autism Spectrum Disorder Surveillance in Quebec report [45], and for HIV we applied the definition described in Durand et al. (2011) [46].

All rules are derived entirely from ICD codes, sometimes relying on consecutive observations of the same ICD code(s) for a patient within a small (i.e. 1 month) time interval. Precise definitions for each rule-based phenotype can be found in the **Supplementary Materials**. Since we aggregated ICD code frequencies over time for a given patient, our model does not use the same temporal information as the rules, making them a good reference for MixEHR-G. Therefore, we used the phenotype labels derived from the rules as gold-standards to evaluate our unsupervised model performance. We do not apply these rule-based phenotyping algorithms to the MIMIC-III dataset as they were intended to identify chronic diseases using longitudinal data, whereas MIMIC-III only contains data for hyperacute ICU visits.

For both the MIMIC-III and PopHR datasets, we ran MixEHR-G on parent phenotypes and subphenotypes separately. We then derived the MixEHR-G phenotype score $\tilde{\boldsymbol{\Theta}}_k$ for a coarse phenotype such as COPD (496) using the corresponding topic from the parent phenotype model, whereas for subphenotypes such as emphysema (496.2) we used the subphenotype model.

4.2. Identification of similar and comorbid phenotypes

We defined the *similarity* between two phenotypes k and k' as the Spearman correlations between their respective MixEHR-G topic distributions $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Phi}_{k'}$ over the V unique EHR codes across data modalities. We defined the *comorbidity* between phenotypes k and k' as the Spearman correlations between their respective patient mixture proportions $\boldsymbol{\theta}_k = [\theta_{dk}]_D$ and $\boldsymbol{\theta}_{k'} = [\theta_{dk'}]_D$. These choices reflect the intuition that two phenotype topics with similar distributions over EHR features are qualitatively similar, whereas two topics with similar mixture probabilities over patients are comorbid. Thus, for each phenotype of interest, we use the trained $\boldsymbol{\Phi}_{K\times V}$ matrix to identify similar *parent* phenotypes, and we use the trained $\boldsymbol{\theta}_{D\times K}$ matrix to identify comorbid *parent* phenotypes.

4.3. Estimation of relative population phenotype prevalence stratified by age and sex

For a given phenotype k, we set the overall phenotype prevalence ρ_k to the estimate from the Global Health Data Exchange, a comprehensive catalog of health-related data compiled by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington [47]. We sought to estimate the relative prevalence stratified by age and sex using the PopHR dataset, which is representative of the general population (**Fig.** 7). First, we set a threshold $\tau^* \geq 0$ to the patient-topic mixtures $\tilde{\boldsymbol{\theta}}_k$ to derive predicted phenotype outcomes $[\hat{Y}_{dk}]_D$ with mean value equal to the known population prevalence ρ_k. In other terms, we denote the empirical quantile function of $\tilde{\boldsymbol{\theta}}_k$ as $\hat{Q}_{\tilde{\boldsymbol{\theta}}_k}(x)$,

$$\tau^* = \hat{Q}_{\tilde{\boldsymbol{\theta}}_k}(\rho_k) \text{ where}$$

$$\hat{Y}_{dk} = \begin{cases} 1 & \text{if} \quad \tilde{\theta}_{dk} \geq \tau^* \\ 0 & \text{otherwise} \end{cases}$$

For a given age t_{age}, we identified the set of patients $d \in \mathcal{S}_{t_{age}}$ such that $T_{min,d} \leq t_{age} \leq T_{max,d}$, where $T_{min,d}$ and $T_{max,d}$ denote the ages at which patient d respectively enters and exits the PopHR dataset. We then estimated the population phenotype prevalence at age t_{age} as the mean predicted phenotype outcome over $\mathcal{S}_{t_{age}}$:

$$\hat{\rho}_{k,t_{age}} = \frac{1}{|\mathcal{S}_{t_{age}}|} \sum_{d \in \mathcal{S}_{t_{age}}} \hat{Y}_{dk}$$

Recall that we only aimed to estimate relative, not absolute, prevalence over time. Given some predefined sequence of ages $\mathcal{T}$ subscripted by t, we estimated the relative prevalence at time t_{age} as

$$\hat{\rho}_{t_{age},rel} = \frac{\hat{\rho}_{t_{age}}}{\sum_{t \in \mathcal{T}} \hat{\rho}_t}$$

We compared our relative prevalence estimations to those from the IHME Global Health Data Exchange [47]. Again, we computed relative prevalence rates from the absolute prevalence rates supplied by the data exchange by normalizing ρ_k over time.

5. Results

5.1. MixEHR-G's phenotype topics are clinically meaningful

We qualitatively explored the inferred phenotype topics from both the PopHR and MIMIC-III datasets by examining the EHR features with the highest topic distribution probabilities ($\mathbf{\Phi}_k$'s) for 9 diverse chronic diseases covering a variety of medical specialities (**Fig.** 2; **Supplementary Fig.** S4). We include top 5 features for all parent phenotypes and subphenotypes in the **Supplementary Materials**. We observe that in PopHR the phenotype topics appear to be predominantly defined by prescription (Rx) codes despite the dataset consisting of 52.6% ICD-9 codes and only 11.6% Rx codes (**Fig.** 2a). Meanwhile, MIMIC-III phenotypes are dominated by ICD-9 codes (**Fig.** 2b) despite that dataset consisting of 63.2% words from notes, 15.9% Rx codes, and only 11.8% ICD-9 codes. This highlights MixEHR-G's ability to leverage diverse EHR modalities to infer phenotype topics

in a data-driven manner as opposed to relying solely on the ICD-9 signatures for each PheCode or relying predominantly on prevalent features. In contrast, the top features identified by sureLDA are mostly ICD-9 codes for PopHR and words from notes for MIMIC-III - in both cases the dataset's most prevalent feature modality (**Supplementary Fig.** S3).

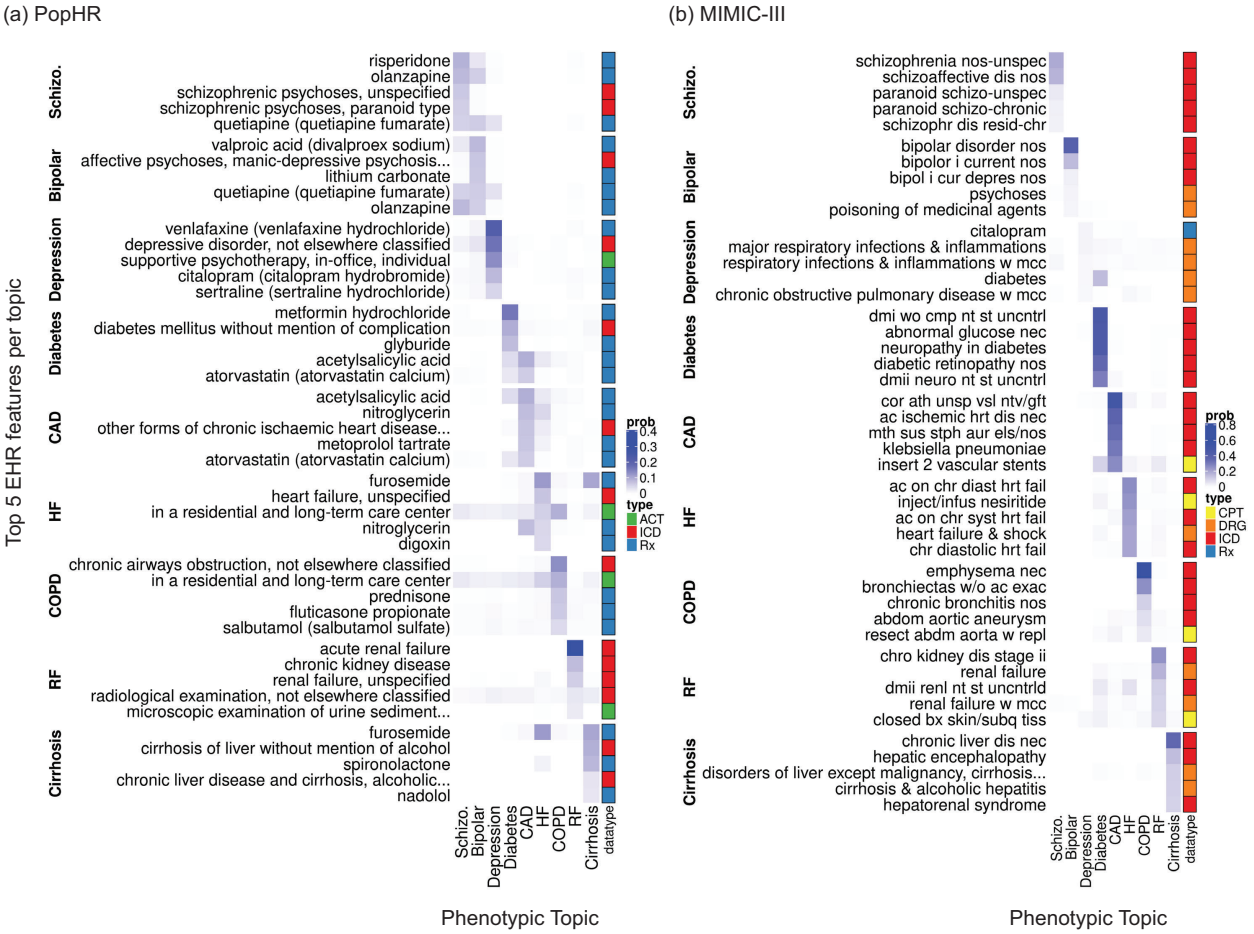


Figure 2: **Top 5 features for each of 9 diverse disease phenotypes.** (**a**) Phenotype distributions inferred from the PopHR dataset. Using MixEHR-G, we inferred phenotype distributions (ϕ) for all PheCode-guided parent phenotypes and subphenotypes and displayed 9 of them. We include distributions for all phenotypes in the Supplementary Materials. Phenotype distributions are inferred over EHR features of 3 data types (ICD, ACT, Rx). (**b**) Phenotype distributions inferred from the MIMIC-III data. Phenotype distributions were inferred over EHR features of 6 data types (CPT, DRG, ICD, Rx, clinical notes, and lab data). Heatmap color intensities reflect the probabilities of features under a given phenotype. The color bar on the right side of each heatmap indicates the data types as shown in the legend. Only data types corresponding to the top EHR codes under each phenotype are visible in the heatmaps. Some row names were truncated due to length. In panel (a) the truncated names are 'affective psychoses, manic-depressive psychosis, other and unspecified', 'other forms of chronic ischaemic heart disease, unspecified', 'microscopic examination of urine sediment and interpretation', 'chronic liver disease and cirrhosis, alcoholic cirrhosis of liver'; in panel (b) the truncated name is 'disorders of liver except malignancy, cirrhosis, alcoholic hepatitis with complications, comorbidities'. The full names for the abbreviated column names are as follows: Schizo.: Schizophrenia; Bipolar: Bipolar Disorder; CAD: Coronary Artery Disease; HF: Heart Failure; COPD: Chronic Obstructive Pulmonary Disease; and RF: Renal Failure.

Qualitatively, MixEHR-G consistently identifies clinically meaningful features for each phenotype topic (**Fig.** 2a). For instance, risperidone and olanzapine are common antipsychotic medications used to treat schizophrenia (and to a lesser extent bipolar disorder), and furosemide is a common diuretic used for treatment of decompensated heart failure and cirrhosis, among other conditions. Likewise, MixEHR-G identifies meaningful associations between phenotype topics and lab results (**Supplementary Fig.** S2). For instance, the schizophrenia and bipolar topics exhibit high odds ratios for abnormal blood valproic acid, a common mood stabilizer, while the COPD topic is found to exhibit high odds ratios for abnormal urine chloride and potassium, reflecting the well known chronic compensatory metabolic alkalosis of these patients. Meanwhile, the COPD topic exhbits low odds ratios for abnormal troponin-1, a marker of ischemic cardiac injury, and NT-proBNP, a marker of heart failure - both diagnoses commonly ruled out in patients presenting with COPD exacerbation. Together these results suggest that while MixEHR-G is anchored to a prior defined by PheCodes, it is able to infer phenotype topics by leveraging other data modalities, including prescriptions and labs. Therefore, MixEHR-G's phenotype topic distributions could be used to augment rule-based phenotyping algorithms and alleviate tedious manual phenotyping efforts.

We also observe qualitatively meaningful disease similarities and comorbidities based on the topic distributions (Φ) and patient-topic mixtures ($\tilde{\Theta}$) (**Fig.** 3). In particular, schizophrenia, bipolar disorder, and depression exhibit similar topic distributions with each other as well as with personality, anxiety, and mood disorders (**Fig.** 3a). Phenotype pairs with known causative relationships also exhibit similar distributions, including diabetes and renal failure, heart failure and pulmonary edema, and cirrhosis and ascites. Interestingly, our analysis also revealed complex disease network connections among some phenotypes. For instance, cirrhosis is found to have a high degree of similarity with renal failure and nonhypertensive congestive heart failure, both of which also cause edema, as well as viral hepatitis, which also causes acute liver failure.

Relative similarity in terms of the patient-topic mixtures ($\tilde{\Theta}$) also replicates the well-known comorbidities among schizophrenia, mood disorders, and anxiety disorders, which have long been described in the medical literature [48–50] (**Fig.** 3b). Other meaningful

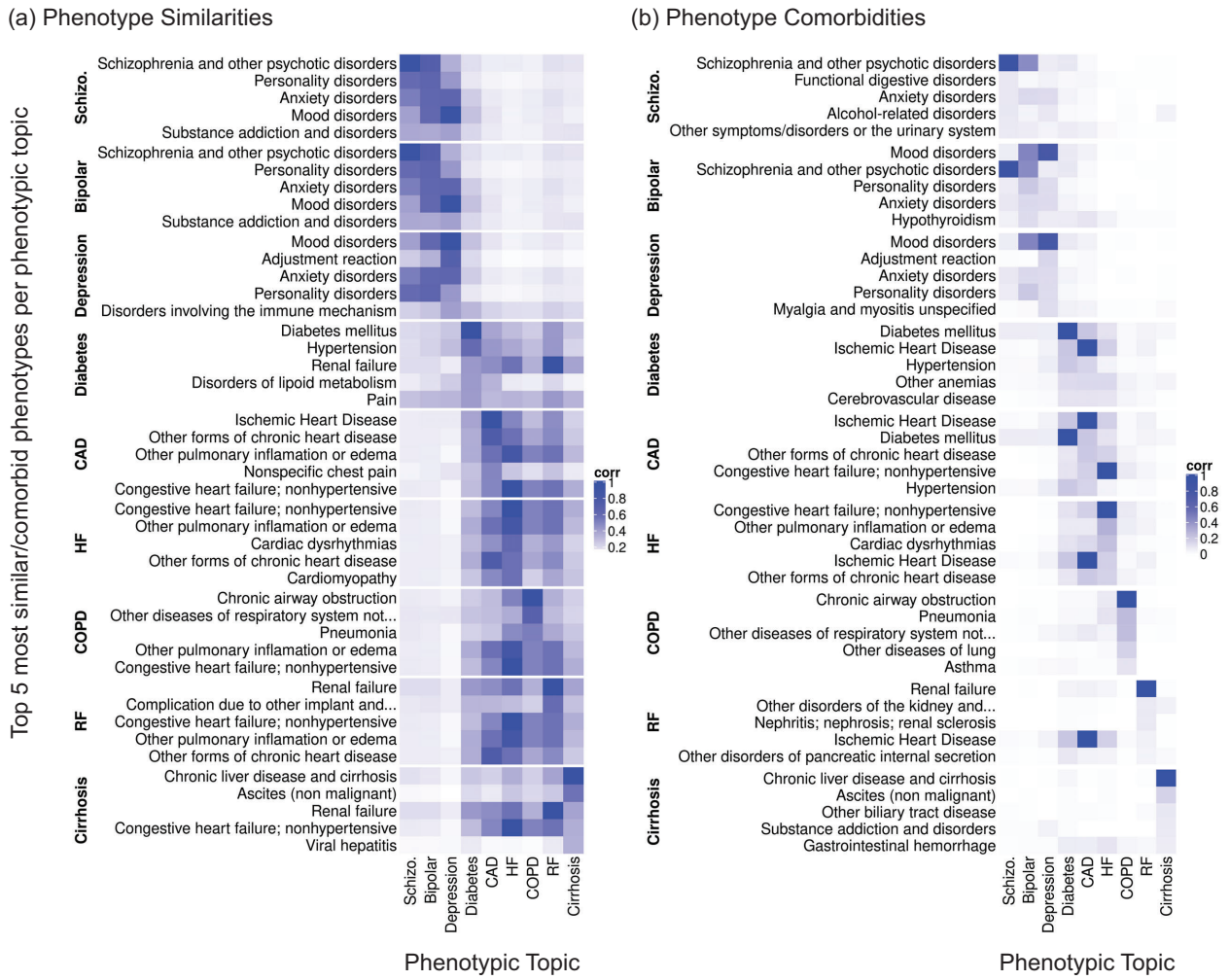


Figure 3: **Predicted similarities and comorbidities of 9 diverse phenotypes.** (**a**) 5 most similar parent phenotypes for each of 9 diverse disease phenotypes, as predicted by *post-hoc* Spearman correlations among MixEHR-G's topic distributions Φ over all multimodal EHR features. (**b**) 5 most comorbid phenotypes per target disease as predicted by *post-hoc* Spearman correlations among MixEHR-G's patient-topic mixtures $\hat{\Theta}$. MixEHR-G was trained on the PopHR dataset. Heatmap color intensities are proportional to the Spearman correlation values. The color bar on the right side of each heatmap indicates the data types as shown in the legend. Only data types corresponding to the top EHR codes under each phenotype are visible in the heatmaps.

combinations include diabetes with ischemic heart disease and hypertension, as well as cirrhosis with ascites, biliary tract disease, and substance disorder. More interestingly, we observe comorbidities between bipolar disorder and hypothyroidism, concurrent with several recent studies positing a potential association between the two [51–53]. Similarly, the association between depression and myalgia/myositis supports the hypothesized clinical association between chronic pain/inflammation and depression [54, 55]. Notably, no two parent phenotypes have particularly strong associations (Spearman correlations <

0.4), reflecting the fact that even comorbid phenotypes like anxiety and mood disorders are very much independent entities.

5.2. Automated phenotyping using MixEHR-G's topic mixture

A salient feature of MixEHR-G is that the K PheCode-guided topics are readily identifiable. Consequently, they can be directly used as *phenotype risk scores (PheRS)* [56]. Compared to alternative phenotyping algorithms, we found that MixEHR-G's $\tilde{\Theta}$ more closely aligns with validated rule-based phenotypes compiled by clinical experts at the Chronic Disease Surveillance Division of the Public Health Agency of Canada and elsewhere (**Methods**; **Fig.** 4; **Table** 1) [57]. In particular, MixEHR-G's scores exhibited significantly higher AUPRCs than the raw PheCodes or the recently published unsupervised phenotyping algorithms Multimodal Automated Phenotyping (MAP) [26] and sureLDA [30]

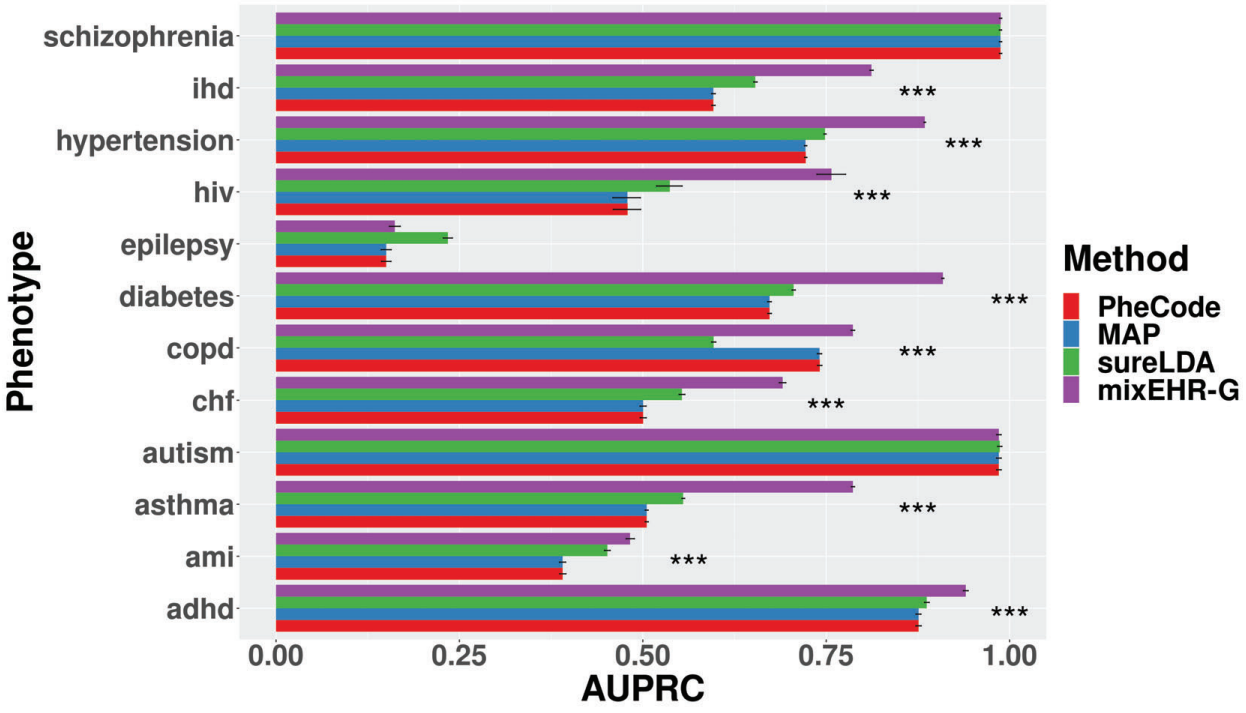


Figure 4: **Predictive performance of $\tilde{\theta}_k$ for phenotypes with expert-derived rule-based labels.** We applied MixEHR-G and two recently published unsupervised phenotyping methods, MAP and sureLDA, to the entire PopHR data without supervision or feature filtering. We also computed the frequency of the raw PheCode counts for each target disease. The barplots display the areas under the precision recall curve (AUPRCs) within the PopHR dataset taking rule-based phenotype labels as gold-standards. 95% confidence intervals and standard errors were estimated by bootstrapping with 1000 subsamples. P-values were estimated using a two-sample t-test with bootstrapped standard error estimates. (*): p < 0.05; (**): p < 0.01; (***): p < 0.001.

	Prevalence	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1
ADHD	0.02	1.00	0.94	1.00	1.00	0.88	1.00	0.93
Acute MI	0.03	0.84	0.48	0.70	0.98	0.52	0.99	0.60
Asthma	0.10	0.95	0.79	0.81	0.95	0.63	0.98	0.71
Autism	0.00	0.99	0.99	0.99	1.00	1.00	1.00	0.99
CHF	0.04	0.89	0.69	0.78	0.98	0.61	0.99	0.69
COPD	0.07	0.89	0.79	0.79	0.99	0.89	0.99	0.84
Diabetes	0.08	0.96	0.91	0.94	0.96	0.71	0.99	0.80
Epilepsy	0.01	0.60	0.16	0.20	1.00	0.67	0.99	0.31
HIV	0.00	0.92	0.76	0.84	1.00	0.57	1.00	0.68
HTN	0.19	0.94	0.88	0.86	0.95	0.81	0.97	0.83
IHD	0.09	0.94	0.81	0.91	0.95	0.64	0.99	0.75
Schizo.	0.01	0.99	0.99	0.99	1.00	1.00	1.00	0.99

Table 1: **Accuracy metrics of MixEHR-G's topic mixture scores $\tilde{\Theta}$ for identification of 12 diverse phenotypes.** For computation of sensitivity, specificity, PPV, NPV, and F1 score, scores were thresholded to achieve NPVs of at least 0.95.

for 9 out of 12 phenotypes evaluated (IHD, hypertension, HIV, diabetes, CPD, CHF, asthma, AMI, and ADHD) at level 0.05. For the remaining 3 phenotypes, either the PheCode already achieves near-perfect accuracy and thus leaves no room for improvement (i.e. schizophrenia, autism), or the phenotype's prevalence is sufficiently low that MixEHR-G's topic remains highly anchored to the PheCode prior (i.e. epilepsy).

MixEHR-G's improvement over sureLDA was particularly notable despite the similarity between the two algorithms, suggesting that MixEHR-G's multimodal topic design and efficient collapsed variational inference algorithm provide tangible benefits over sureLDA's unimodal topics and Gibbs sampling training procedure [30, 34]. The only phenotype for which sureLDA outperformed MixEHR-G was epilepsy, for which no phenotyping algorithm aligned particularly well with the rule-based labels (i.e. all AUPRCs were below 0.25). Nonetheless, MixEHR-G consistently equaled or outperformed the raw PheCode counts and MAP. Since both PheCodes and MAP only draw upon a handful of ICD-9 codes per phenotype, the superior performance of MixEHR-G further corroborates the benefit of leveraging the vast amount of additional information in patients' EHRs.

To investigate the concordance and discrepancy between MixEHR-G's inferred topics and the existing phenotyping rules, we focused within the PopHR dataset on chronic obstructive pulmonary disease (COPD), a notoriously difficult phenotype to define with

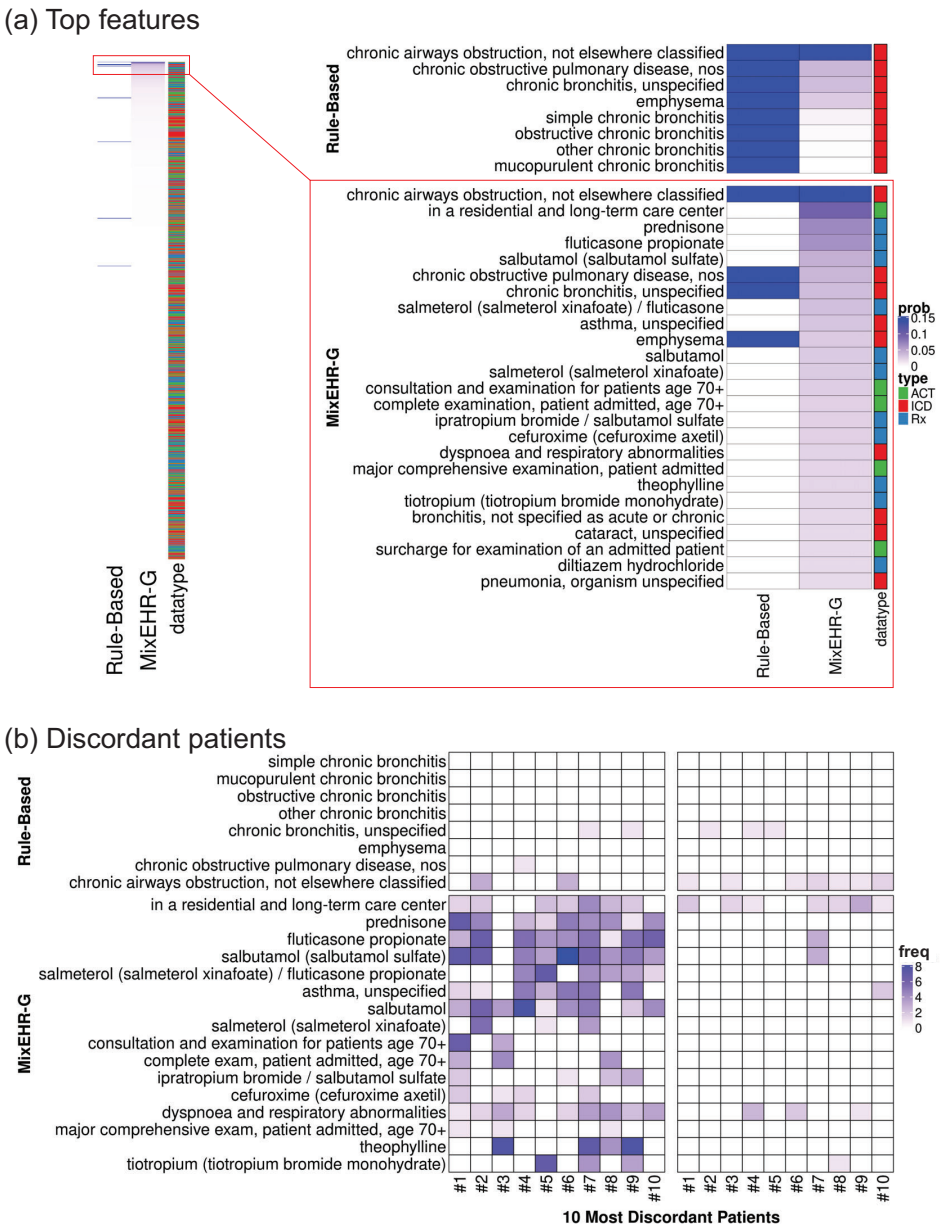


Figure 5: **In-depth analysis of MixEHR-G's phenotyping results for chronic obstructive pulmonary disease (COPD) using the PopHR dataset.** (**a**) All 11,283 PopHR features (left) and the top 25 EHR features (right) with corresponding data types (color coded) per the COPD topic distribution Φ_{COPD}. We compared these features with the features included in the expert-curated rule-based COPD algorithm obtained from the Chronic Disease Surveillance Division of the Public Health Agency of Canada [43] described in the **Methods**. Features included in the rule-based algorithm are shown in the heatmap on top. (**b**) Profiles of the 20 most discordant patients. The heatmap on the left displays the 10 patients for whom MixEHR-G predicts the highest mixture scores for the COPD topic ($\tilde{\theta}_{\text{COPD}}$) but the rule-based algorithm predicts them as negative. The heatmap on the right displays the 10 patients for whom MixEHR-G predicts the lowest COPD topic mixture scores but the rule-based algorithm predicts them as positive. The color intensity indicates the frequency count of each EHR code for the selected patients.

rules (**Fig.** 5; **Supplementary Fig.** S5) [58]. We first examined the top 25 codes under the COPD topic's distribution $\boldsymbol{\phi}_{\text{COPD}}$ (**Fig.** 5a; **Supplementary Fig.** S5). Four out of eight ICD codes that constitute the COPD rule were present among the top 10 features under the COPD topic distribution. However, MixEHR-G also incorporates other meaningful features including prednisone (commonly used for treatment of COPD exacerbations), salbutamol (used for asthma and COPD maintenance but more commonly COPD), salmeterol (also used for asthma and COPD maintenance but more commonly COPD), and tiotropium (used almost exclusively for COPD maintenance). Indeed, MixEHR-G identifies COPD patients using all available information in the EHR as opposed to just the ICD-9 codes defined by the rule.

We further examined the top 10 rule-negative patients with the highest MixEHR-G COPD PheRS scores (**Fig.** 5b left panel) as well as the top 10 rule-positive patients with the lowest MixEHR-G COPD PheRS scores (**Fig.** 5b right panel). Many of the patients in the former category have numerous observations of prednisone, fluticasone, salbutamol, salmeterol, theophylline, or tiotropium, suggesting that they may indeed be COPD patients (i.e. potential false negatives for the rule). Conversely, many low-scoring patients in the latter category exhibit no evidence of any relevant prescriptions, making them dubious COPD cases (i.e. potential false positives). These results suggest that the predefined disease validation rules may be too rigid since no ICD code or prescription is adequately sensitive or specific. On the other hand, MixEHR-G provides richer and largely complementary information to help re-define these rules in a data-driven manner. To formally assess this, we would need chart-reviewed gold-standard labels, which are not available from our current data.

5.3. Estimating disease prevalence rates in Quebec using MixEHR-G's topic mixture

We used MixEHR-G's patient-phenotype mixtures $\tilde{\boldsymbol{\Theta}}$ to identify which phenotypes are most strongly associated with age and sex (**Fig.** 6). We observe that hypertension and osteoporosis are strongly positively associated with age, whereas otitis media and tonsilitis are strongly negatively associated. Likewise, pathologies of the prostate are associated with male sex, whereas breast and cervical pathologies are associated with female sex.

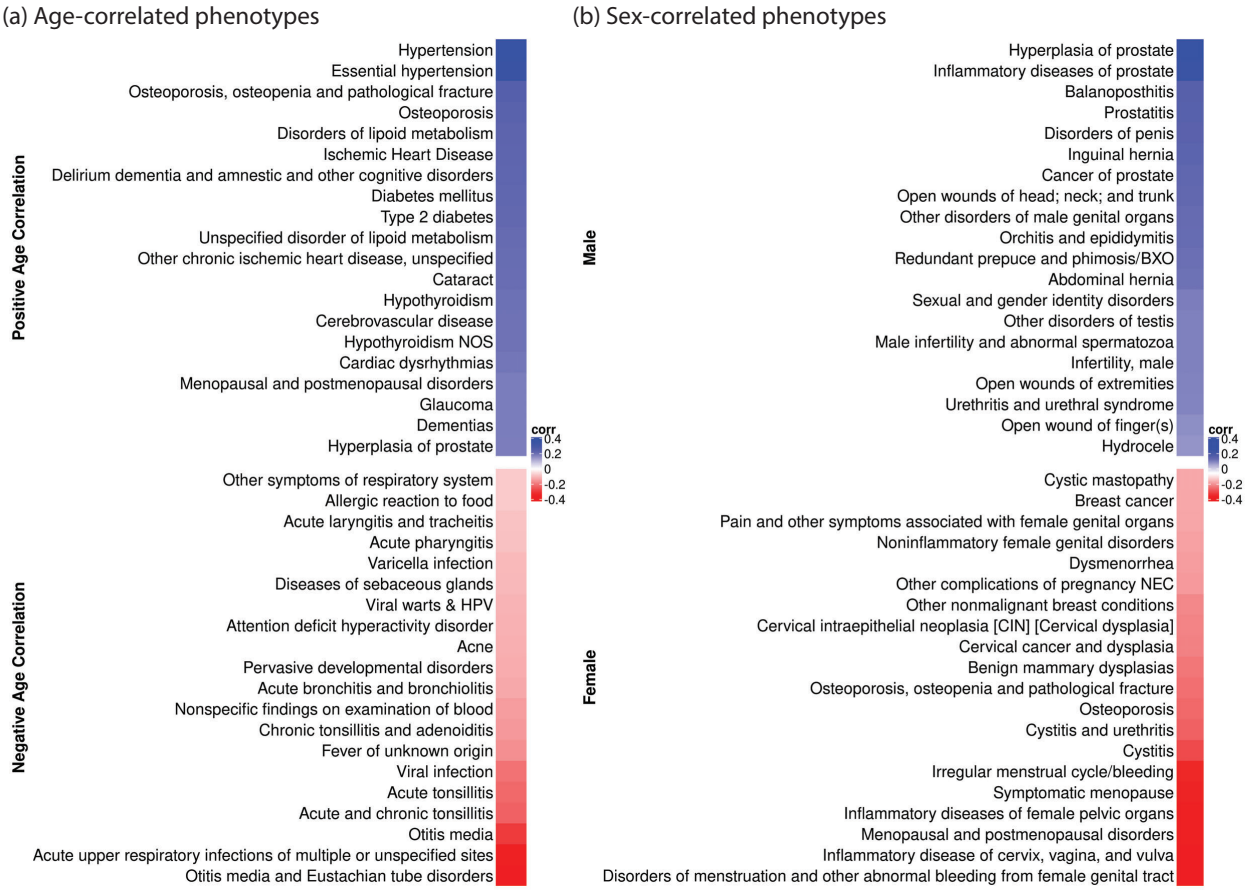


Figure 6: **Top 20 phenotypes correlated with age and sex as predicted in our analysis.** (**a**) The 20 phenotypes with the most positive (top) and most negative (bottom) correlations to patient age. (**b**) The top 20 sex-correlated phenotypes. For each panel, we computed Spearman correlations between MixEHR-G's patient-topic mixtures $\tilde{\Theta}$ and patients' (a) mean age or (b) sex, taking MALE as SEX=1.

These simple proof-of-concepts paved the way for further exploratory epidemiological analyses.

We used $\tilde{\Theta}$ to investigate the relative prevalence for 8 diverse phenotypes stratified by age and sex, which we compare to gold-standard estimates from the Global Health Data Exchange (GHDx) [47] (**Fig.** 7; **Methods**). We observe high concordance between our estimates and the gold-standard. Consistent with the gold-standard, our analysis indicates that asthma peaks in prevalence around age 10; leukemia and lung cancer around ages 65-75; Parkinson's disease, colon cancer, and melanoma around ages 75-85, and COPD and Alzheimer's disease just increase monotonically with age. Likewise, MixEHR-G's estimates support the inference that asthma, COPD, Parkinson's disease, colon cancer, and melanoma are significantly more prevalent in men, whereas Alzheimer's disease (AD)

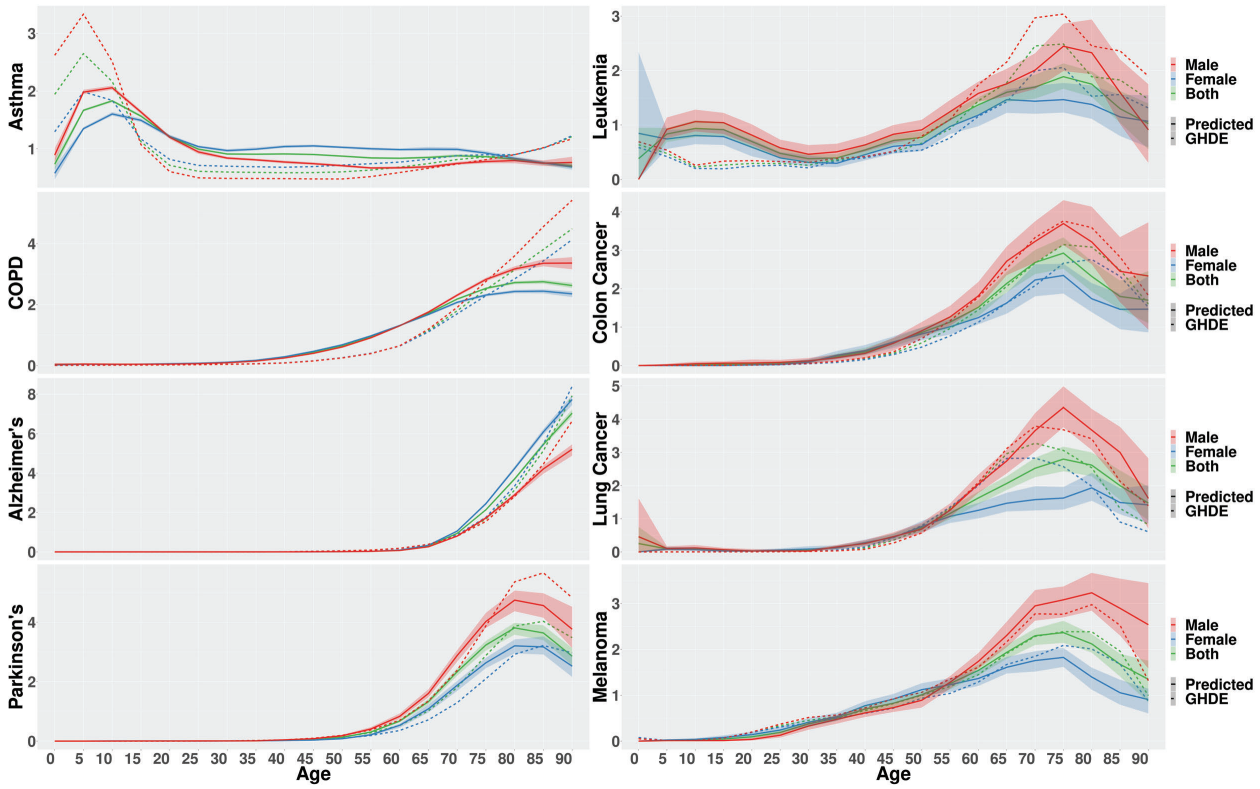


Figure 7: **Estimates of relative phenotype prevalence.** Based on the patient-phenotype mixtures $\tilde{\mathbf{\Theta}}$, we computed phenotype prevalence stratified by age and sex for 8 diverse disease phenotypes with variable natural functions normalized by the mean estimate over time (**Methods**). We compared our predicted prevalences MixEHR-G (solid lines) with the estimates from the Global Health Data Exchange (GHDE, dotted lines). 95% confidence intervals for MixEHR-G's estimates were computed by bootstrapping with 100 replicates.

predominates in women, likely reflecting increased survivorship among female Alzheimer's patients. These diseases were arbitrarily chosen as a proof-of-concept based on our knowledge about the diverse onsets of these diseases and the availability of prevalence data from GHDE. We leave a more comprehensive analysis of the entire phenome as future work.

6. Discussion

The effective application of machine learning algorithms to EHR data will lay the foundation for the development of modern digital medicine. One important and fruitful application is predicting a patient's disease phenotypes using a supervised learning model such as logistic regression or neural network [59]. However, supervised learning methods are not scalable for simultaneously predicting thousands of phenotype labels [18]. Model-

based unsupervised methods such as non-negative matrix factorization [60], autoencoder [31], and topic modeling [34] are at the other end of the spectrum. While the latent factors or topics inferred by these unsupervised methods can provide clinical insights, they are not identifiable as inferred topics cannot be directly mapped to known phenotypes. Moreover, while supervised topic models such as MixEHR-S [61] can simultaneously infer topic distributions and fit a predictive function of a target disease, they are not scalable to predicting multiple disease labels simultaneously. Therefore, an efficient method is needed to achieve simultaneous phenotype inference while providing interpretable topic distributions over heterogeneous EHR data.

In this study, we present a PheCode-guided topic modeling algorithm called MixEHR-G to simultaneously model up to 1515 well-defined phenotypes as a function of high-dimensional, multimodal EHR data. MixEHR-G is highly scalable due to its use of memory-efficient sparse matrices and time-efficient variational Bayesian inference implemented with parallelization in C++. Indeed, it can perform inference on 1.3 million patient records containing a total of 70.7 million non-zero feature observations over 24,192 unique features in $\sim$10 hours on a 2.2 GHz processor with 10 cores, using a maximum of 60 MB of RAM and 130% CPU (1.3 cores).

In proof-of-concept analyses, we observe that the MixEHR-G-inferred topics are well-aligned with the phenotypes they represent and complementary to rule-based phenotyping algorithms. We used MixEHR-G trained on the PopHR database to derive meaningful insights about disease similarities, comorbidities, and prevalences in the Quebec population [62].

We envision two major applications for the trained MixEHR-G model. The first application is phenotype labeling for cohort selection or clinical research. MixEHR-G more closely aligns with expert-curated phenotype rules than other recently published phenotyping methods including MAP and sureLDA across 9 diverse phenotypes (**Fig.** 4). Our case study on COPD suggests that effectively incorporating prescription data produces more informative rules (**Fig.** 5). For a large-scale downstream study that requires labels across many phenotypes, such as a Phenome-Wide Association Study (PheWAS), MixEHR-G's patient-topic mixtures $\tilde{\Theta}$ can be directly used as continuous phenotype risk scores (PheRS)

[56], obviating the need for labor-intensive chart review or rule generation.

For a study focusing on a single phenotype, MixEHR-G's topic distributions $\boldsymbol{\phi}$ can be further refined with subdivided groups of specific EHR codes for sub-phenotyping [56]. When associating these sub-phenotypes with genotype data using population data such as UK Biobank [63], we can investigate the genetic epidemiology of the refined sub-phenotypes and gain insights into the diverse outcomes of the same primary disease (e.g., drug response, survival time). These phenotype labels can then be used for downstream epidemiological research or to identify cohorts for inclusion in a clinical trial.

For patient outcome classification, the inferred phenotype topic probabilities $\tilde{\boldsymbol{\Theta}}$ can be treated as phenotype risk scores or PheRS [56] derived from the training patient dataset. As a post-processing step, they can then be incorporated as strongly predictive features into a supervised classifier for applications such as short-term mortality prediction [34, 64], sequential diagnosis prediction [34], and first-time drug usage (e.g., insulin regime adoption by diabetic patients [61]). Given the interpretability of the topics, we can then use them to understand the target outcome based on their corresponding importance scores derived from the trained classifiers.

MixEHR-G's second major application is to derive population-level insights about phenotypes. In this study, we validated the use of MixEHR-G's patient-topic mixtures $\tilde{\boldsymbol{\Theta}}$ to infer relative phenotype prevalence stratified by age and sex. For prevalence estimation specifically, MixEHR-G's estimates can either elucidate the general shape of a prevalence function over a stratifying variable such as age, or be calibrated using a small set of gold-standard labels and a validated risk estimation method such as SCORNET [65] to produce valid absolute prevalence estimates. Moreover, MixEHR-G's outputs can identify potential epidemiological directions for more rigorous follow-up. For instance, our analysis of associations in $\boldsymbol{\phi}$ supports the recently hypothesized comorbidity between bipolar disease and hypothyroidism (**Fig.** 3b); ascertaining the nature of this association could be an interesting subject for further investigation. On the other hand, we can incorporate the prior knowledge of the phenotypic prevalence by sex and age (e.g., from Global Health Data Exchange) into the Bayesian topic prior for each patient (i.e., the Dirichlet topic hyperparameters). Therefore, depending on the age and sex of patient d, we will have a

different set of prior weights over his/her topic mixture membership θ_d. This will presumably improve phenotyping accuracy especially for small sample size and sparse observation in the EHR data.

As a future direction, we will extend MixEHR-G to model 1) continuous data modalities, and 2) time. In its current form, MixEHR-G must threshold continuous variables such as lab tests into categorical values, and it rolls up observations over time into feature counts to infer longitudinal patient parameters. While many continuous clinical variables can be safely reduced to low-dimensional categoricals without significant loss of information (i.e. certain lab results being represented as 'normal', 'high', or 'low'), many other variables (i.e. age) cannot be easily reduced in this way. Moreover, being able to model the progression of patients' phenotype topics over time could yield valuable insights about longitudinal risk, the natural course of diseases, and causal relationships. We leave these as future extensions for the MixEHR model family.

We also envision MixEHR-G being modified to handle multi-center data such that it can simultaneously incorporate EHR data across medical centers. In MixEHR-G's current form, incorporating multi-center data would be challenging with potential biases given meaningful differences in not just patient populations but also how diseases are coded and treated. Being able to leverage commonalities across centers in a manner similar to mixed effects models could yield powerful insights on a scale larger than any individual medical center.

7. Conclusion

In summary, by modeling multimodal phenotypes with the guided topic modeling strategy, MixEHR-G achieves accurate simultaneous phenotype predictions for up to 1515 phenotypes and improved interpretability of phenotype topics. Altogether, we believe that MixEHR-G will enable more efficient use of EHR data toward the goals of personalized digital medicine and precision public health.

Ethics

The use of the PopHR data for this research was approved by the McGill Faculty of Medical and Health Sciences Institutional Review Board.

Acknowledgements

Y.L. is supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2019-0621), Fonds de recherche Nature et technologies (FRQNT) New Career (NC-268592), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative New Investigator start-up award (G249591). D.B. is supported by a Canada Research Chair (Tier 1) in Health Informatics and Data Science. The authors declare no conflicts of interest.

Author contributions

Y.A., D.B., and Y.L. conceived the study. Y.A. and Y.L. developed the methodology and implemented the software. Y.A. ran the majority of the data analyses with help from Y.Z. and A.V.. Y.A., D.B., and Y.L. analyzed the results. Y.A. and Y.L. wrote the initial draft of the manuscript with critical comments from D.B. All of the authors reviewed and wrote the final version of the manuscript.

References

[1] I. Kohane, S. Churchil, and S. Murphy, "A translational engine at the national scale: Informatics for integrating biology and the bedside," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 181–185, 2012.

[2] G. Hripcsak and D. Albers, "Next-generation phenotyping of electronic health records," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 2013.

[3] D. Charles, M. Gabriel, and M. Furukawa, "Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012," *ONC Data Brief*, vol. 9, pp. 1–9, 2013.

[4] J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel, "Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015," *ONC Data Brief*, vol. 35, pp. 1–9, 2016.

[5] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: Towards better research applications and clinical care,” *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[6] J. Denny, L. Bastarache, M. Ritchie, *et al.*, “Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data,” *Nature Biotechnology*, vol. 31, no. 12, pp. 1102–1110, 2013.

[7] J. Denny, M. Ritche, M. Basford, *et al.*, “Phewas: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations,” *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, 2010.

[8] K. Liao, T. Cai, V. Gainer, *et al.*, “Electronic medical records for discovery research in rheumatoid arthritis,” *Arthritis care Research*, vol. 62, no. 8, pp. 1120–1127, 2010.

[9] C. Cipparone, M. Withiam-Leitch, K. Kimminau, *et al.*, “Inaccuracy of icd-9 codes for chronic kidney disease: A study from two practice-based research networks (pbrns),” *Journal of the American Board of Family Medicine*, vol. 28, no. 5, pp. 678–682, 2010.

[10] R. Carroll, W. Thompson, A. Eyeler, *et al.*, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Journal of the American Medical Informatics Association*, vol. 19, no. e1, e162–169, 2016.

[11] K. Liao, A. Ananthakrishnan, V. Kumar, *et al.*, “Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts,” *PLoS One*, vol. 10, no. 8, e0136651, 2015.

[12] B. Beaulieu-Jones and C. Greene, “Semi-supervised learning of the electronic health record for phenotype stratification,” *Journal of Biomedical Informatics*, vol. 64, pp. 168–178, 2013.

[13] K. Newton, P. Peissing, A. Kho, *et al.*, “Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the emerge network,” *Journal of the American Medical Informatics Association*, vol. 20, no. e1, e147–154, 2013.

[14] A. Ananthakirshnan, T. Cai, G. Savova, *et al.*, “Improving case definition of crohn’s disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach,” *Inflammatory Bowel Disease*, vol. 19, no. 7, pp. 1411–1420, 2013.

[15] Z. Xia, E. Secor, L. Chibnik, *et al.*, “Modeling disease severity in multiple sclerosis using electronic health records,” *PLoS One*, vol. 8, no. 11, e78927, 2013.

[16] K. Liao, T. Cai, G. Savova, *et al.*, “Development of phenotype algorithms using electronic medical records and incorporating natural language processing,” *British Medical Journal*, vol. 350, h1885, 2015.

[17] J. Kirby, P. Speltz, L. Rasmussen, *et al.*, “Phekb: A catalog and workflow for creating electronic phenotype algorithms for transportability,” *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1046–1052, 2016.

[18] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[19] J. Ho, J. Ghosh, and J. Sun, “Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization,” *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD ’14*, pp. 115–124, 2014.

[20] J. C. Ho *et al.*, “Limestone: High-throughput candidate phenotype generation via tensor factorization,” *Journal of Biomedical Informatics*, vol. 52, pp. 199–211, 2014.

[21] J. Henderson *et al.*, “Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping,” in *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics*, Institute of Electrical and Electronics Engineers Inc, 2017, pp. 214–223.

[22] R. Pivovarov, A. Perotte, E. Grave, *et al.*, “Learning probabilistic phenotypes from heterogeneous ehr data,” *Journal of Biomedical Informatics*, vol. 58, pp. 156–165, 2015.

[23] M. B. Mayhew, B. K. Petersen, A. P. Sales, J. D. Greene, V. X. Liu, and T. S. Wasson, "Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models," *Journal of Biomedical Informatics*, vol. 78, 2018.

[24] R. Henao, J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin, "Electronic health record analysis via deep poisson factor models," *Journal of Machine Learning Research*, vol. 17, no. 186, pp. 1–32, 2016.

[25] S. Yu, Y. Ma, J. Gronsbell, *et al.*, "Enabling phenotypic big data with phenorm," *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 54–60, 2018.

[26] K. Liao, J. Sun, T. Cai, *et al.*, "High-throughput multimodal automated phenotyping (map) with application to phewas," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1255–1262, 2019.

[27] S. Yu, K. Liao, S. Shaw, *et al.*, "Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 993–1000, 2015.

[28] S. Yu, A. Chakrabortty, K. Liao, *et al.*, "Surrogate-assisted feature extraction for high-throughput phenotyping," *Journal of the American Medical Informatics Association*, vol. 24, no. e1, e143–149, 2017.

[29] M. Levine, D. Albers, and G. Hripcsak, "Methodological variations in lagged regression for detecting physiologic drug effects in ehr data," *Journal of Biomedical Informatics*, vol. 86, pp. 149–159, 2018.

[30] Y. Ahuja, D. Zhou, Z. He, *et al.*, "Surelda: A multidisease automated phenotyping method for the electronic health record," *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1235–1243, 2020.

[31] R. Miotto, L. Li, B. Kidd, *et al.*, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, p. 26 094, 2016.

[32] J. K. De Freitas *et al.*, "Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records," *Patterns (NY)*, vol. 2, no. 9, 2021.

[33] V. Agarwal, T. Podchiyska, J. Banda, *et al.*, "Learning statistical models of phenotypes using noisy labeled training data," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1166–1173, 2016.

[34] Y. Li, P. Nair, X. Lu, *et al.*, "Inferring multimodal latent topics from electronic health records," *Nature Communications*, vol. 11, no. 2536, 2020.

[35] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Science USA*, vol. 101, pp. 5228–5235, 2004.

[36] A. Asuncion, M. Welling, P. Smyth, and Y. Teh, "On smoothing and inference for topic models," *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence UAI'09*, pp. 27–34, 2009.

[37] A. Shaban-Nejad, M. Lavigne, A. Okhmatovskaia, and D. L. Buckeridge, "PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data.," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 44–53, 2016. DOI: 10.1111/nyas.13271. [Online]. Available: http://doi.wiley.com/10.1111/nyas.13271.

[38] A. Johnson *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, pp. 160 035–1 600 359, 2016.

[39] Y. Halpern, Y. Choi, S. Horng, *et al.*, "Using anchors to estimate clinical state without labeled data," *AMIA Annual Symposium Proceedings 2014*, pp. 606–615, 2014.

[40] W.-Q. Wei *et al.*, "Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record," *PloS one*, vol. 12, no. 7, e0175508, 2017.

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

[42] P. Wu, A. Gifford, X. Meng, *et al.*, “Mapping icd-10 and icd-10-cm codes to phecodes: Workflow development and initial evaluation,” *JMIR Medical Informatics*, vol. 7, no. 4, e14325, 2019.

[43] L. M. Lix, J. Ayles, S. Bartholomew, *et al.*, “The canadian chronic disease surveillance system: A model for collaborative surveillance,” *International Journal of Population Data Science*, vol. 3, no. 3, p. 433, 2018.

[44] H.-M. Vasiliadis, F. B. Diallo, L. Rochette, *et al.*, “Temporal trends in the prevalence and incidence of diagnosed adhd in children and young adults between 1999 and 2012 in canada: A data linkage study,” *Canadian Journal of Psychiatry*, vol. 62, no. 12, pp. 818–826, 2017.

[45] F. B. Diallo, L. Rochette, E. Pelletier, and A. Lesage, *Autism Spectrum Disorder Surveillance in Quebec*. Institut national de sante publique Quebec, 2017.

[46] M. Durand, O. Sheehy, J.-G. Baril, J. Lelorier, and C. L. Tremblay, “Association between hiv infection, antiretroviral therapy, and risk of acute myocardial infarction: A cohort and nested case-control study using québec’s public health insurance database,” *Journal of Acquired Immune Deficiency Syndromes*, vol. 57, no. 3, pp. 245–253, 2011.

[47] I. for Health Metrics and E. (IHME), “Epi visualization,” 2020.

[48] P. Buckley, B. Miller, D. Lehrer, and D. Castle, “Psychiatric comorbidities and schizophrenia,” *Schizophrenia Bulletin*, vol. 35, no. 2, 2009.

[49] S. Young, D. Pfaff, K. Lewandowski, *et al.*, “Anxiety disorder comorbidity in bipolar disorder, schizophrenia and schizoaffective disorder,” *Psychopathology*, vol. 46, pp. 176–185, 2013.

[50] C. Correll, D. Ng-Mak, D. Stafkey-Mailey, *et al.*, “Cardiometabolic comorbidities, readmission, and costs in schizophrenia and bipolar disorder: A real-world analysis,” *Annals of General Psychiatry*, vol. 16, no. 9, 2017.

[51] S. Chakrabarti, “Thyroid functions and bipolar affective disorder,” *Journal of Thyroid Research*, 2011.

[52] Z. Gan, X. Wu, Z. Chen, *et al.*, “Rapid cycling bipolar disorder is associated with antithyroid antibodies, instead of thyroid dysfunction,” *BMC Psychiatry*, vol. 19, no. 378, 2019.

[53] A. Bocchetta, F. Traccis, E. Mosca, A. Serra, G. Tamburini, and A. Loviselli, “Bipolar disorder and antithyroid antibodies: Review and case series,” *International Journal of Bipolar Disorders*, vol. 4, no. 5, 2016.

[54] A. Reddy, B. Birur, R. Shelton, and L. Li, “Major depressive disorder following dermatomyositis: A case linking depression with inflammation,” *Psychopharmacology Bulletin*, vol. 48, no. 3, 2018.

[55] K. Hannibal and M. Bishop, “Chronic stress, cortisol dysfunction, and pain: A psychoneuroendocrine rationale for stress management in pain rehabilitation,” *Physical Therapy & Rehabilitation Journal*, vol. 94, no. 12, pp. 1816–1825, 2014.

[56] L. Bastarache, “Using phecodes for research with the electronic health record: From phewas to phers,” *Annual Review of Biomedical Data Science*, vol. 4, pp. 1–19, 2021.

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

[58] V. Martucci, N. Liu, V. Kerchberger, *et al.*, “A clinical phenotyping algorithm to identify cases of chronic obstructive pulmonary disease in electronic health records,” *bioRxiv*, 2021.

[59] J. Wong, M. M. Horwitz, L. Zhou, and S. Toh, “Using machine learning to identify health outcomes from electronic health record data,” *Current epidemiology reports*, vol. 5, no. 4, pp. 331–342, 2018.

[60] S. Gunasekar, J. Ho, J. Ghosh, *et al.*, “Phenotyping using structured collective matrix factorization of multi–source ehr data,” *arXiv*, 2016.

[61] Z. Song *et al.*, “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*, ser. BCB ’21, Gainesville, Florida: Association for Computing Machinery, 2021, ISBN: 9781450384506. DOI: 10.1145/3459930.3469543. [Online]. Available: https://doi.org/10.1145/3459930.3469543.

[62] M. Yuan, G. Powell, M. Lavigne, A. Okhmatovskaia, and D. L. Buckeridge, “Initial usability evaluation of a knowledge-based population health information system: The population health record (pophr),” in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2017, 2017, p. 1878.

[63] C. Bycroft *et al.*, “The uk biobank resource with deep phenotyping and genomic data,” *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[64] Z. Wen *et al.*, “Mining heterogeneous clinical notes by multi-modal latent topic model,” *PloS one*, vol. 16, no. 4, e0249622, 2021.

[65] Y. Ahuja, L. Liang, S. Huang, and T. Cai, “Semi-supervised calibration of risk with noisy event times (scornet) using electronic health record data,” *bioRxiv*, 2021. DOI: 10.1101/2021.01.08.425976. eprint: https://www.biorxiv.org/content/early/2021/01/09/2021.01.08.425976.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2021/01/09/2021.01.08.425976.

Supplementary Materials

MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record

Yuri Ahuja[1,2,*], **Yuesong Zou**[3], **Aman Verma**[4], **David Buckeridge**[4], **Yue Li**[3,*]

[1]Department of Biostatistics, Harvard TH Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA [2]Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA [3]School of Computer Science, McGill University [4]School of Population and Global Health, McGill University, 2001 McGill College Avenue, Montreal, Québec H3A 1G1, Canada

[*]Correspondence: `yuri_ahuja@hms.harvard.edu`, `david.buckeridge@mail.ca`, `yueli@cs.mcgill.ca`

S1. Supplementary Figures

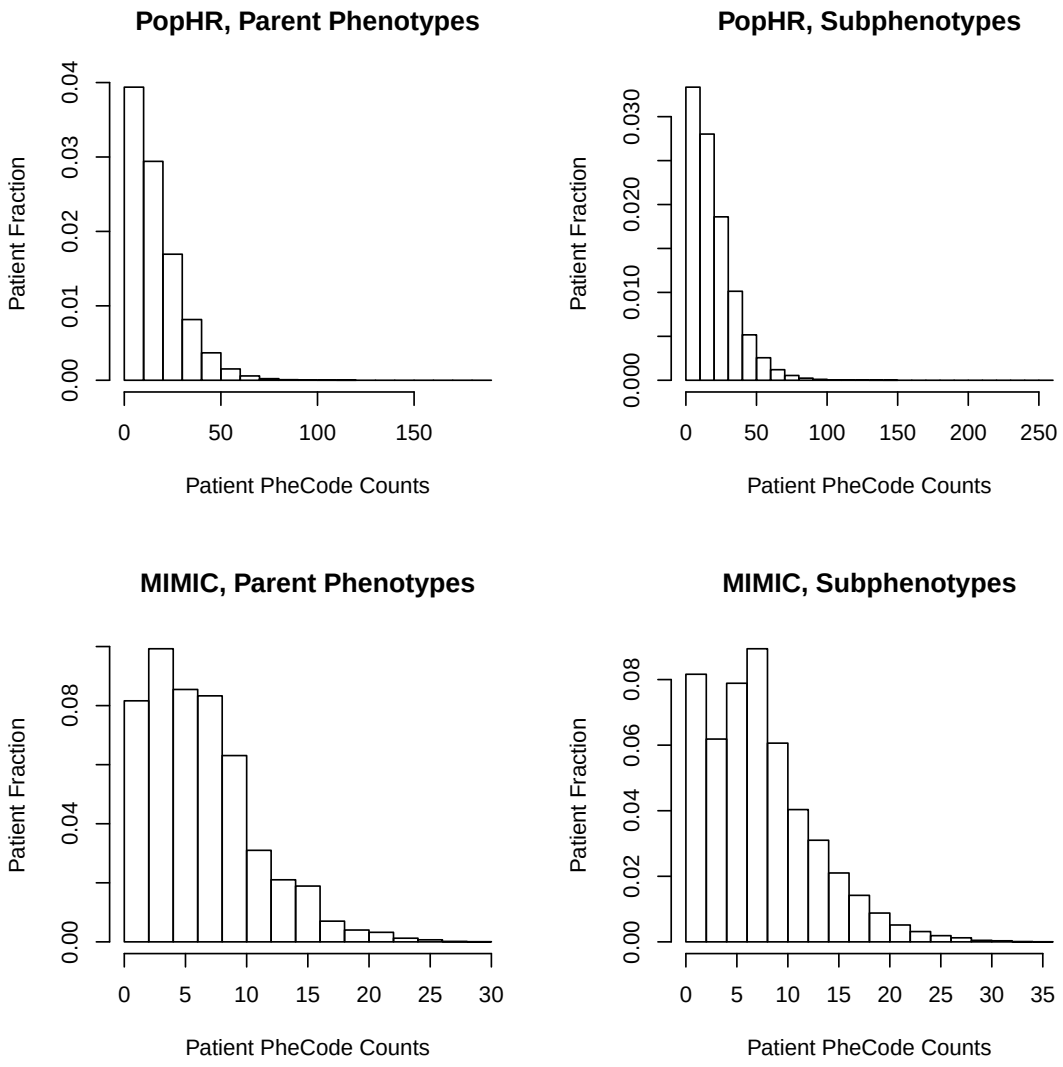


Figure S1: Distribution of PheCode anchors per patient (including both 3-integer and 3-integer+1-decimal codes) for the PopHR and MIMIC-III datasets.

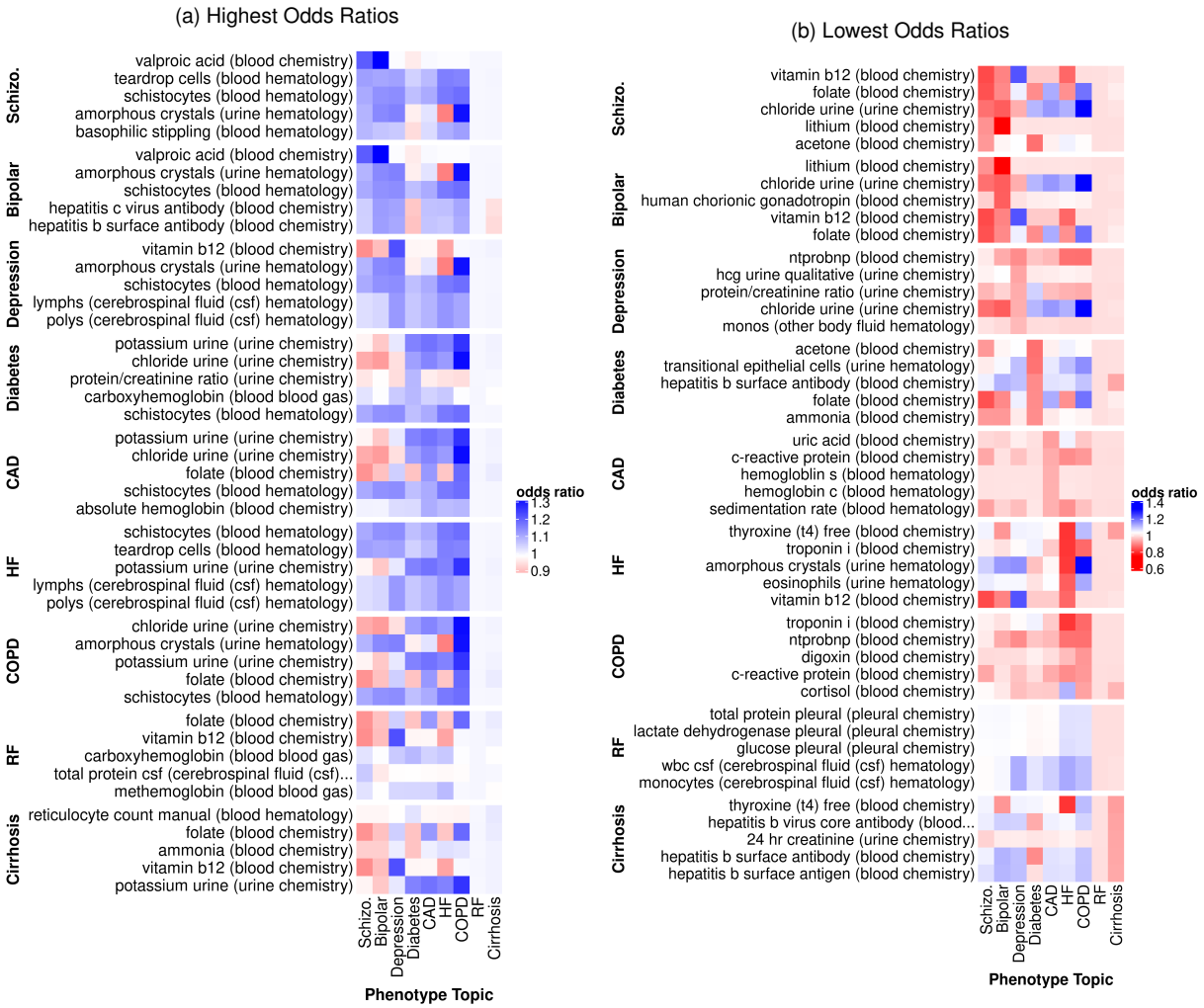


Figure S2: Lab tests with the (a) highest and (b) lowest odds ratio of abnormal result for each of 9 diverse disease phenotypes as ascertained by MixEHR-G trained on the MIMIC-III dataset.

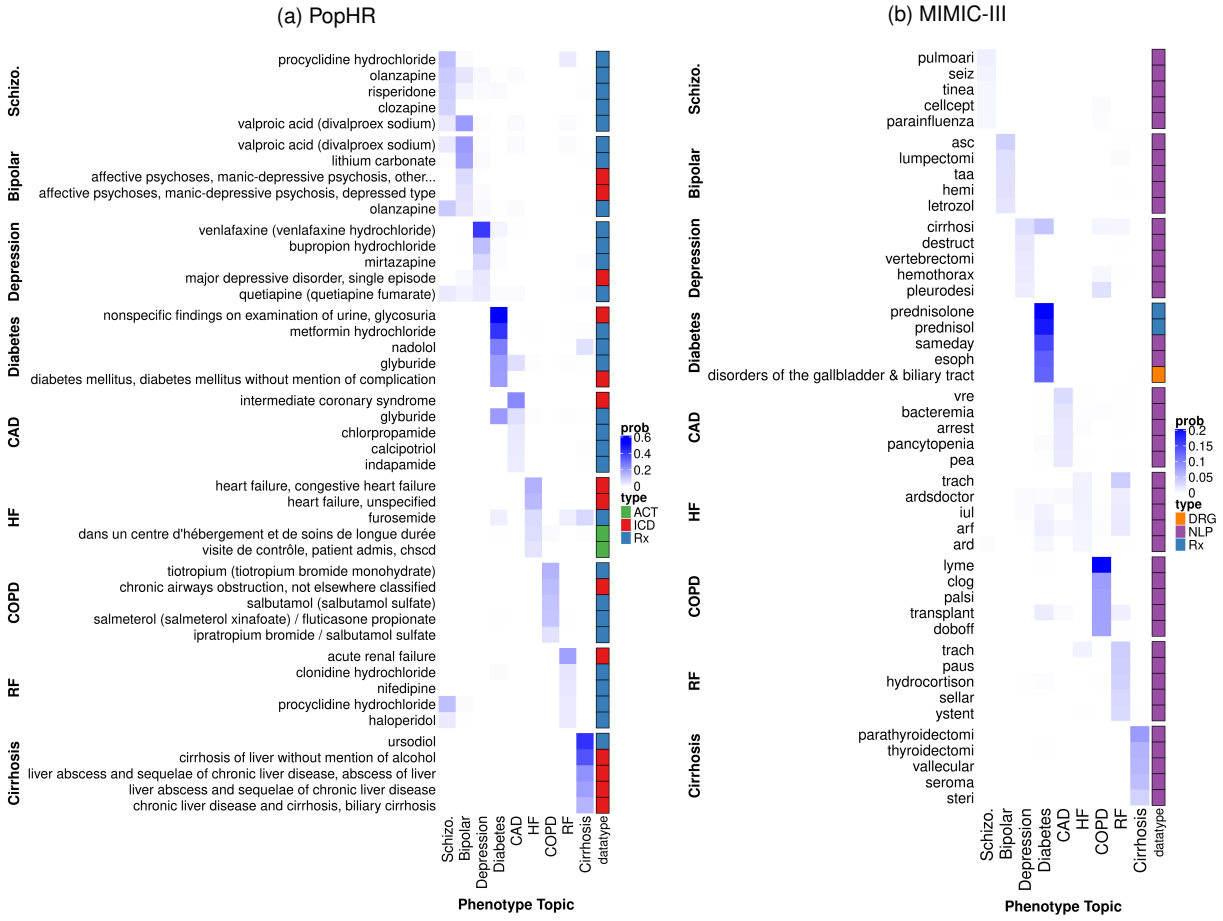
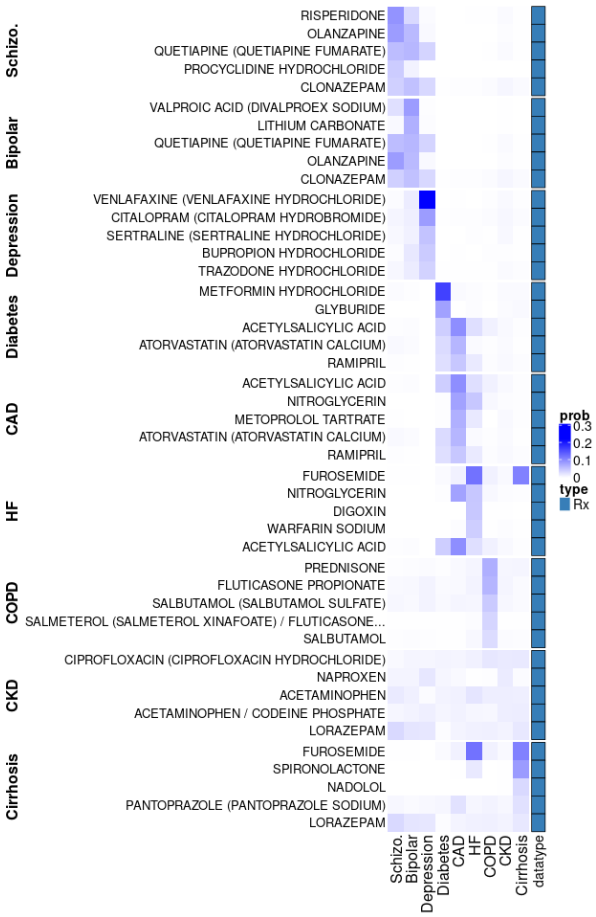


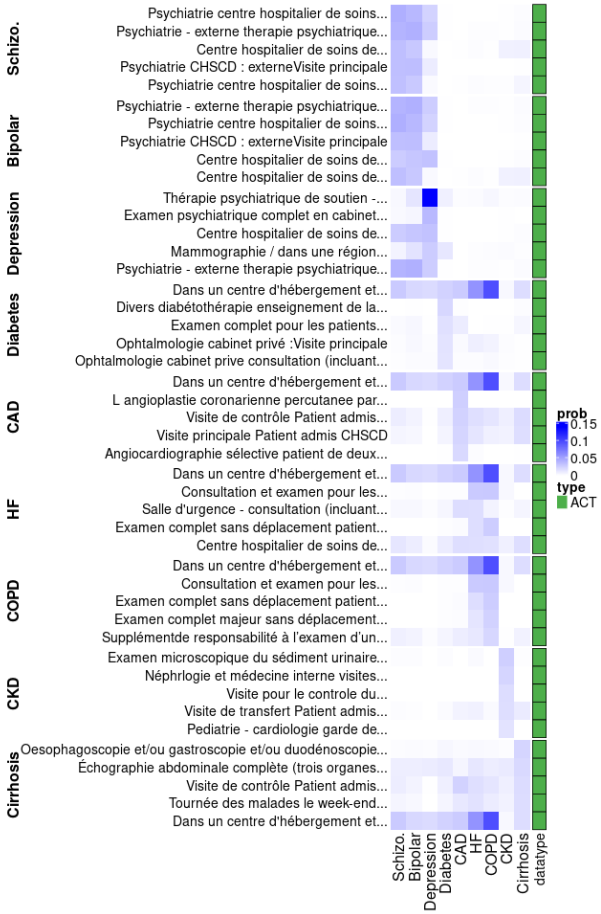
Figure S3: Top 5 features for each of 9 diverse disease phenotypes as ascertained by sureLDA, for comparison with **Fig.** 2. We present results for two datasets: (a) PopHR, which has 3 data modalities, and (b) MIMIC-III, which has 5.

(a) PopHR/ICD-9

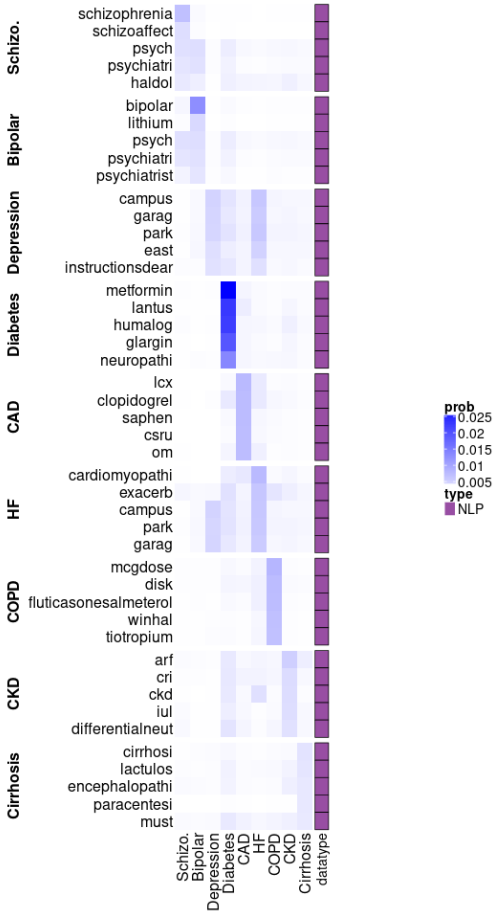
(b) PopHR/ACT



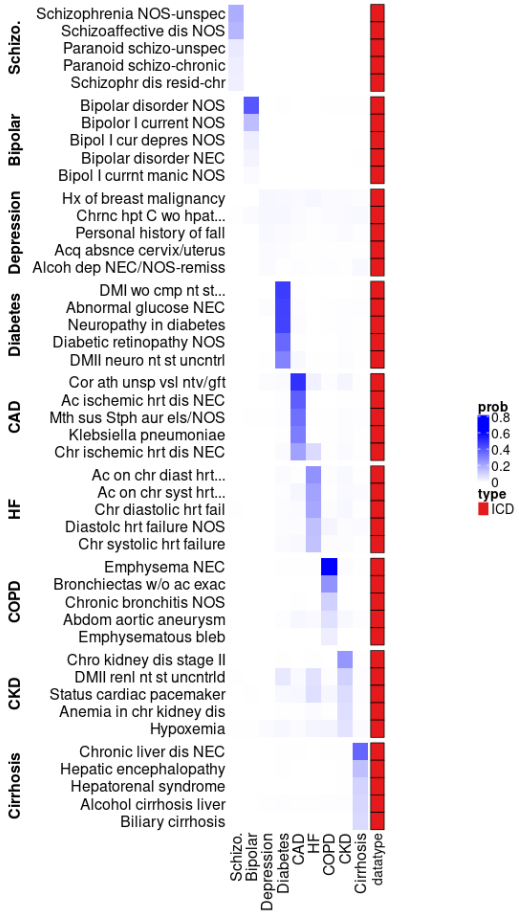
(c) PopHR/Rx



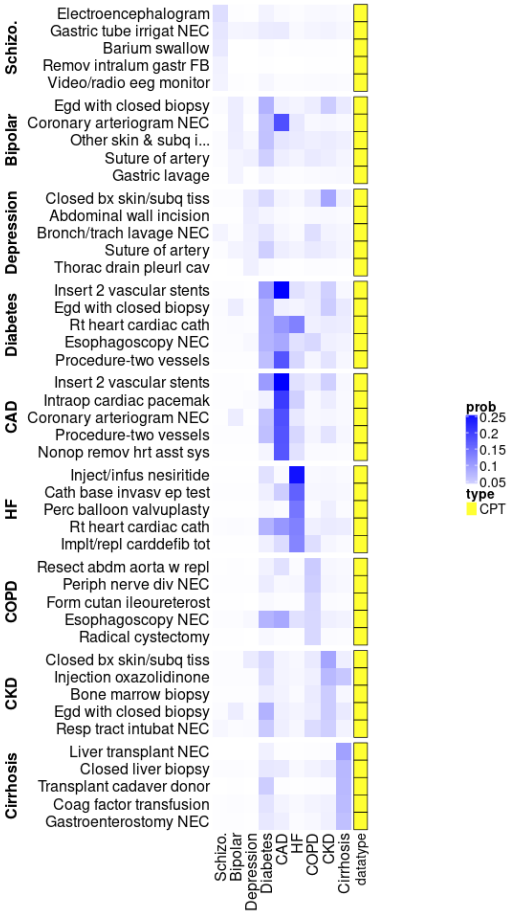
(d) MIMIC-III/NLP



(e) MIMIC-III/ICD



(f) MIMIC-III/CPT



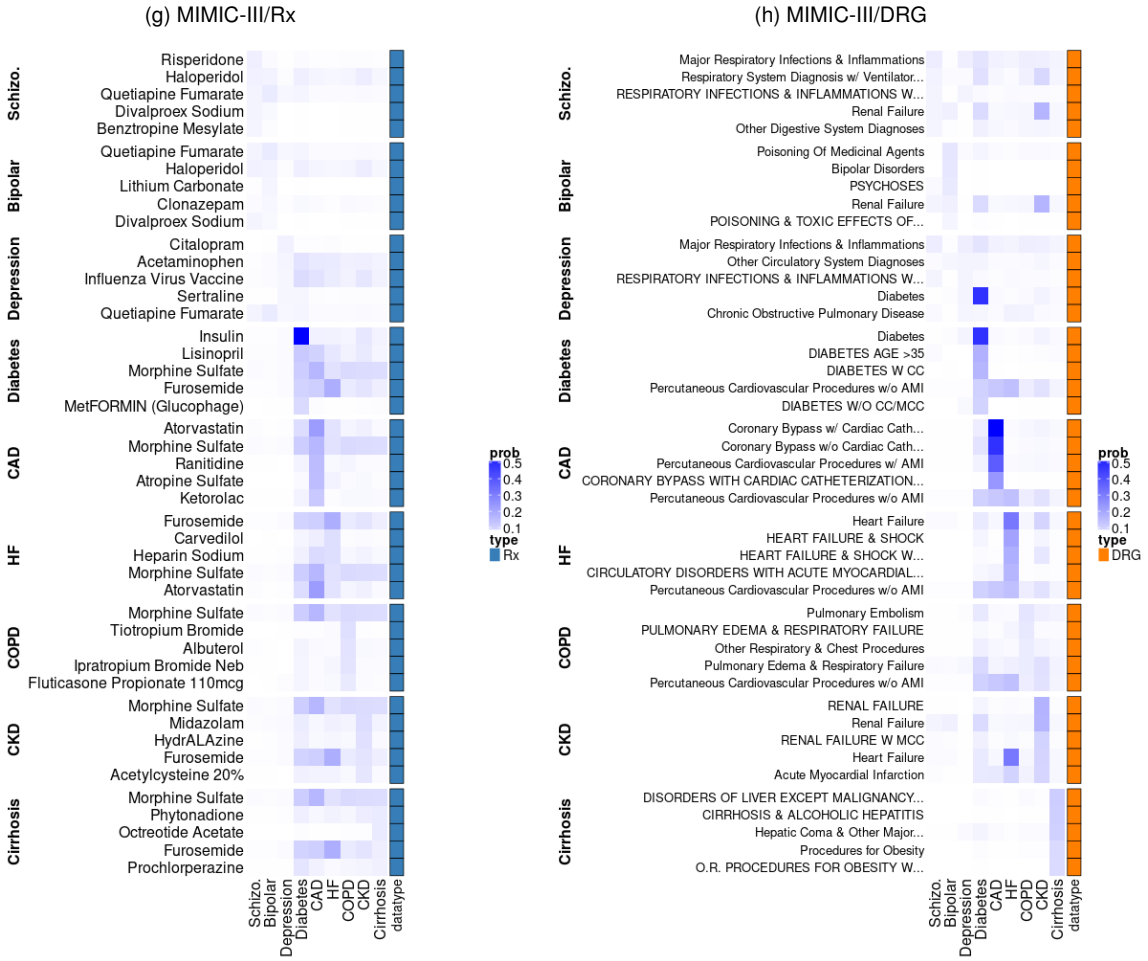


Figure S4: Top 5 features stratified by data modality for each of 9 diverse disease phenotypes as ascertained by MixEHR-G. We present results for two datasets: (a-c) PopHR, which has 3 multinomial data modalities (ICD-9, ACT, and Rx), and (d-h) MIMIC-III, which has 5 (NLP, ICD-9, CPT, Rx, and DRG). Note that while MIMIC-III also contains lab observations and results, MixEHR-Gdoes not model these modalities as multinomials, and thus their corresponding distributions cannot be interrogated in this way.

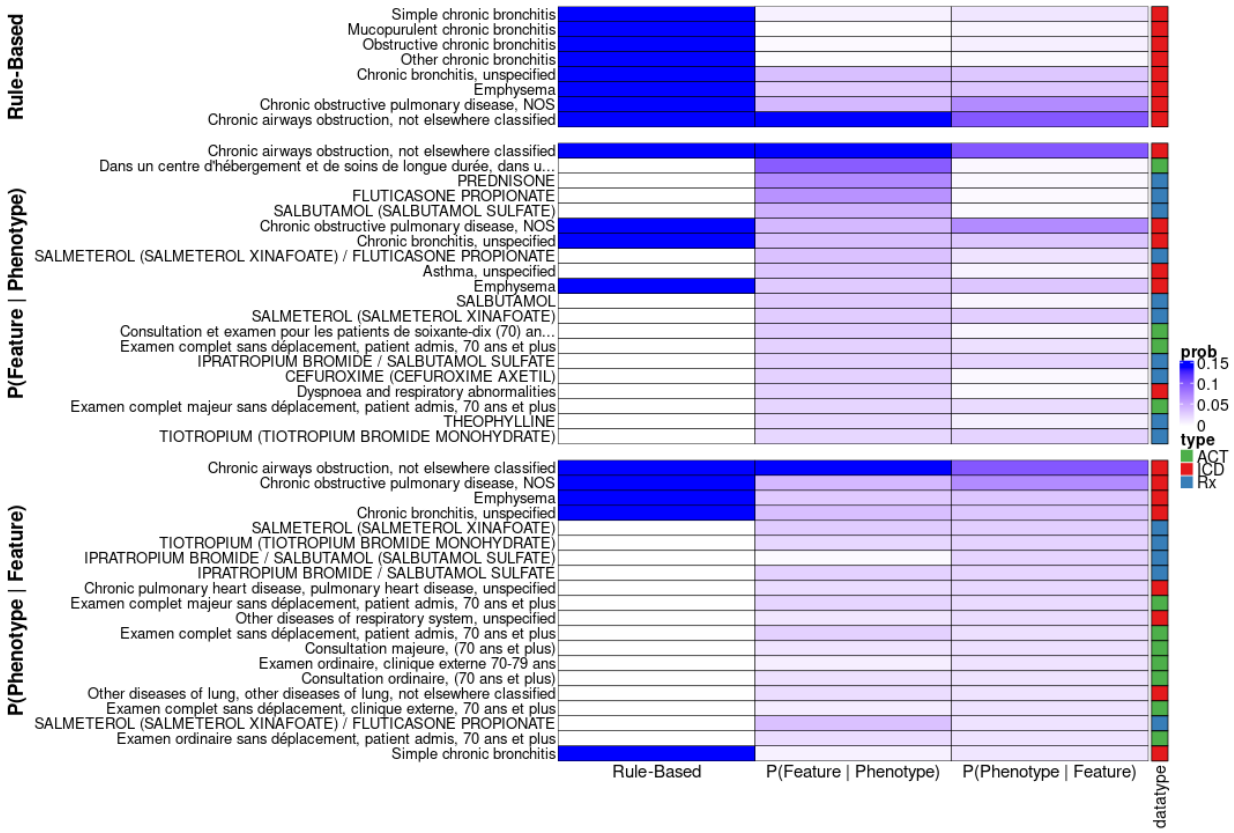


Figure S5: Top 25 features with the highest (1) sensitivities {P(Feature | Phenotype)} and (2) positive predictive values {P(Phenotype | Feature)} for COPD based on MixEHR-G's trained topic distributions $\boldsymbol{\phi}$, as compared to features included in the rule-based COPD algorithm.