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

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

[1]Department of Biostatistics, Harvard TH Chan School of Public Health;
Harvard Medical School
[2]Department of Computer Science, McGill University
[3]School of Population and Global Health, McGill University
[*]Correspondence to yuri_ahuja@hms.harvard.edu, david.buckeridge@mail.ca,
yueli@cs.mcgill.ca

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 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.

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 [18–26]. 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 that can be subsequently used to guide phenotype estimation in a “weakly supervised" manner [20–22, 25, 26]. 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 [20, 21], 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 [26–29]. 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 [26, 27]. 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 [27]. 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 [26]. 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 propose MixEHR-Guided (MixEHR-G), an adaptation of MixEHR that employs the guided topic modeling approach of sureLDA to align inferred phenotype topics with well-defined disease phenotypes. We apply MixEHR-G to the PopHR dataset [30], which contains longitudinal administrative claims data for a random sample of 1.3 million residents of Quebec, Canada (**Methods**), and to the MIMIC-III dataset [31], 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 EHR data including ICD codes and prescriptions. 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 Results

2.1 MixEHR-G and EHR data overview

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 Phenotype Codes (PheCodes), expert-defined groupings of ICD-9 codes that have been shown to more closely align with disease phenotypes described in clinical practice than individual ICD-9 codes [32], 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 more interpretable and well-defined phenotypes.

We applied MixEHR-G with phenotype topics defined by the 1515 unique integer and 1-decimal PheCodes to the PopHR and MIMIC-III datasets described in the Introduction [30, 31]. Besides the distinct population coverage of the two datasets, they also involve different feature modalities and levels of disease acuity (**Methods**). 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 (i.e. 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 used these two datasets to demonstrate the generalizability of our MixEHR-G and gain contrasting insights between the datasets regarding the characterization of a common set of diseases.

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. 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.

2.2 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 (ϕ_k's) for 9 diverse chronic diseases covering a variety of medical specialities (**Fig.** 2; **Supplementary Fig.** S2). We observe that the phenotype topics appear to be predominantly defined by prescription (Rx) codes in the PopHR dataset (**Fig.** 2a), whereas in the MIMIC-III dataset they are dominated by ICD-9 codes (**Fig.** 2b). This highlights the ability of MixEHR-G 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. In contrast, the top features identified by sureLDA are mostly ICD-9 codes for PopHR and words from clinical notes for MIMIC-III, for which words account for 33,388 of 53,432 total features (**Supplementary Fig.** S1).

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. While MixEHR-G is anchored to a prior defined by PheCodes, it is able to infer phenotype topics by leveraging prescription data from PopHR. 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.

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 [33–35] (**Fig.** 3b). Other meaningful 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 [36–38]. Similarly, the association between depression and myalgia/myositis supports the hypothesized clinical association between chronic pain/inflammation and depression [39, 40].

2.3 Automated phenotyping using MixEHR-G's topic mixture

A salient feature of MixEHR-G is that the 1515 PheCode-guided topics are readily identifiable. Consequently, they can be directly used as phenotype prediction scores. 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 (**Methods**; **Fig.** 4; **Table** 1). In particular, MixEHR-G's scores exhibited higher AUPRCs than the raw PheCodes or the recently published unsupervised phenotyping algorithms Multimodal Automated Phenotyping (MAP) [22] and sureLDA [26].

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 [26, 27]. 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 outperformed the baseline methods including the raw PheCode counts and MAP. Since both baselines 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 rules (**Fig.** 5; **Supplementary Fig.** S3) [41]. We first examined the top 25 codes under the COPD topic's distribution ϕ_{COPD} (**Fig.** 5a; **Supplementary Fig.** S3). 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 scores (**Fig.** 5b left panel) as well as the top 10 rule-positive patients with the lowest 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.

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

We used MixEHR-G's patient-phenotype mixtures $\tilde{\Theta}$ 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. 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 [42] (**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) predominates in women, likely reflecting increased survivorship among female Alzheimer's patients.

3 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 [43]. However, supervised learning methods are not scalable for simultaneously predicting thousands of phenotype labels. Model-based unsupervised methods such as non-negative matrix factorization [44], autoencoder [5], and topic modeling [27] 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 [45] can simultaneously infer topic distributions and 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 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).

As a proof-of-concept analysis, 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 [46].

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, obviating the need for labor-intensive chart review or rule generation. For a study focusing on a single phenotype, MixEHR-G's topic distributions ϕ can be manually analyzed to improve upon rules, and $\tilde{\Theta}$ can potentially be incorporated into an ensemble phenotyping algorithm as a strongly predictive feature. These phenotype labels can then be used for downstream epidemiological research or to identify cohorts for inclusion in a clinical trial.

MixEHR-G's second major application is to derive population-level insights about phenotypes. In this study, we validate the use of MixEHR-G's patient-topic mixtures $\tilde{\Theta}$ 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 [47] 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.

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 MixEHR-G.

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 wrought 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.

In summary, by modeling multimodal phenotypes with the guided topic modeling strategy, MixEHR-G achieves accurate simultaneous phenotype predictions for 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.

4 Methods

4.1 MixEHR-G

MixEHR-G models EHR data using the generative latent topic model MixEHR as previously described [27] but with topic hyperparameters constrained to align with reference diseases using a procedure similar to that detailed in Ahuja et al. [26]. The 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 [27] 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 [48], we treat each patient record as a document and the frequency of 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). EHR codes are often filtered by feature selection methods such as the Surrogate-Assisted Feature Extraction (SAFE) method [24], 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 $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)'$, using a modified MAP algorithm [22]. 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 [32]. We mapped ICD codes to PheCodes using the established PheWAS mapping [49]. 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). In total, there were 499 parent phenotypes and 1016 subphenotypes (1515 in total). We then run 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.

Step 3: training MixEHR-G MixEHR models EHR data using a generative latent topic model inspired by Latent Dirichlet Allocation [27, 48] (**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 $z^{(t)}_{dj}$ from a multinomial distribution with probability vector θ_d. Given $z^{(t)}_{dj}$, we then draw an EHR code $x^{(t)}_{d,j}$ from a multinomial distribution with probability vector $\phi^{(t)}_{z^{(t)}_{d,j}}$.

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. [27]. The key variational topic inference step is:

$$\gamma^{(t)}_{d,j,k} \propto \left(\alpha^*_k + \tilde{n}^{-(d,j)}_{d,.,k}\right) \left(\frac{\beta^{(t)}_{x^{(t)}_{d,j}} + [\tilde{n}^{(t)}_{.,x^{(t)}_{d,j},k}]^{-(d,j)}}{\sum_w \beta^{(t)}_w + [\tilde{n}^{(t)}_{.,w,k}]^{-(d,j)}} \right) \tag{1}$$

where $\gamma^{(t)}_{d,j,k}$ denotes the variational probabilities of the k^{th} topic assigned to EHR code j of patient d: $q(z^{(t)}_{d,j} = k) \equiv \gamma^{(t)}_{d,j,k}$ 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:

$$\tilde{n}^{-(d,j)}_{d,.,k} = \sum_{t=1}^{T} \sum_{j' \neq j}^{M^{(t)}_d} \gamma^{(t)}_{d,j',k}, \quad [\tilde{n}^{(t)}_{.,x^{(t)}_{d,j},k}]^{-(d,j)} = \sum_{d' \neq d}^{D} \sum_{j=1}^{M^{(t)}_{d'}} [x^{(t)}_{d',j'} = x^{(t)}_{d,j}] \gamma^{(t)}_{d',j,k} \tag{2}$$

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:

$$\alpha'_k \leftarrow \frac{a_\alpha - 1 + \alpha_k \sum_d \Psi(\alpha_k + \tilde{n}_{dk} + \tilde{m}_{dk}) - \Psi(\alpha_k)}{b_\alpha + \sum_i \Psi(\tilde{n}_{dk} + \sum_k \alpha_k) - \Psi(\sum_k \alpha_k)} \tag{3}$$

$$\alpha^*_k \leftarrow \pi_k \alpha'_k \frac{\sum_j \alpha'_j}{\sum_j \pi_j \alpha'_j} \tag{4}$$

The remainder of the algorithm is identical to that detailed in Li et al. [27] and thus omitted here. When MixEHR-G converges, we obtain the count of total EHR features assigned to topic k for subject d as the patient-level topic 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.

4.2 Implementation of existing methods

We implemented sureLDA [26], MAP [22], and PheNorm [21] 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 [27] using the source code from the github page: `https://github.com/li-lab-mcgill/mixehr`.

4.3 Application to the MIMIC-III dataset

Medical Information Mart for Intensive Care (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 [31]. 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.

We ran MixEHR-G on parent phenotypes and subphenotypes separately. We then derived the MixEHR-G phenotype score $\tilde{\Theta}_k$ 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.

4.4 Application to the PopHR dataset

The Population Health Record (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 [30]. 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 6,120 unique ICD-9 codes, 6,920 ACT codes, and 11,152 DIN codes for a total of 24,192 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*.

As with the MIMIC-III dataset, we again ran the guided MixEHR procedure on parent phenotypes and subphenotypes separately. Likewise, we derived MixEHR-G scores for coarse phenotypes (i.e. COPD) from the parent phenotype model and for subphenotypes (i.e. emphysema) from the subphenotype model.

4.5 Phenotyping accuracy evaluation

For the PopHR data [30], we have rule-based phenotyping algorithms for 12 diseases that were derived solely from ICD codes [50]. Some of these rules rely on consecutive observations of the same ICD code(s) for a patient within a small (i.e. 1 month) time interval. 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.

4.6 Identification of similar and comorbid phenotypes

We define 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 types. We define 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 phenotypes, and we use the trained $\boldsymbol{\theta}_{D\times K}$ matrix to identify comorbid phenotypes.

4.7 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 [42]. 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 ρ_k:

$$\tau^* \leftarrow \min_{\tau} \left| \frac{1}{D} \sum_d \hat{Y}_{dk} - \rho_k \right|$$

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

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 [42]. Again, we computed relative prevalence rates from the absolute prevalence rates supplied by the data exchange by normalizing by ρ_k.

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. 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.

References

1. IS Kohane, SE Churchil, and SN Murphy. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2):181–185, 2012.

2. G Hripcsak and DJ Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

3. D Charles, M Gabriel, and MF Furukawa. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC Data Brief*, 9:1–9, 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*, 35:1–9, 2016.

5. R Miotto, L Li, BA Kidd, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6:26094, 2016.

6. JC Denny, L Bastarache, MD Ritchie, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1110, 2013.

7. JC Denny, MD Ritche, MA Basford, et al. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

8. KP Liao, T Cai, V Gainer, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care Research*, 62(8):1120–1127, 2010.

9. CW Cipparone, M Withiam-Leitch, KS 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*, 28(5):678–682, 2010.

10. RJ Carroll, WK Thompson, AE 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*, 19(e1):e162–169, 2016.

11. KP Liao, AN 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*, 10(8):e0136651, 2015.

12. BK Beaulieu-Jones and CS Greene. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, 64:168–178, 2013.

13. KM Newton, PL Peissing, AN 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*, 20(e1):e147–154, 2013.

14. AN 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*, 19(7):1411–1420, 2013.

15. Z Xia, E Secor, LB Chibnik, et al. Modeling disease severity in multiple sclerosis using electronic health records. *PLoS One*, 8(11):e78927, 2013.

16. KP Liao, T Cai, GK Savova, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *British Medical Journal*, 350:h1885, 2015.

17. JC Kirby, P Speltz, LV Rasmussen, et al. Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6):1046–1052, 2016.

18. Y Halpern, Y Choi, S Horng, et al. Using anchors to estimate clinical state without labeled data. *AMIA Annual Symposium Proceedings 2014*, pages 606–615, 2014.

19. Y Halpern, S Horng, Y Choi, et al. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–730, 2016.

20. V Agarwal, T Podchiyska, JM Banda, et al. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173, 2016.

21. S Yu, Y Ma, J Gronsbell, et al. Enabling phenotypic big data with phenorm. *Journal of the American Medical Informatics Association*, 25(1):54–60, 2018.

22. KP Liao, J Sun, TA Cai, et al. High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11):1255–1262, 2019.

23. S Yu, KP Liao, SY Shaw, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993–1000, 2015.

24. S Yu, A Chakrabortty, KP Liao, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association*, 24(e1):e143–149, 2017.

25. ME Levine, DJ Albers, and G Hripcsak. Methodological variations in lagged regression for detecting physiologic drug effects in ehr data. *Journal of Biomedical Informatics*, 86:149–159, 2018.

26. 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*, 27(8):1235–1243, 2020.

27. Y Li, P Nair, XH Lu, et al. Inferring multimodal latent topics from electronic health records. *Nature Communications*, 11(2536), 2020.

28. TL Griffiths and M Steyvers. Finding scientific topics. *Finding scientific topics*, 101:5228–5235, 2004.

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

30. Arash Shaban-Nejad, Maxime Lavigne, Anya Okhmatovskaia, and David 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*, 1387(1):44 – 53, 2016.

31. AEW Johnson et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035–1600359, 2016.

32. WQ Wei, LA Bastarache, RJ Carroll, et al. Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record. *PLoS One*, 2017.

33. PF Buckley, BJ Miller, DS Lehrer, and DJ Castle. Psychiatric comorbidities and schizophrenia. *Schizophrenia Bulletin*, 35(2), 2009.

34. S Young, D Pfaff, KE Lewandowski, et al. Anxiety disorder comorbidity in bipolar disorder, schizophrenia and schizoaffective disorder. *Psychopathology*, 46:176–185, 2013.

35. CU Correll, DS 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*, 16(9), 2017.

36. S Chakrabarti. Thyroid functions and bipolar affective disorder. *Journal of Thyroid Research*, 2011.

37. Z Gan, X Wu, Zhongcheng Chen, et al. Rapid cycling bipolar disorder is associated with antithyroid antibodies, instead of thyroid dysfunction. *BMC Psychiatry*, 19(378), 2019.

38. 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*, 4(5), 2016.

39. A Reddy, B Birur, RC Shelton, and Li Li. Major depressive disorder following dermatomyositis: A case linking depression with inflammation. *Psychopharmacology Bulletin*, 48(3), 2018.

40. KE Hannibal and MD Bishop. Chronic stress, cortisol dysfunction, and pain: A psychoneuroendocrine rationale for stress management in pain rehabilitation. *Physical Therapy Rehabilitation Journal*, 94(12):1816–1825, 2014.

41. VL Martucci, N Liu, VE Kerchberger, et al. A clinical phenotyping algorithm to identify cases of chronic obstructive pulmonary disease in electronic health records. *bioRxiv*, 2021.

42. Institute for Health Metrics and Evaluation (IHME). Epi visualization. 2020.

43. Jenna Wong, Mara Murray Horwitz, Li Zhou, and Sengwee Toh. Using machine learning to identify health outcomes from electronic health record data. *Current epidemiology reports*, 5(4):331–342, 2018.

44. S Gunasekar, JC Ho, J Ghosh, et al. Phenotyping using structured collective matrix factorization of multi–source ehr data. *arXiv*, 2016.

45. Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. Supervised multi-specialist topic model with applications on large-scale electronic health record data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB ’21, New York, NY, USA, 2021. Association for Computing Machinery.

46. Mengru Yuan, Guido Powell, Maxime Lavigne, Anya Okhmatovskaia, and David L Buckeridge. Initial usability evaluation of a knowledge-based population health information system: The population health record (pophr). In *AMIA Annual Symposium Proceedings*, volume 2017, page 1878. American Medical Informatics Association, 2017.

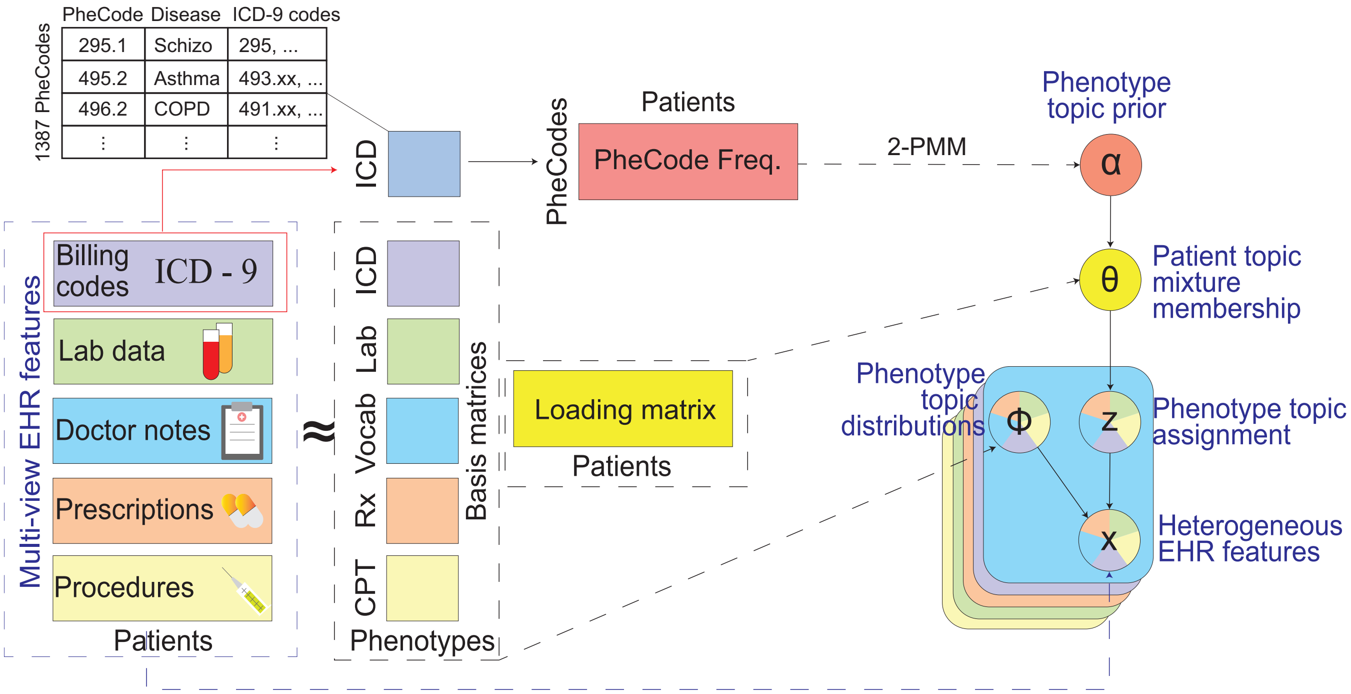
47. Yuri Ahuja, Liang Liang, Selena Huang, and Tianxi Cai. Semi-supervised calibration of risk with noisy event times (scornet) using electronic health record data. *bioRxiv*, 2021.

48. DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

49. 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*, 7(4):e14325, 2019.

50. 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*, 34 Suppl 1:1–30, 2014.

5 Figures



PheCode	Disease	ICD-9 codes
295.1	Schizo	295, ...
495.2	Asthma	493.xx, ...
496.2	COPD	491.xx, ...
⋮	⋮	⋮

Figure 1: **Schematic of the MixEHR-G algorithm.** We factorize the heterogeneous EHR data matrices into phenotype topic distributions (i.e. basis matrices) and 1 common patient topic mixture (i.e. loading 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 ($\mathbf{z}$) and phenotype topic distributions ($\boldsymbol{\phi}$) using a collapsed mean-field variational inference algorithm.

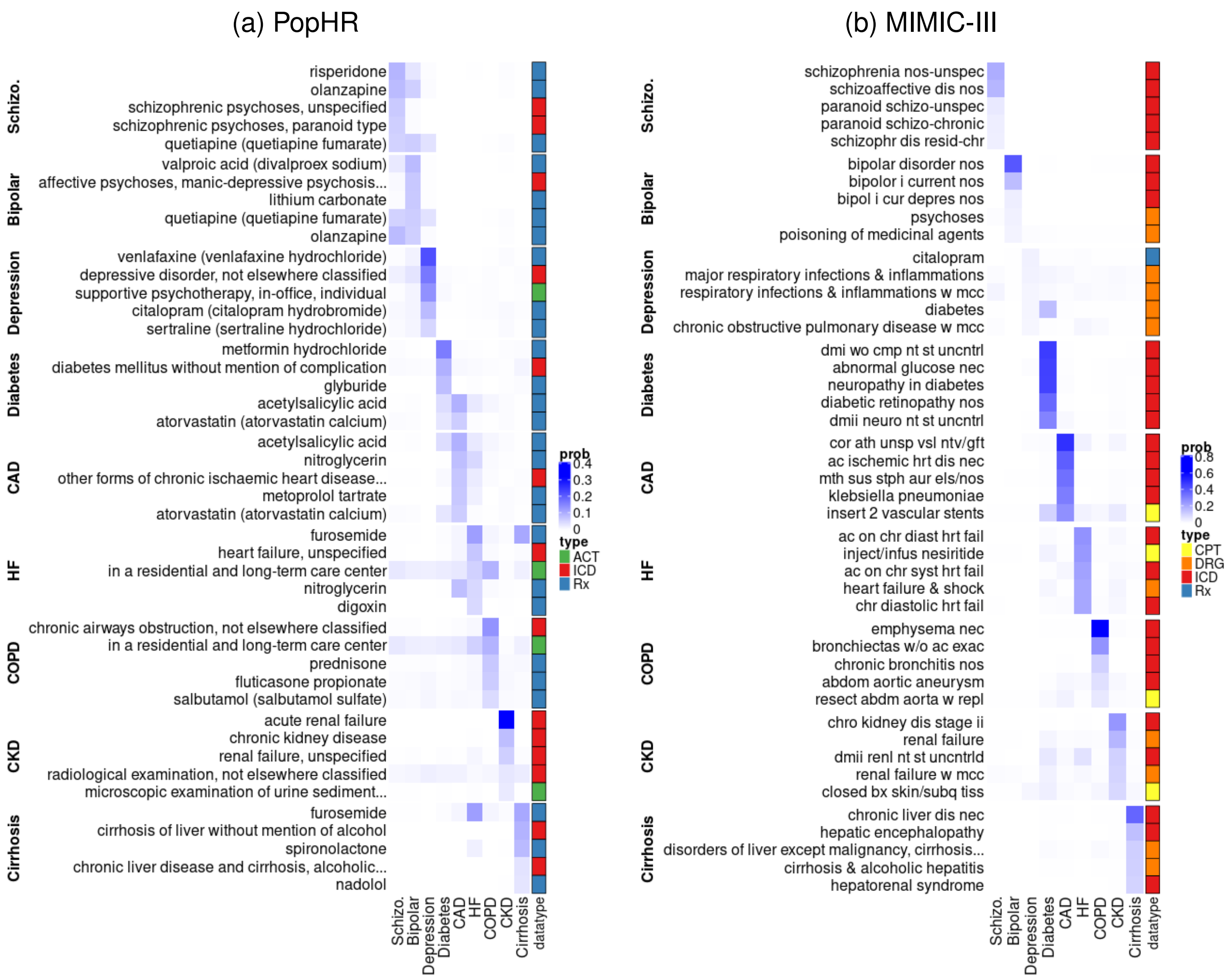


Figure 2: **Top 5 features for each of 9 diverse disease phenotypes.** (**a**) Phenotype distributions inferred from the PopHR dataset. Using MixEHR-G, we inferred phenotype distributions (ϕ) for 1515 PheCode-guided phenotypes and displayed 9 of them. Phenotype distributions are inferred over EHR features 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 CKD: Chronic Kidney Disease.

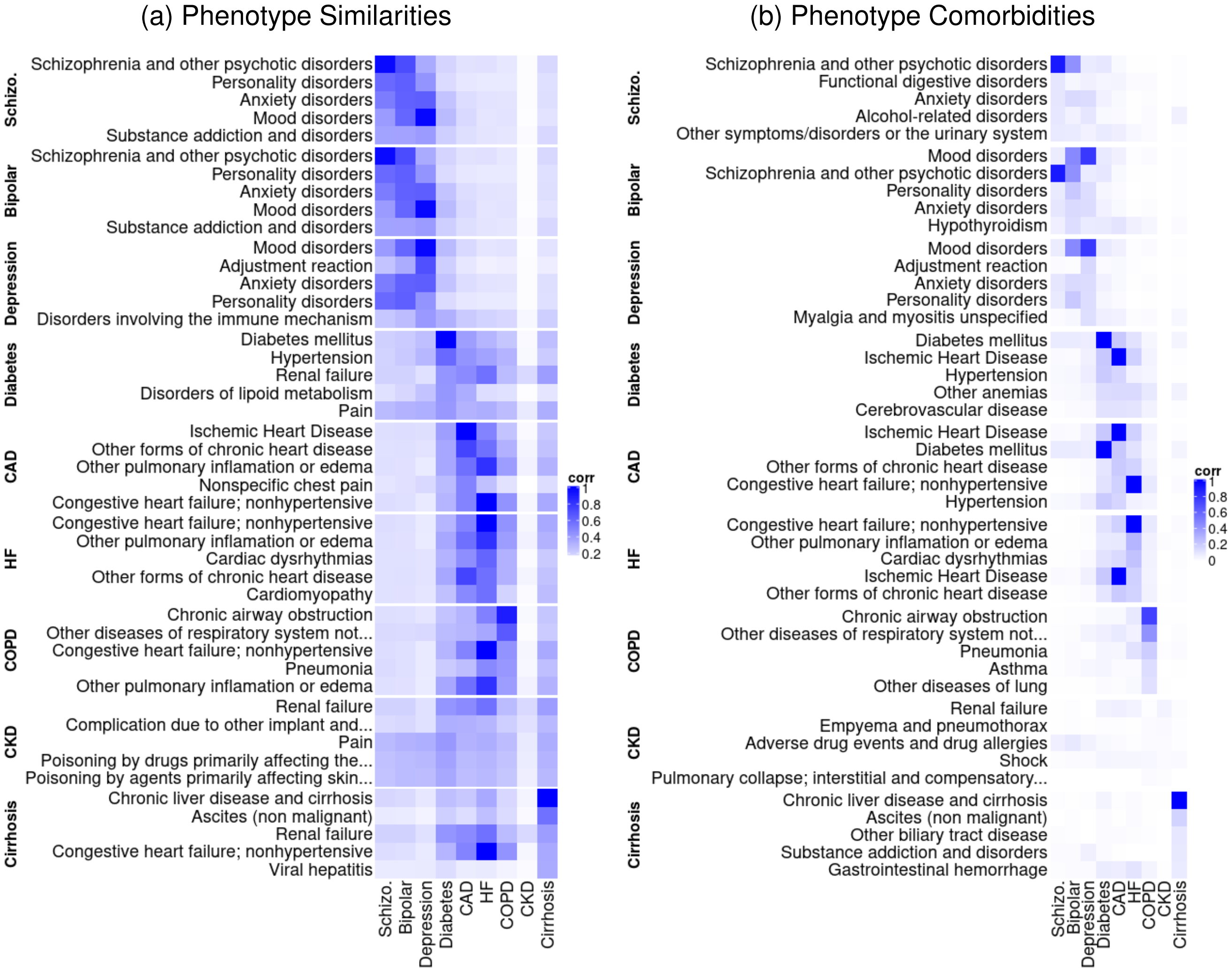


Figure 3: **Predicted similarities and comorbidities of 9 diverse phenotypes.** (**a**) 5 most similar phenotypes for each of 9 diverse disease phenotypes, as predicted by *post-hoc* Spearman correlations among MixEHR-G's patient-topic mixtures $\tilde{\Theta}$ (**b**) 5 most comorbid phenotypes per target disease as predicted by *post-hoc* Spearman correlations among their topic distributions Φ over EHR codes. 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 indicate 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.

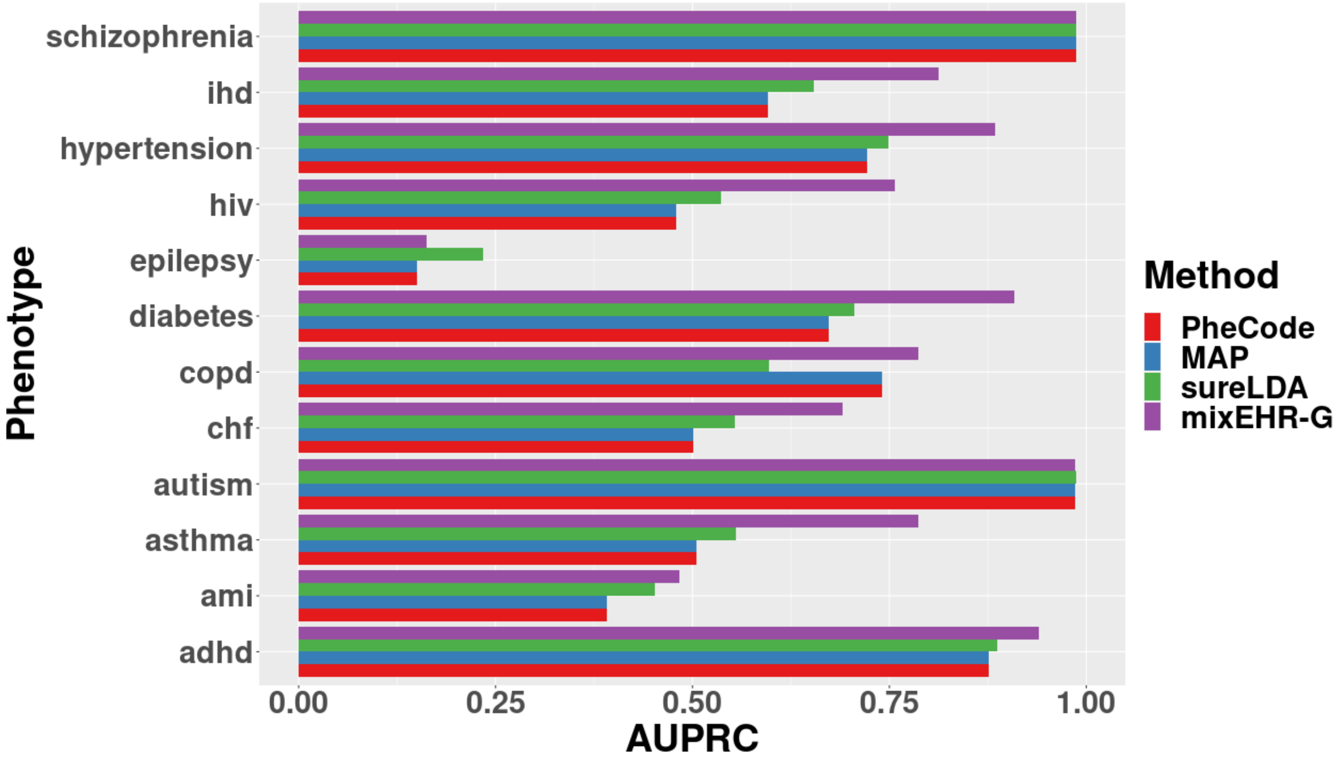


Figure 4: **Predictive performance of $\tilde{\Theta}_k$ for phenotypes with expert-derived rule-based labels.** We applied MixEHR-Gand two recently published unsupervised phenotyping methods, MAP and sureLDA, to the entire PopHR data without supervision or feature filtering. As a baseline, we also computed 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.

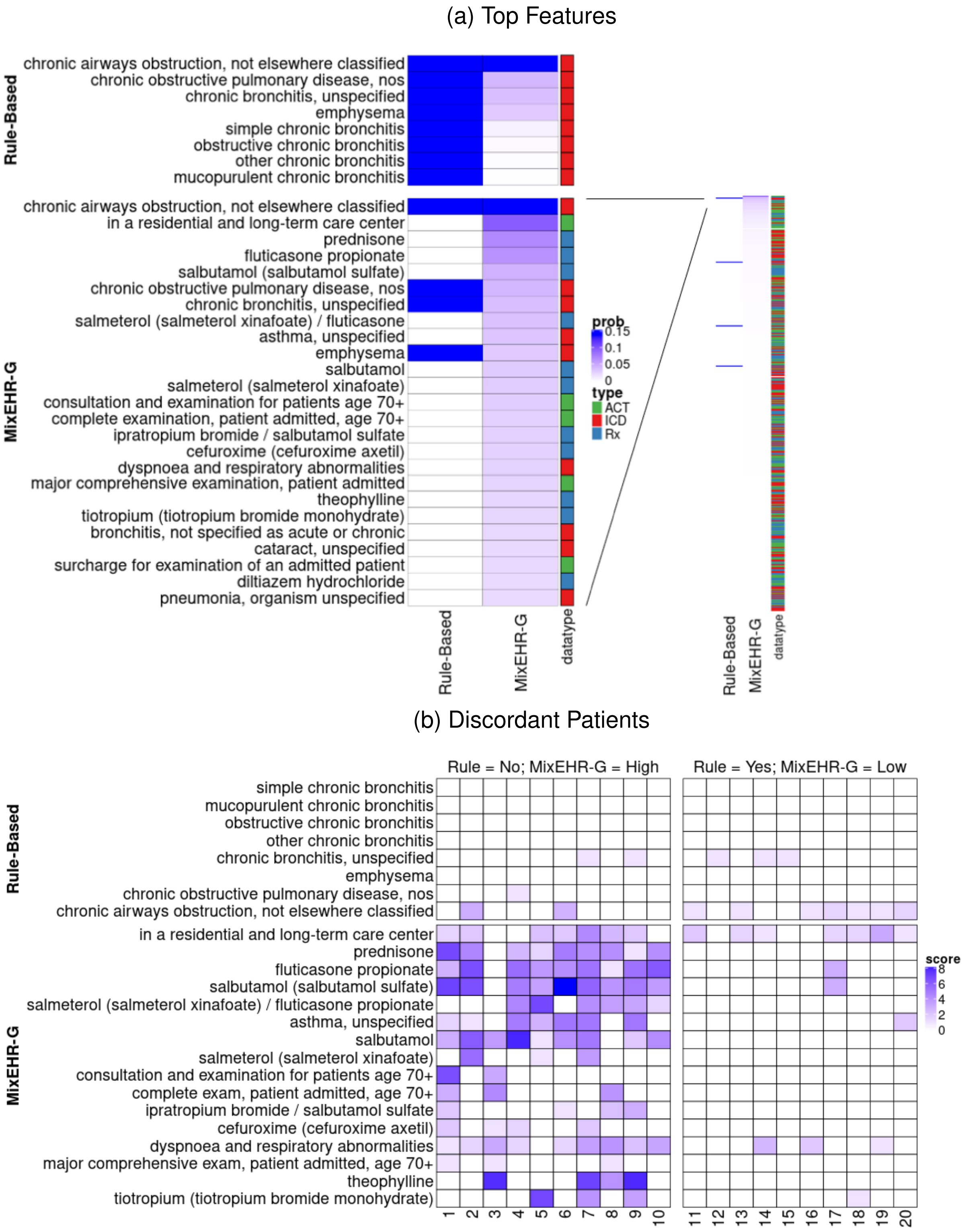
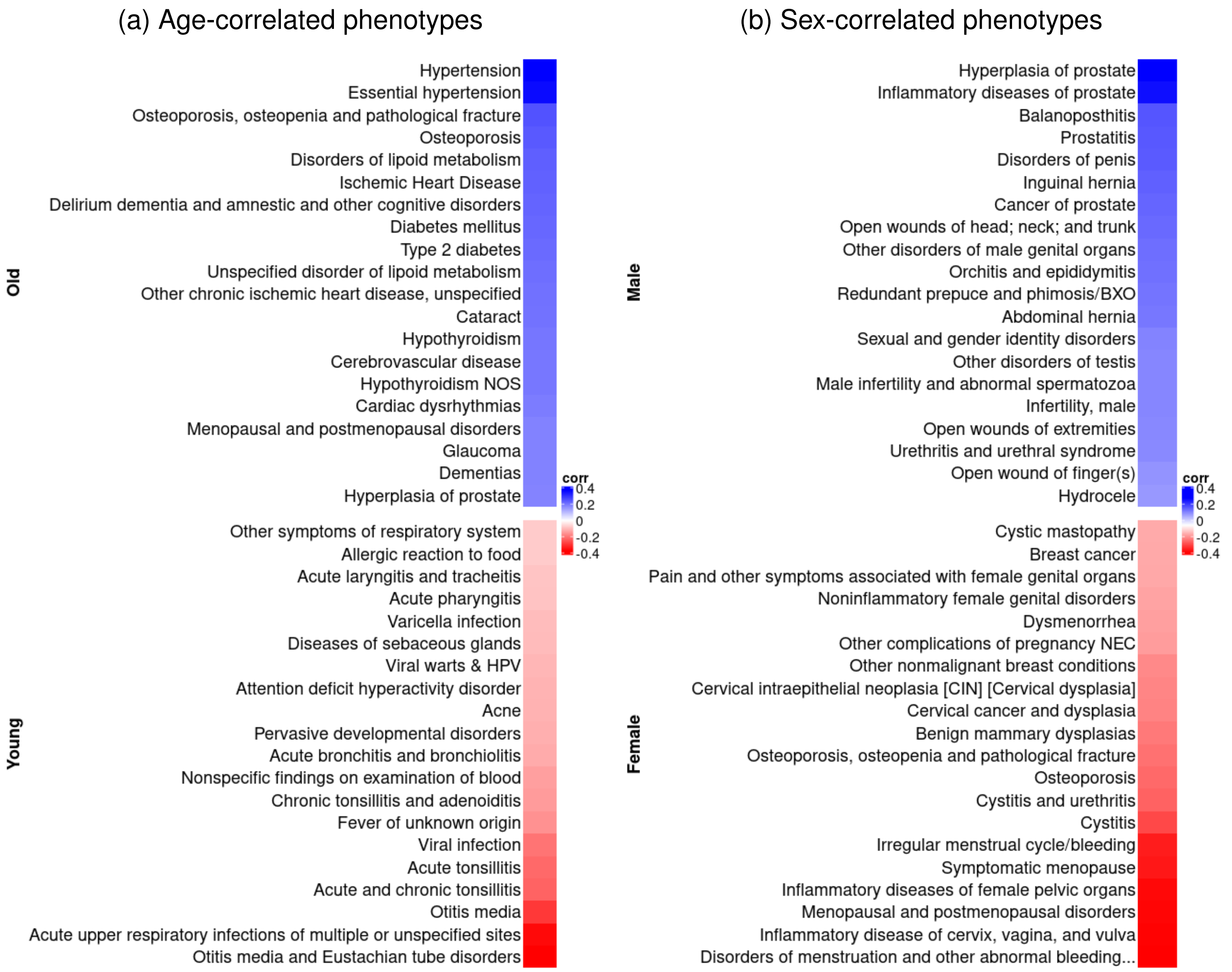


Figure 5: **In-depth analysis of MixEHR-G's phenotyping results for chronic obstructive pulmonary disease (COPD) using the PopHR dataset.** (**a**) The top 25 features and corresponding datatypes per the COPD topic distribution ϕ_{COPD}. We compared these features with the features included in the rule-based COPD algorithm shown in the heatmap on top. (**b**) Profiles of the 20 most discordant patients. The left heatmap displays the 10 patients for whom MixEHR-G predicts the highest probability of COPD but the rules predict no. The right heatmap displays the 10 for whom MixEHR-G predicts the lowest probability but the rules predict yes.



(a) Age-correlated phenotypes

(b) Sex-correlated phenotypes

Figure 6: **Top 20 phenotypes correlated with age and sex as predicted in our analysis.** (**a**) The top 20 age-correlated phenotypes. (**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.

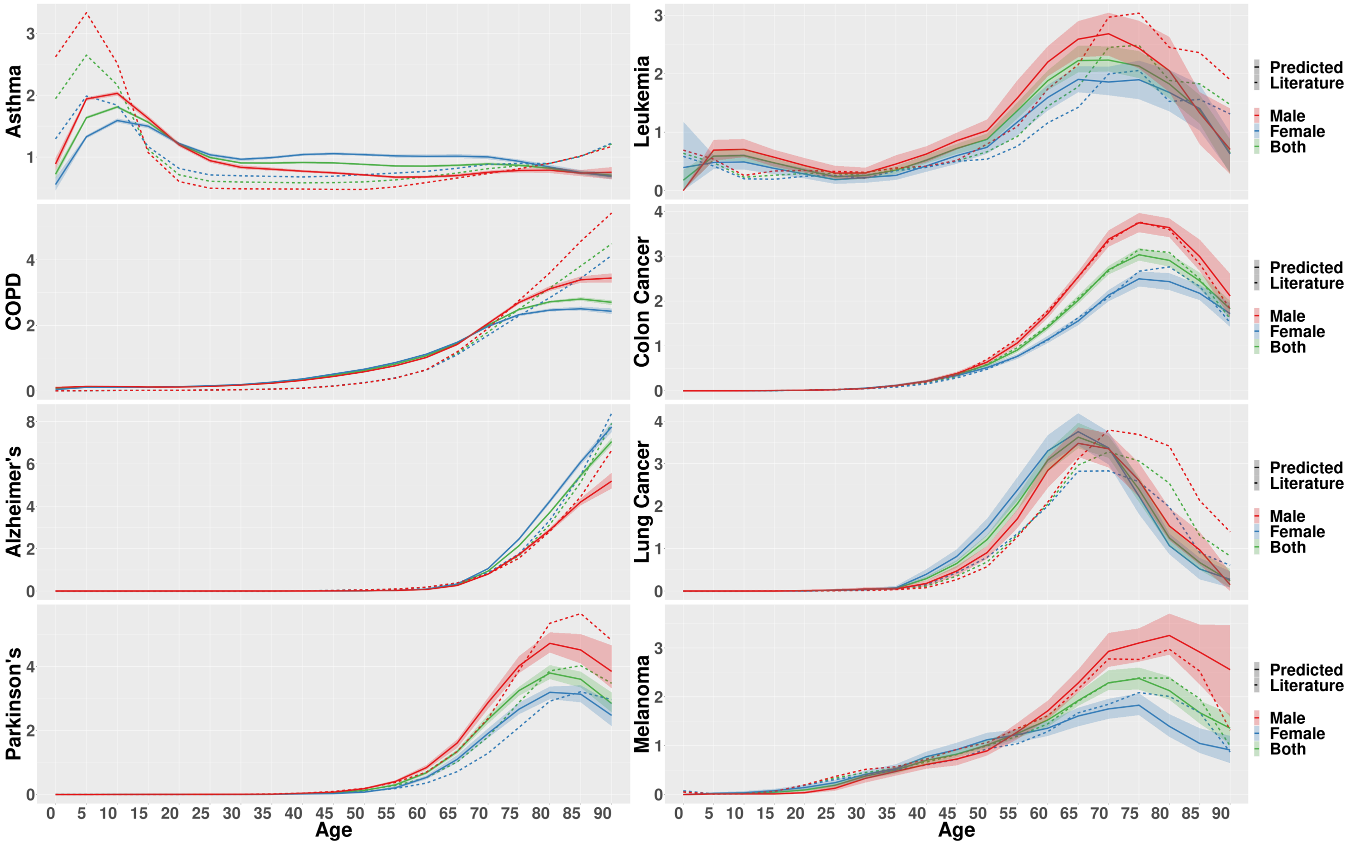


Figure 7: **Estimates of relative phenotype prevalence.** Based on the patient-phenotype mixtures $\tilde{\Theta}$, we computed phenotype prevalence stratified by age and sex for 8 diverse disease phenotypes with variable natural functions normalized by the mean estimate over time (**Methods**). We compared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

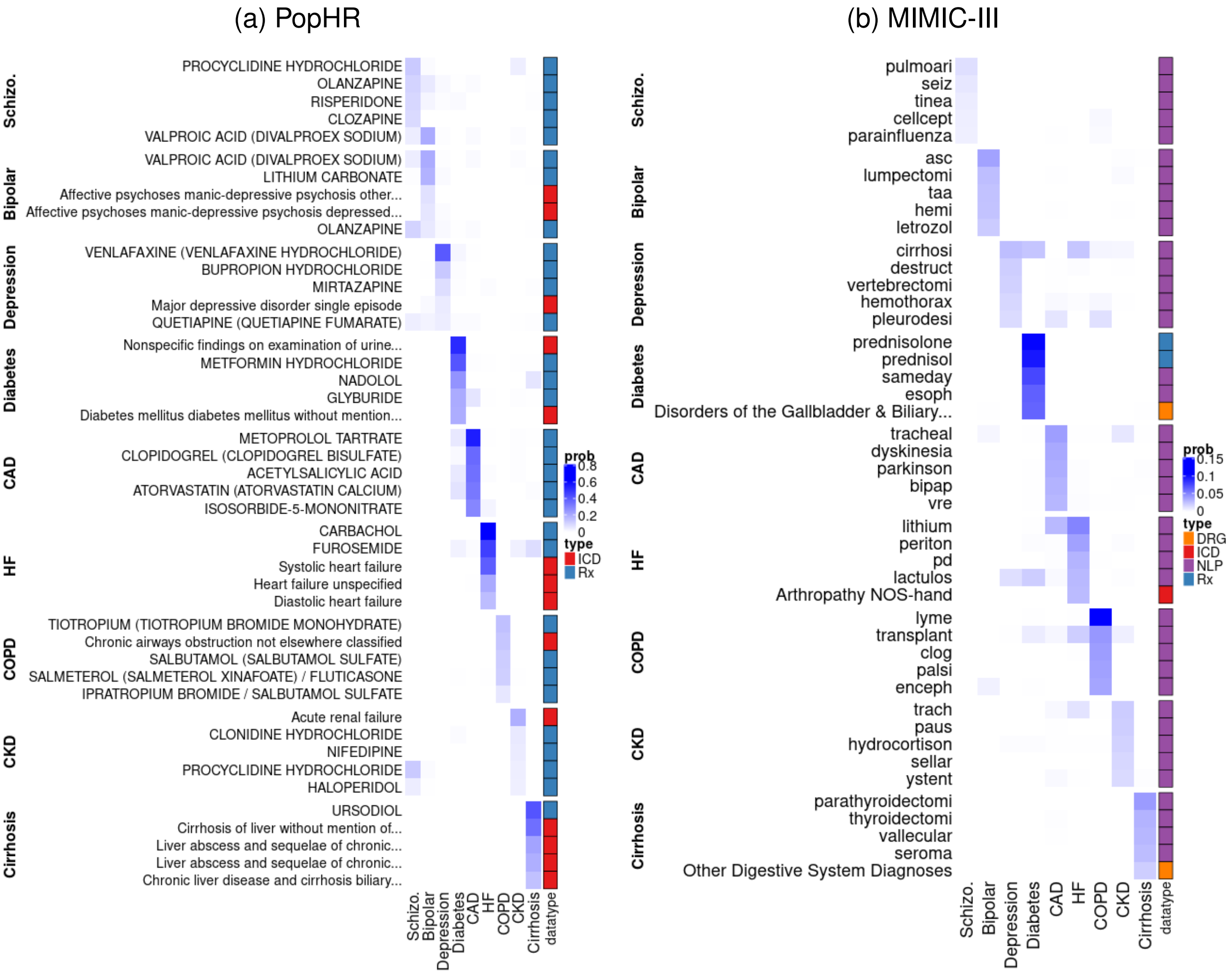
6 Tables

	Prevalence	AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F-score
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 F-score, scores were thresholded to achieve NPVs of at least 0.95.

Supplementary Materials

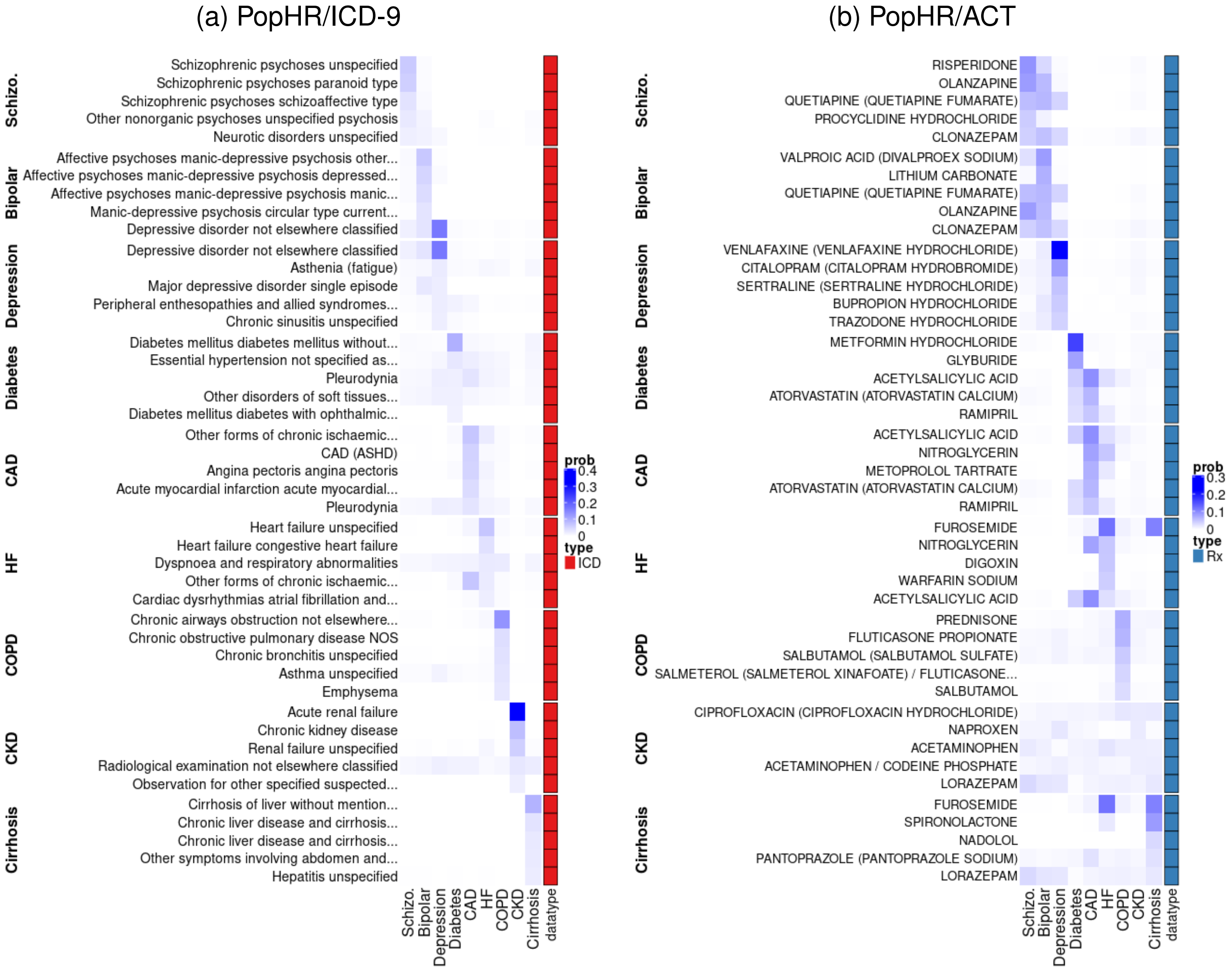
S1 Supplementary Figures

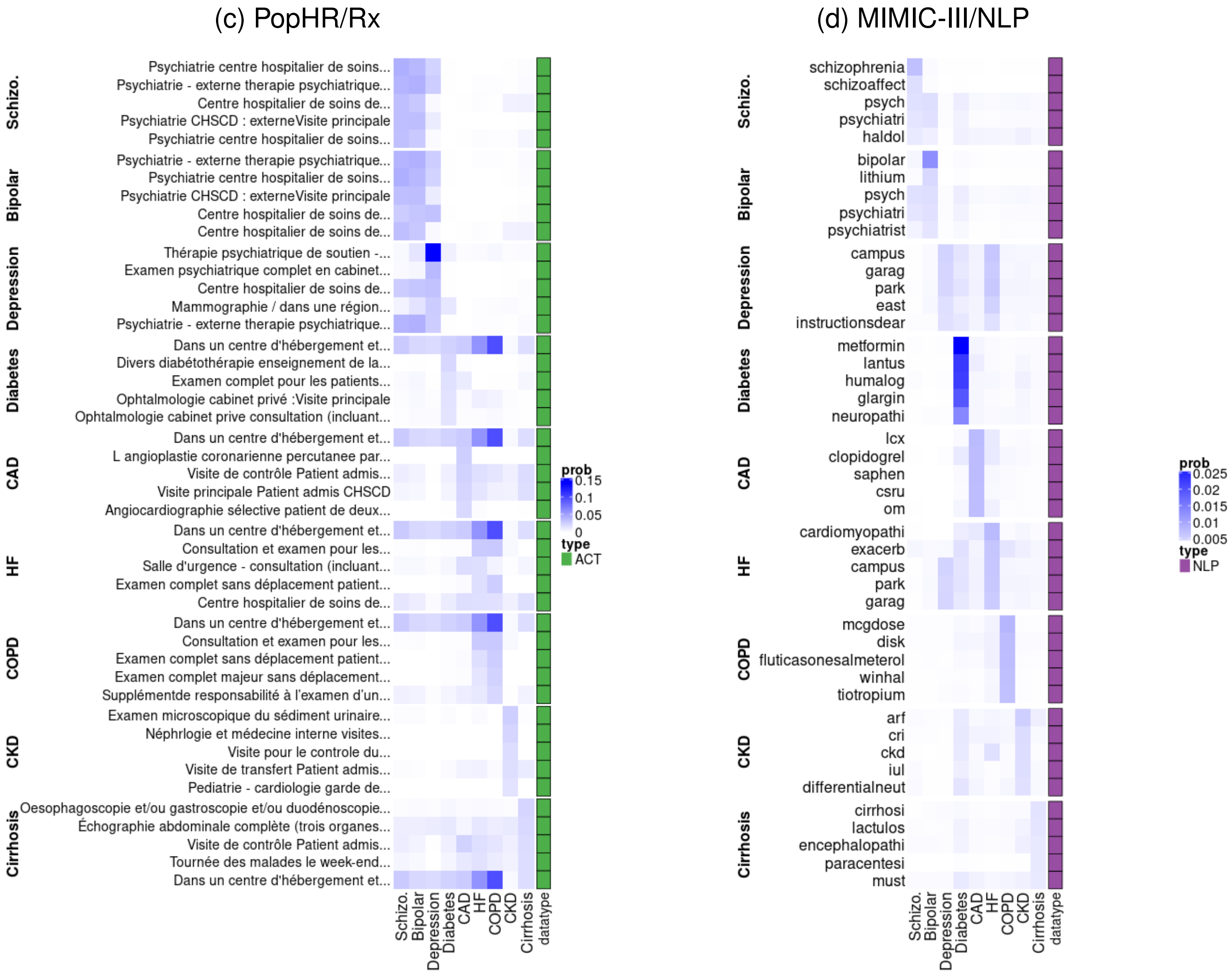


(a) PopHR

(b) MIMIC-III

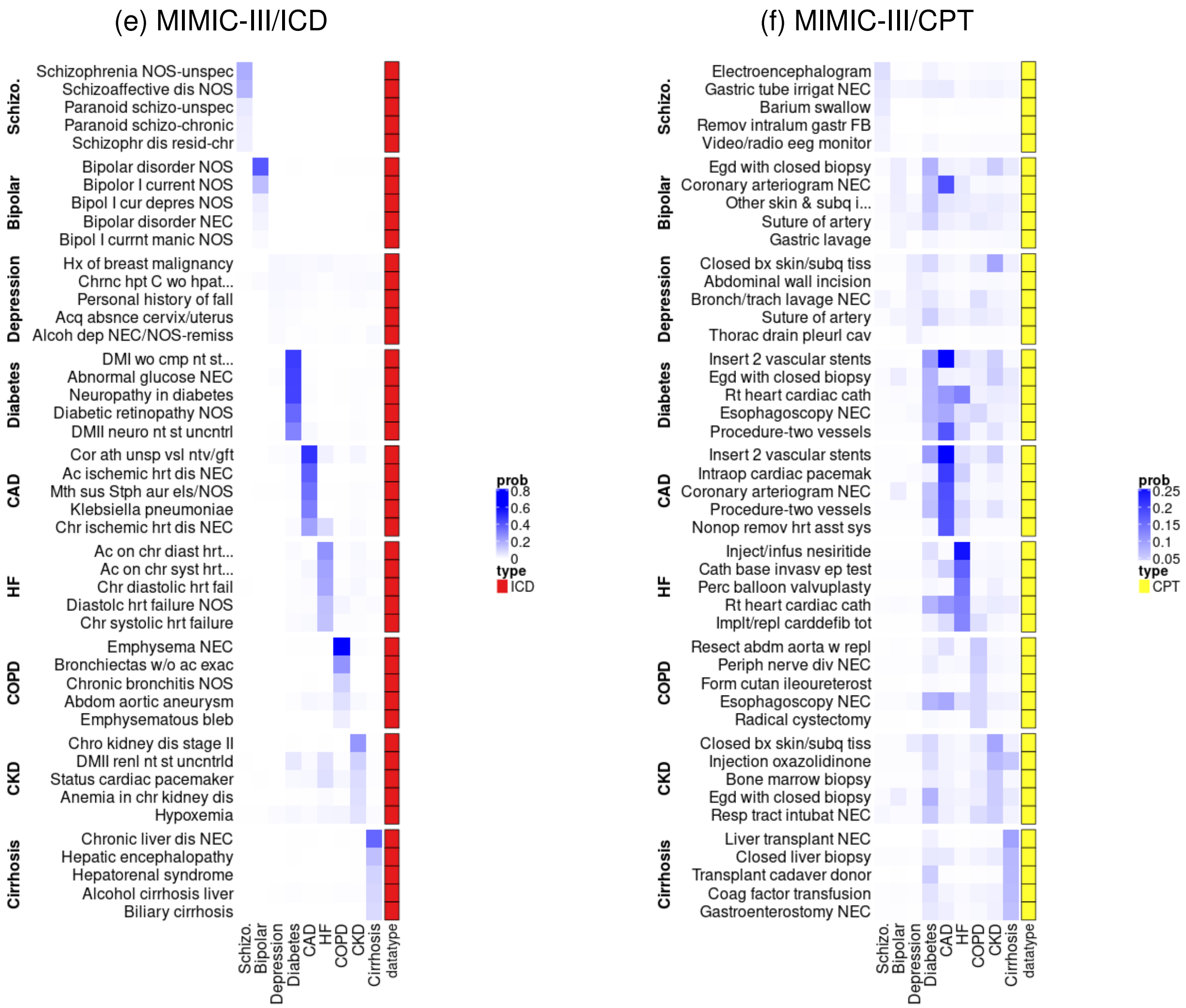
Figure S1: Top 5 features for each of 9 diverse disease phenotypes as ascertained by sureLDA, for comparison with **Fig.** 2. We present results for two datasets: (a) PopHR, which has 3 data modalities, and (b) MIMIC-III, which has 5.


(a) PopHR/ICD-9
Schizo.
Schizophrenic psychoses unspecified
Schizophrenic psychoses paranoid type
Schizophrenic psychoses schizoaffective type
Other nonorganic psychoses unspecified psychosis
Neurotic disorders unspecified
Bipolar
Affective psychoses manic-depressive psychosis other...
Affective psychoses manic-depressive psychosis depressed...
Affective psychoses manic-depressive psychosis manic...
Manic-depressive psychosis circular type current...
Depressive disorder not elsewhere classified
Depression
Depressive disorder not elsewhere classified
Asthenia (fatigue)
Major depressive disorder single episode
Peripheral enthesopathies and allied syndromes...
Chronic sinusitis unspecified
Diabetes
Diabetes mellitus diabetes mellitus without...
Essential hypertension not specified as...
Pleurodynia
Other disorders of soft tissues...
Diabetes mellitus diabetes with ophthalmic...
CAD
Other forms of chronic ischaemic...
CAD (ASHD)
Angina pectoris angina pectoris
Acute myocardial infarction acute myocardial...
Pleurodynia
HF
Heart failure unspecified
Heart failure congestive heart failure
Dyspnoea and respiratory abnormalities
Other forms of chronic ischaemic...
Cardiac dysrhythmias atrial fibrillation and...
COPD
Chronic airways obstruction not elsewhere...
Chronic obstructive pulmonary disease NOS
Chronic bronchitis unspecified
Asthma unspecified
Emphysema
CKD
Acute renal failure
Chronic kidney disease
Renal failure unspecified
Radiological examination not elsewhere classified
Observation for other specified suspected...
Cirrhosis
Cirrhosis of liver without mention...
Chronic liver disease and cirrhosis...
Chronic liver disease and cirrhosis...
Other symptoms involving abdomen and...
Hepatitis unspecified
Schizo.
Bipolar
Depression
Diabetes
CAD
HF
COPD
CKD
Cirrhosis
datatype
prob
0.4
0.3
0.2
0.1
0
type
ICD
(b) PopHR/ACT
Schizo.
RISPERIDONE
OLANZAPINE
QUETIAPINE (QUETIAPINE FUMARATE)
PROCYCLIDINE HYDROCHLORIDE
CLONAZEPAM
Bipolar
VALPROIC ACID (DIVALPROEX SODIUM)
LITHIUM CARBONATE
QUETIAPINE (QUETIAPINE FUMARATE)
OLANZAPINE
CLONAZEPAM
Depression
VENLAFAXINE (VENLAFAXINE HYDROCHLORIDE)
CITALOPRAM (CITALOPRAM HYDROBROMIDE)
SERTRALINE (SERTRALINE HYDROCHLORIDE)
BUPROPION HYDROCHLORIDE
TRAZODONE HYDROCHLORIDE
Diabetes
METFORMIN HYDROCHLORIDE
GLYBURIDE
ACETYLSALICYLIC ACID
ATORVASTATIN (ATORVASTATIN CALCIUM)
RAMIPRIL
CAD
ACETYLSALICYLIC ACID
NITROGLYCERIN
METOPROLOL TARTRATE
ATORVASTATIN (ATORVASTATIN CALCIUM)
RAMIPRIL
HF
FUROSEMIDE
NITROGLYCERIN
DIGOXIN
WARFARIN SODIUM
ACETYLSALICYLIC ACID
COPD
PREDNISONE
FLUTICASONE PROPIONATE
SALBUTAMOL (SALBUTAMOL SULFATE)
SALMETEROL (SALMETEROL XINAFOATE) / FLUTICASONE...
SALBUTAMOL
CKD
CIPROFLOXACIN (CIPROFLOXACIN HYDROCHLORIDE)
NAPROXEN
ACETAMINOPHEN
ACETAMINOPHEN / CODEINE PHOSPHATE
LORAZEPAM
Cirrhosis
FUROSEMIDE
SPIRONOLACTONE
NADOLOL
PANTOPRAZOLE (PANTOPRAZOLE SODIUM)
LORAZEPAM
Schizo.
Bipolar
Depression
Diabetes
CAD
HF
COPD
CKD
Cirrhosis
datatype
prob
0.3
0.2
0.1
0
type
Rx

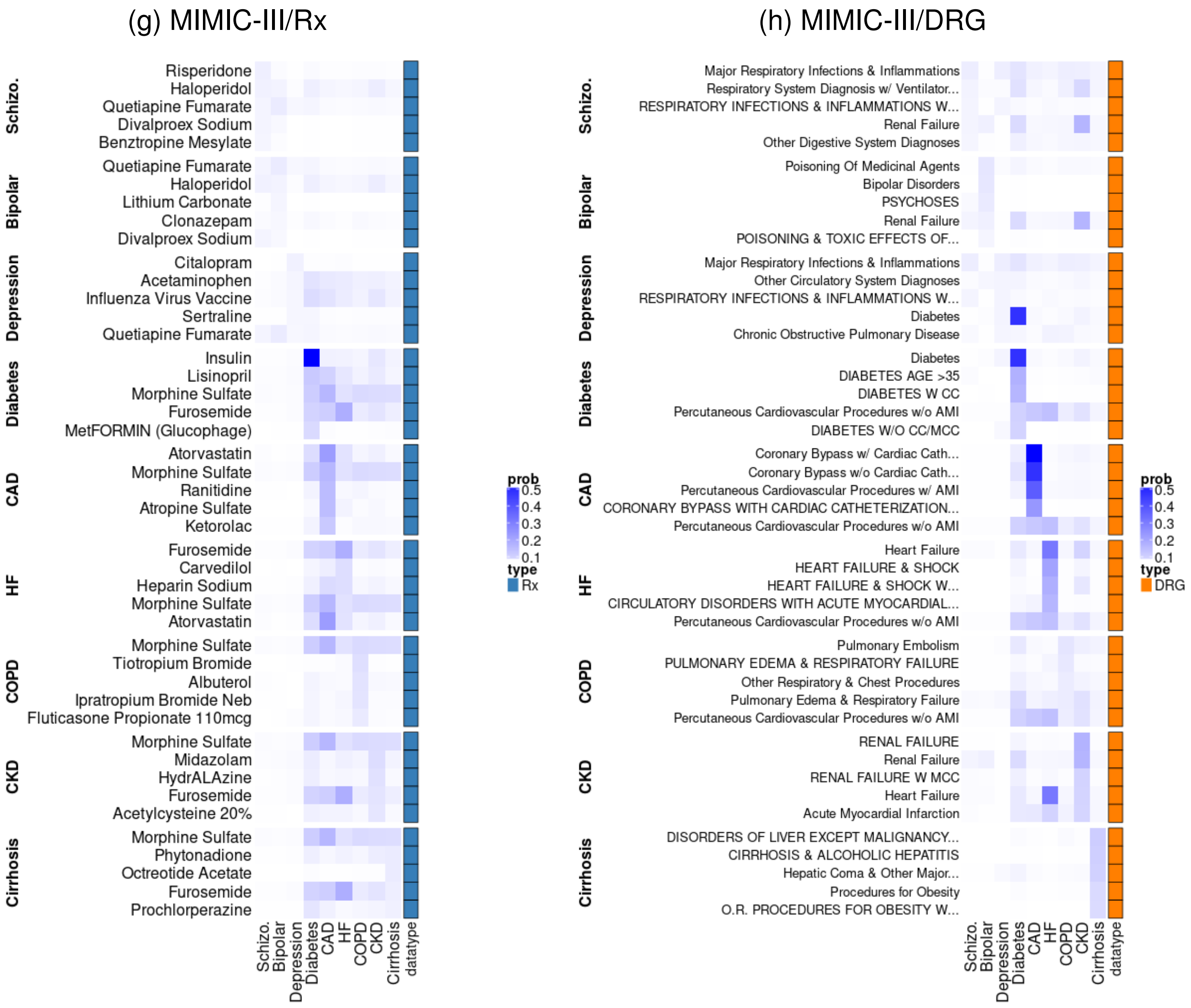


(c) PopHR/Rx

(d) MIMIC-III/NLP


(e) MIMIC-III/ICD

(f) MIMIC-III/CPT


(g) MIMIC-III/Rx

(h) MIMIC-III/DRG

Figure S2: 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.

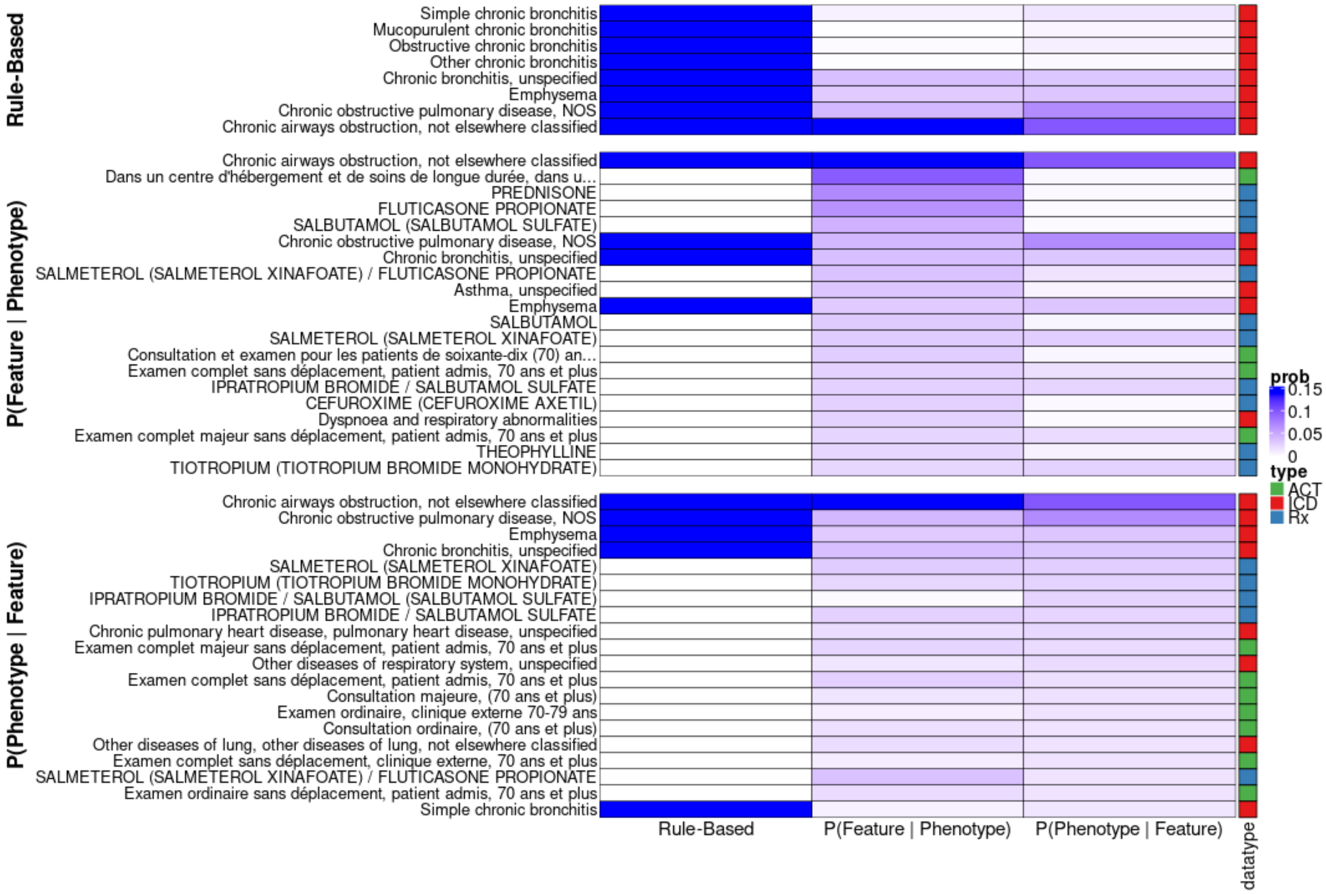


Figure S3: Top 25 features with the highest (1) sensitivities {P(Feature | Phenotype)} and (2) positive predictive values {P(Phenotype | Feature)} for COPD based on MixEHR-G's trained topic distributions ϕ, as compared to features included in the rule-based COPD algorithm.