Phenotype Inference with Semi-Supervised Mixed Membership Models

Victor A. Rodriguez victor.a.rodriguez@columbia.edu
Columbia University
Department of Biomedical Informatics
New York City, NY, USA

Adler Perotte adler.perotte@columbia.edu
Columbia University
Department of Biomedical Informatics
New York City, NY, USA

Abstract

Disease phenotyping algorithms are designed to sift through clinical data stores to identify patients with specific diseases. Supervised phenotyping methods require significant quantities of expert-labeled data, while unsupervised methods may learn spurious or non-disease phenotypes. To address these limitations, we propose the Semi-Supervised Mixed Membership Model (SS3M) – a probabilistic graphical model for learning disease phenotypes from partially labeled clinical data. We show SS3M can generate interpretable, disease-specific phenotypes which capture the clinical features of the disease concepts specified by the labels provided to the model. Furthermore, SS3M phenotypes demonstrate competitive predictive performance relative to commonly used baselines.

1. Introduction

Phenotypes are powerful tools for working with observational clinical data in the absence of reliable disease labels (Hripcsak and Albers, 2012). Disease-specific phentoypes allow researchers to sift through large-scale clinical data stores to identify patients with evidence of specific clinical conditions. By answering the question of who has what disease, phenotypes power essential tasks such as cohort selection, trial recruitment and clinical outcome prediction (Hripcsak and Albers, 2012; Richesson et al., 2013, 2016; Pathak et al., 2013).

Traditionally, phenotypes were developed by groups of clinical experts who painstakingly hand-tuned rule-based algorithms. The limited scalability of this approach has led to the development of automated methods for learning phenotypes directly from clinical data. Many studies in this vein utilize supervised machine learning methods to build phenotyping algorithms (Bergquist et al., 2017; Esteban et al., 2017). Though this approach avoids laborious expert knowledge engineering, it requires significant amounts of labeled clinical data generated by manual chart review.

To avoid costly, expert-generated disease labels, many authors have utilized unsupervised methods to cluster patients according to underlying patterns in their clincal data (Joshi et al., 2016; Ho et al., 2014a,b; Wang et al., 2015; Miotto et al., 2016). In this setting, such patterns play the role of phenotypes. Unsupervised phenotyping methods often learn multiple phenotypes simultaneouly, which may confer evidence of specific diseases. However, such phenotypes are generally not guaranteed to represent single disease concepts. This complicates their evaluation and use in downstream tasks.

In this paper we propose the Semi-Supervised Mixed Membership Model (SS3M), a probabilistic graphical model which utilizes relatively few disease labels to learn multiple disease-specific phenotypes from multi-modal observational clinical data. SS3M addresses the limitations of supervised phenotyping by reducing the amount of labeled data needed to learn disease phenotypes; disease labels are not required for all patients, and labeled patients need not possess labels for all diseases. SS3M also addresses the limitations of unsupervised phenotyping by associating disease labels with the phenotypes to be learned; a label specifies which disease a phenotype is meant to represent. This simplifies both the qualitative and quantitative evaluation of SS3M phenotypes. Qualitatively, phenotype labels inform us as to what content we should expect to be well represented within a learned phenotype. Quantitatively, we can evaluate how well learned phenotypes predict labels on a held-out patient cohort using standard performance metrics.

Technical Significance SS3M introduces a mechanism for semi-supervised learning within a class of unsupervised mixed membership models developed for inferring phenotypes from multi-modal clinical data. In so doing, SS3M permits the researcher to specify which phenotypes she would like the model to infer. Importantly, this added input is minimal: only a subset of cases need to be labeled as positive or negative for the model to learn a disease-specific phenotype.

Clinical Relevance The secondary use of observational clinical often depends critically upon the creation of disease-specific phentoypes. SS3M provides a method for constructing phenotypes efficiently. Instead of hand tuning phenotype definitions, the user need only identify a handful of patients which have or do not have a specific disease. Thus, SS3M serves as a tool for widening this persistent bottleneck in clinical research and development of clinical applications.

2. Cohort

We train all our models using clinical data extracted from the Medical Information Mart for Intensive Care version III (MIMIC-III) (Johnson et al., 2016).

2.1. Cohort Selection

Our dataset is restricted to adult patients where adults are defined as patients who are 18 years of age or older upon admission. Age upon admission is calculated by subtracting each patient's recorded date of birth from their time of admission. This constraint yields a cohort of 38,549 individual patients.

2.2. Data Extraction

Each patient is represented by multiple sets of clinical observations (one set per data source) and a set of labels (possibly empty). We limit ourselves to the clinical observations and labels associated with each patient's first hopsital admission.

Clincal observations are drawn from clinical notes, labs and medications. We refer to these data types as data sources. Notes were restricted to the following types: "Physician", "General", and "Discharge Summary". No restrictions were placed on clinical labs and medications.

In this work, we lack a set of true, expert-generated, gold-standard disease labels for our patient cohort. For this reason we make use of readily availble ICD9 diagnosis codes to contstruct our label set. Our labels correspond to a variety of disease conditions from the single-level definitions of the Health Cost and Utilization (HCUP) Clinical Classification Software (CCS). The HCUP CCS conditions are defined by groups of related ICD9 diagnosis codes. Relative to raw ICD9 codes, HCUP CCS code groups are significantly less noisy, which makes them attractive for phenotype prediction tasks in the absence of a true gold-standard. We apply all HCUP CCS single-level definitions to the ICD9 cdoes for our cohort and consider conditions with a least 10^3 positive cases (prevalence $\approx 2.5\%$). As MIMIC-III is a critical care database, we further limit ourselves to well represented acute conditions. This process led us to retain a total of 40 conditions for use in our experiments (See Table 3 and Figure 5 for our full list of HCUP CCS conditions). For each patient, we record a binary label for each of these disease conditions specifying its presence (1) or absence (0). We treat this label set as ground truth.

2.3. Feature Choices

For a given patient, we concatenate all associated clinical observations within each data source. These observations are tokenized to yield a patient's raw token representation in terms of words (from notes), lab names and medication names.

Tokenized notes are further preprocessed to remove English stop words as well as any word token with 20 appearances or less over the entire notes corpus. This latter step is intended to a filter out the large quantity of misspelled words observed in the unfiltered token vocabulary.

The notes vocabularly is further constrainted by applying a term-frequency/inverse-document-frequence (TF-IDF) filter. For each patient d, and each token t observed in their tokenized set of notes we calculate a tf-idf weight w_{dt} as follows.

$$w_{dt} = N_{dt} \log_2 \frac{D}{N_t}, \tag{1}$$

where N_{dt} is the number of times token t appears in patient d's tokenized notes, N_d is the total number of patient d's note tokens, and D is the total number of patients. Next, we average these weights over all patients and retain the top 10^4 mean weighted tokens. No additional preprocessing was applied to clinical labs and medications post-tokenization.

Algorithm 1 Generative process for SS3M

```
Initialize: $\alpha$, $\beta$, $\beta^*$, $\{\gamma_s\}_{s=1}^{S}$

# Sample global variables
Sample $B^* \sim \text{Gamma}(\beta^*)$
for each phenotype $p = 1$ to $P$ do
    Sample $\boldsymbol{B}_p \sim \text{Gamma}(\beta)$
    Sample $\boldsymbol{C}_p \sim \text{Beta}(\alpha)$

    for each data source $s = 1$ to $S$ do
        Sample $\boldsymbol{\Phi}_{s,p} \sim \text{Dirichlet}(\gamma_s)$
    end
end
# Sample local variables
for each patient $d = 1$ to $D$ do
    for each phenotype $p = 1$ to $P$ do
        Sample $\boldsymbol{A}_{d,p} \sim \text{Bernoulli}(\boldsymbol{C}_p)$
    end
    Sample $\boldsymbol{\Theta}_d \sim \text{Dirichlet}(\boldsymbol{A}_{d,:} \odot \boldsymbol{B}_{:} + (\mathbb{1} - \boldsymbol{A}_{d,:})B^*)$
    for each data source $s = 1$ to $S$ do
        for each observation $n = 1$ to $N_{sd}$ do
            Sample $\boldsymbol{Z}_{sdn} \sim \text{Categorical}(\boldsymbol{\Theta}_d)$
            Sample $\boldsymbol{W}_{sdn} \sim \text{Categorical}(\boldsymbol{\Phi}_{s\boldsymbol{Z}_{sdn}})$
        end
    end
end
```

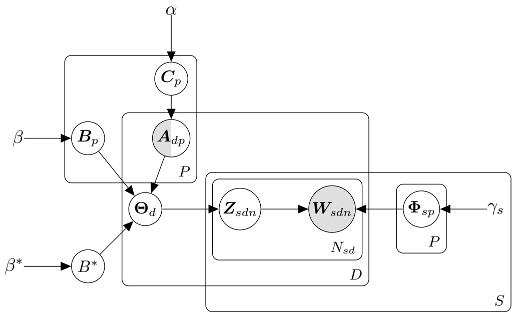


Figure 1: Graphical model for SS3M

Variable	Description
D	Number of patients
S	Number of data sources
P	Number of phenotypes
N_{sd}	Number of tokens of source s for patient d
V_s	Size of vocabulary for source s
$\boldsymbol{A}$	Phenotype activations (partially observed)
$\boldsymbol{B}$	Active phenotype parameters
B^*	Inactivate phenotype parameter
$\boldsymbol{C}$	Phenotype prevalences
$\boldsymbol{\Theta}$	Patient-phenotype distribution parameters
$\boldsymbol{\Phi}$	Phenotype-token distribution parameters
$\boldsymbol{Z}$	Phenotype assignments
$\boldsymbol{W}$	Token observations
α	Hyperparameters
β	
β^*	
γ	

Table 1: SS3M variable descriptions

3. Methods

SS3M is a model for bag-of-words of data from multiple data sources. It is closely related to Latent Dirichlet Allocation (LDA) (Blei et al., 2003), as well as its multisource (Pivovarov et al., 2015) and supervised extensions (Ramage et al., 2009). When applied to clinical data, SS3M treats patients as individual instances, where each instance is comprised of observations from multiple clinical data sources (i.e. clinical notes, labs and medications) as well as a set of disease labels. SS3M processes clinical observations to infer phenotypes which capture the clinical characteristics of the conditions specified by labels its given.

Model Description Here we provide a detailed description of SS3M's structure. The model's generative process and graphical model provide complementary perspectives and are detailed in Algorithm 1 and Figure 1 respectively. Table 1 provides descriptions of all model variables. Here and in the rest of the paper, we use bold capital letters to indicate groups of variables and indices to refer to subsets or specific elements. A bold capital letter without indices indicates all variables within the group. A colon within a variable subscript indicates all elements within the corresponding dimension.

Let D, S, and P be the number of patients, clinical data sources and phenotypes, respectively. Each patient $d \in \{1, ..., D\}$ is associated with several sets of tokenized clinical observations $\boldsymbol{W}_{sd:}$ (e.g. medication names), one for each data source $s \in \{1, .., S\}$. In addition, each patient has a set of partially observed binary labels. Patient d's labels specify the values of her phenotype activations, $\boldsymbol{A}_{d,:}$, thereby indicating for her which phenotypes $p \in \{1, ..., P\}$ are set to be "on" or "off". A latent phenotype assignment $\boldsymbol{Z}_{sdn}$ is assigned to each observation $\boldsymbol{W}_{sdn}$. Each assignment is drawn from a categorical patient-phenotype distribution parameterized by a normalized P-dimensional vector $\boldsymbol{\Theta}_d$. A phenotype assignment specifies which categorical phenotype-token distribution an observation was drawn from. Each $\boldsymbol{\Phi}_{sp}$ is a normalized V_s-dimensional vector parameterizing a categorical phenotype-token distribution, where V_s is the size of the observered vocabulary for data source s.

A patient's label set directly impacts her patient-phenotype distribution, and thereby all her assignments. This is due to the roles of $\boldsymbol{A}$, $\boldsymbol{B}$ and B^* in parameterizing the Dirichlet distributions on the elements of $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta}_d \sim \text{Dirichlet}(\boldsymbol{A}_{d,:} \odot \boldsymbol{B}_{:} + (\mathbb{1} - \boldsymbol{A}_{d,:})B^*)$$

where $\odot$ indicates element-wise multiplication. When $\boldsymbol{A}_{dp} = 1$, patient d has phenotype p "on"; $\boldsymbol{B}_p$ is used to parameterize the p^{th} dimension of the Dirichlet on $\boldsymbol{\Theta}_d$. When $\boldsymbol{A}_{dp} = 0$, the phenotype is "off", and B^* is used instead. The hyperparameters β and β^* parameterize the gamma distributions on $\boldsymbol{B}$ and B^* such that the model is encouraged to sample values of $\boldsymbol{B}_p$ and B^* to maintain $\boldsymbol{B}_p > 1$ and $B^* < 1$. In this setting, when $\boldsymbol{A}_{dp} = 1$ the values of $\boldsymbol{\Theta}_d$ push the patient-phenotype distribution toward allocating more probablity mass for phenotype p. This in turn, results in a larger proportion of patient d's observations being assigned to phenotype p. During inference, this mechansim forms the connection between labels, activations and the content of phenotypes. Labels set phenotype activations "on" or "off" for each patient. For each patient, phenotypes that are "on" account for the majority of phenotype assignments. Thus, labels, by way of activations, significantly influence the quantity of observations that are funneled toward learning any given phenotype.

Activations are *partially observed*. If a patient d has an observed binary label for phenotype p, then the value of $\boldsymbol{A}_{dp}$ is held fixed at the observed value. If the label is unobserved, then the model samples the value of $\boldsymbol{A}_{dp}$ during inference. In this latter case, $\boldsymbol{A}_{dp}$ is modeled as a binary variable drawn from a Bernoulli distribution parameterized by $\boldsymbol{C}_p$ – a beta distributed latent variable controlling the likelihood of phenotype p being "on" within the patient population (i.e. estimates its prevalence). This handling of partially observed labels is what allows SS3M to function as a semi-supervised model.

SS3M can handle both semi-supervised phenotypes for which we have some number of labels, as well as unsupervised phenotypes that lack labels all together. This is a useful property when applying the model to clinical data. In this setting we are unlikely to have labels for all the conditions represented in our dataset. The structure of the conditions we lack labels for can be targeted by SS3M's unsupervised phenotypes during inference. Moreover, including unsupervised phenotypes can help semi-supervised phenotypes "focus" on capturing the phenotypes that align with their labels.

Inference We implement a collapsed Gibbs sampler to obtain posterior estimates of our model's latent variables. The variables $\boldsymbol{C}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Phi}$ are easily integrated out of the joint distribution due to conjugate relationships between their distributions and those on $\boldsymbol{A}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ respectively. The collapsed joint's complete conditional distributions for the elements of $\boldsymbol{A}$ and $\boldsymbol{Z}$ are discrete, easily normalized, and can be sampled from directly. However, the complete conditionals for $\boldsymbol{B}$ and B^* do not have closed forms. We use Hamiltonian Monte Carlo to sample from these (Neal et al., 2011). We set out path length to $L = 15$ and step size to $\epsilon = 10^{-3}$, as these parameters yielded stable trajectories with high acceptance rates in preliminary experiments. See the appendix for further details regarding our inference algorithm.

4. Results

4.1. Experiments with Simulated Data

Here we evaluate SS3M's ability to recover ground truth phenotypes when provided observations and labels from a cohort of simulated patients.

4.1.1. Data Simulation

We create simulated cohorts by drawing observations and labels from our model. To begin, we define 10 ground truth phenotypes, $\boldsymbol{\Phi}_{\text{true}}$, in a manner inspired by (Griffiths and Steyvers, 2004). Each phenotype is a set of three categorical distributions defined over three seperate

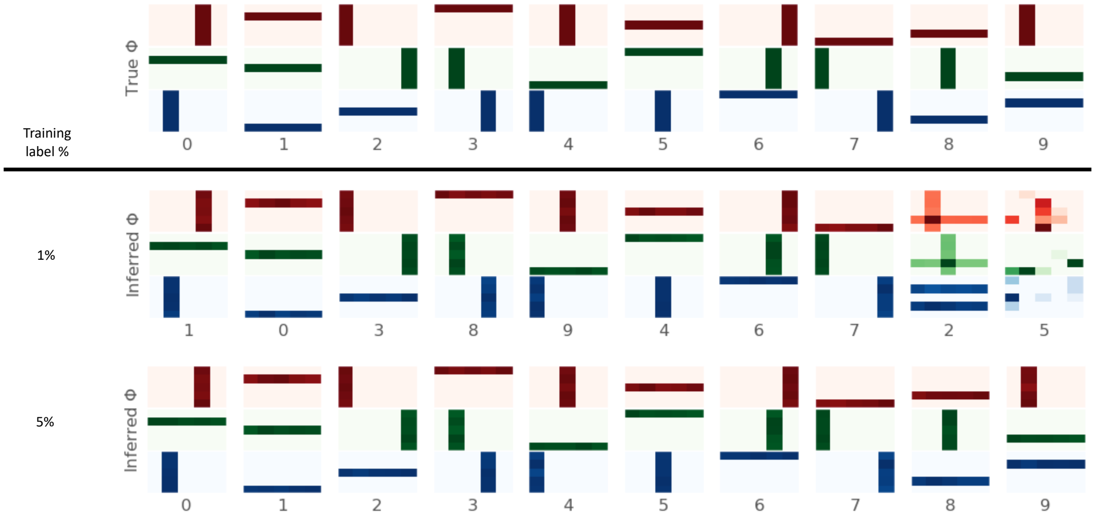


Figure 2: Phenotype inference for simulated patient cohort. **Top row**: Ground truth phenotypes. **Middle & Bottom rows**: Phenotypes inferred using 1% and 5% of ground truth labels for training. When training on 5% of available labels, SS3M recovers ground truth phenotypes; phenotype-token distributions are recovered *and* in the correct order.

vocabularies each of length 25. This allows us to visualize each phenotype as a set of three 5×5 grids. Each component of a phenotype places uniform probability mass over 5 tokens corresponding to a row or column in the grid. Next, we simulate ground truth labels for each patient. This is done by first drawing values for $\boldsymbol{C}$ from Beta distributions parameterized with $\gamma = (10^2, 10^3)$. This ensures each phenotype is active in about 10% of the cohort. We then use the values of $\boldsymbol{C}$ to draw an array of ground truth labels $\boldsymbol{A}_{\text{true}}$. The values of $\boldsymbol{B}$ and B^* are set to $(10., ..., 10.)$ and 10^{-2}, respectively. These values ensure observations are highly likely to be drawn from active phenotypes. Finally, we draw values for $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ which, along with $\boldsymbol{\Phi}_{\text{true}}$, are used to generate our simulated observations, $\boldsymbol{W}$.

4.1.2. Phenotype Inference

We expose SS3M to simulated data and run our inference algorithm to recover the ground truth phenotypes. We use the same training set of observations for each of our experiments. We produce a label set for each experiment by downsampling ground truth labels in $\boldsymbol{A}_{\text{true}}$. We run 2 experiments in which we retain 1% and 5% of positive labels. We then sample negative labels to match the total number of positive lables for a given phenotype.

4.2. Simulated Data Results

Figure 2 contains the results of simulated studies. When training on 1% of available labels, SS3M struggles to recover ground truth. Some of the inferred phenotypes appear to be superpositions of multiple ground truth phenotypes. Though some of the phenotype-token distributions do indeed mirror ground truth, many of the indices are mismatched. Full recovery of ground truth requires both the recovery of phenotype structure as well as phenotype identity. Both of these requirements are met when SS3M is exposed to just 5% of available lables. For our dataset of 1000 simulated patients, 5% of labels corresponds to 14-15 labeled patients per phenotype – half labeled positive and half labeled negative.

4.3. Experiments with Clinical Data

We are interested in evaluating SS3M's ability to learn clincally meaningful phenotypes and perform phentoype prediction on held-out patient data. Moreover, we aim to evaluate SS3M's performance in these tasks when trained on various proportions of labeled patient data.

In the present setting, each patient has a full set of binary labels for each of our 40 HCUP CCS condition targets. These labels are treated as ground truth, and we train SS3M with subsets of them. To obtain each subset, we first specify a percentage of the training cohort for which we wish to retain labels. We then sample the corresponding number of patients from the training cohort and ensure the prevalence of each label in the labeled subset is similar to that in the total training cohort. During training, we use the full set of labels for each patient in the labeled subset. We carry out this process for various percentages of the training cohort including 1%, 5%, 25%, 50%, 75%, and 100%.

As described in Section 3, SS3M handles both semi-supervised and unsupervised phenotypes. In preliminary experiments, we observed SS3M's performance depended in part on the total number of phenotypes modeled, P. To characterize this dependency, we train

SS3M on the label subsets described above with P set to 40 (i.e. no unsupervised phenotypes), 80 or 160.

Our total training cohort is comprised of 60% of the patient cohort described in Section 2. The remaining 40% is reserved for validation (20%) and testing (20%).

4.3.1. Quantitative evaluation

For each labeled subset and value of P, we obtain posterior estimates of SS3M's global latent variables ($\boldsymbol{B}$, B^*, $\boldsymbol{C}$, and $\boldsymbol{\Phi}$) by running our collapsed Gibbs sampler on the training data. These global variables are then passed to untrained SS3M models for which we run a partially collapsed Gibbs sampler (only $\boldsymbol{\Theta}$ is integrated out of the joint distribution) over the local latent variables ($\boldsymbol{A}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$) on the test set. Within the held out set, the complete conditional likelihoods on each activation ($\boldsymbol{A}_{dp}$) are used as label prediction probabilities which we evaluate using the areas under the receiver operating characteristic and precision-recall curves (AUC-ROC, AUC-PR).

We compare SS3M's predictive performance to that of several commonly used baselines. These include k-nearest neighbors (KNN) and random forests (RF), which we train as multilabel classifiers. We also compare against L1-regularized logistic regression (LR) trained as a set of 40 one-versus-rest classifiers, one for each target. Unlike, SS3M, our baselines were not developed to handle partially labeled training data. Thus, for any given configuration of the training cohort, we train baselinee on data for only those patients whose labels are included within the labeled subset.

Performance curves and baselines are estimated using the Scikit-learn Python library (Pedregosa et al., 2011).

4.3.2. Qualitative evaluation

Here we ask clinical experts to asses the quality of SS3M phenotypes relative to phenotypes inferred with a Multi-Channel Mixed Membership Model (MC3M), a closely related unsupervised model developed for phenotype inference (Pivovarov et al., 2015). Like SS3M, MC3M learns mulitple phenotypes jointly from multi-source clinical data. We implement a collapsed Gibbs sampler for MC3M, and run inference on note, lab and medication data for the full training cohort.

We evaluate the quality of phenotypes learned with each model along three axis: *coherence*, *granularity*, and *label quality*. These axis and the methods for their evaluation are detailed in Pivovarov et al..

Coherence A coherent phenotype is defined as a phenotype containing observations typical of a single disease while omitting observations atypical of said disease. The clinical expert was asked to rate the coherence of individual phenotypes using a five-point Likert scale, with 1 and 5 signifying low and and high coherence, respectively.

Granularity Phenotype granularity is defined in terms of three categories: *(1)* non-disease, *(2)* mixture of diseases, *(3)* single disease. We asked our expert to assign each phenotype to one of these categories.

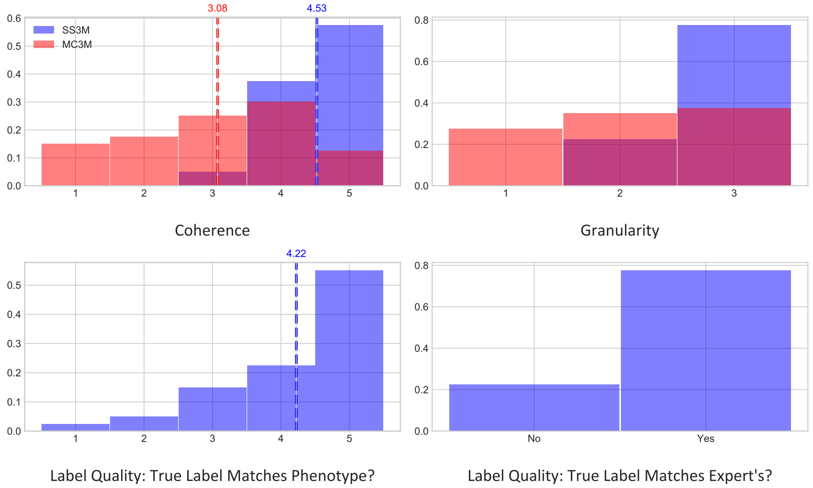


Figure 3: Qualitative evaluation results. Evaluator responses are aggregated within each evaluation type. Shown are the proportions of each possible response (as defined in Section 4.3.2). Means, where appropriate, are shown with vertical hashed lines. Interrater reliabilities (Cohen's κ): Coherence - 0.28; Granularity - 0.14; True label matches phenotype? - 0.04; True label matches expert's? - 0.50.

Label quality We asked our clinical expert to generate a label for each phenotype. If no such label came to mind, the expert was asked to omit this step. If the phenotype in question was learned using SS3M, the expert was asked if their label was equivalent to the phenotype's true label. In addition, the expert was asked to specify how well the true label matched its learned phenotype using a five-point Likert scale with 1 indicating no match and 5 a perfect match.

The phenotypes for our qualitative evaluations are learned using SS3M and MC3M models with $P = 160$. For SS3M, we use phenotypes learned using a labeled subset containing 75% of the training cohort. Individual phenotypes from each model are visualized as sets of three word clouds, one for each data source (See Figures 4 and 5). Word clouds are generated using the WordCloud Python library (Mueller, 2019).

We collaborate with two clinical experts to carry out our evaluation. Both evaluators are medical doctors who have completed or are near completing residency training in internal medicine.

To set up our evaluation we first randomly mix together the individual visualizations of the 40 semi-supervised SS3M phenotypes and 40 randomly chosen MC3M phenotypes, making sure to anonymize their model of origin. These visualizations are then given separately to each clinical expert along with a set of instructions. Each evaluator is also provided a spreadsheet for recording their evaluations. This spreadsheet specifies the order in which phenotypes are to be evaluated, and, for SS3M phenotypes, contains all the ground truth phenotype labels. Where applicable, we ensure evaluators are not exposed to a phenotype's ground truth label until they have completed its granularity and coherence assessments and suggested their own expert label.

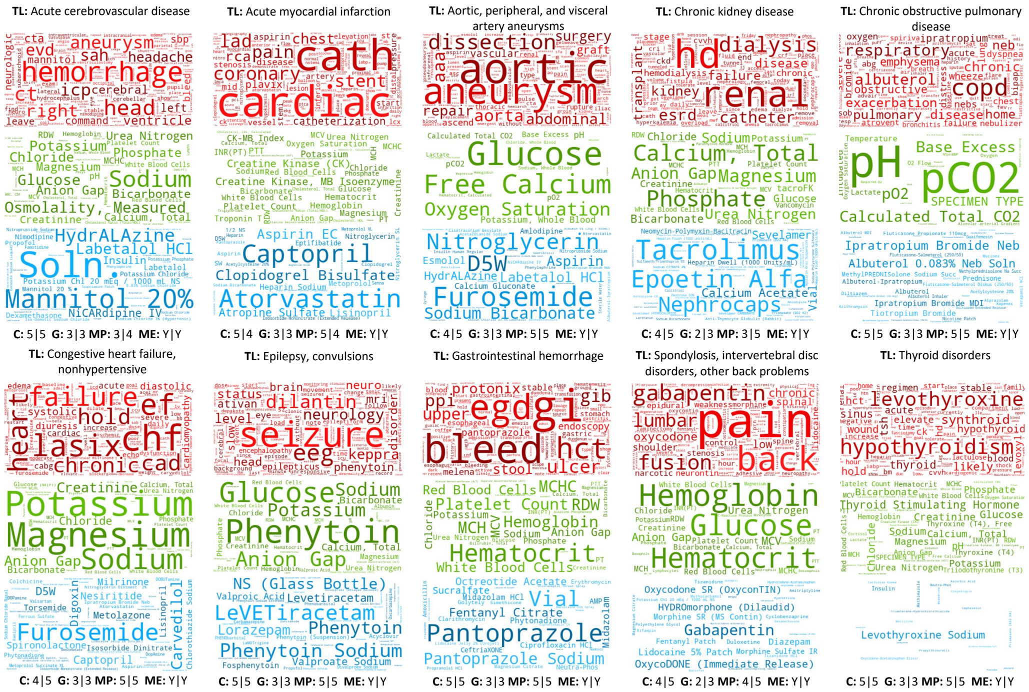
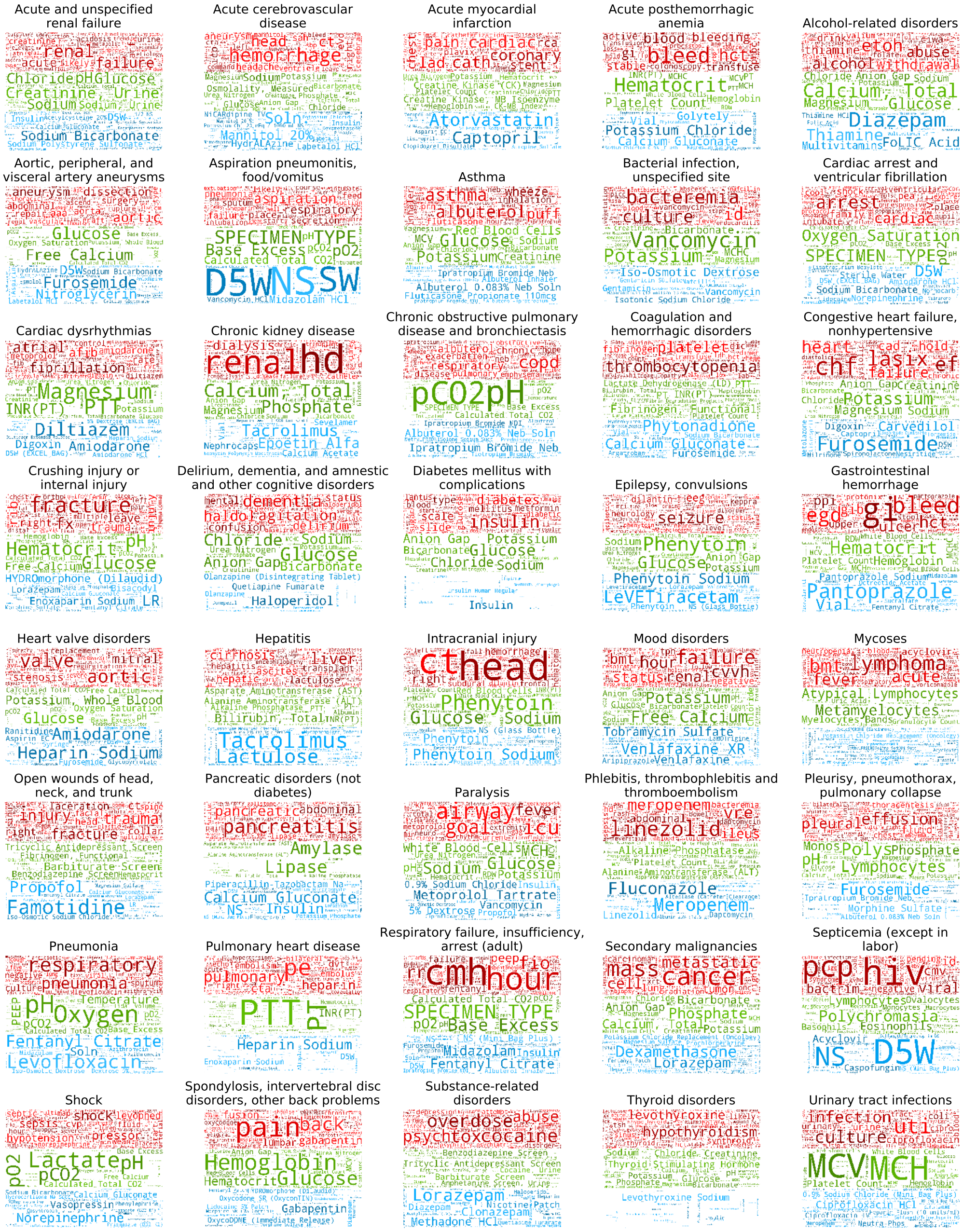


Figure 4: Sample of evaluated SS3M phenotypes. Token size is proportional to token likelihood within a phenotype. Red - words from clinical notes; Green - clinical lab names; Blue - medication names. Evaluations from both clincal experts are presented below each phenotype. **TL** - True label; **C** - Coherence; **G** - Granularity; **MP** - True label matches phenotype?; **ME** - True label matches expert's?

We aggregate evaluations from each of our clinical experts and use Cohen's Kappa to calculate their interrater reliability within each evaluative task.

4.4. Clinical Data Results

Qualitative Evaluation Results Figure 3 summarizes the results of our qualitative evaluation. On average, SS3M outperforms MC3M in terms of phenotype coherence and granularity. Over 90% of SS3M semi-supervised phenotypes showed high coherence (scores of 4 or 5) and nearly 80% were considered to have single-disease granularity. Meanwhile, unsupervised MC3M phenotypes had a more uniform distribution over all levels of coherence and granularity. In terms of label quality, about 75% of SS3M phenotypes were found to match well with their ground truth labels (scores of 4 or 5). Noteably, for nearly 80% of SS3M phenotypes, our expert evaluators were able to suggest a label that matched the ground truth label. This finding suggests that the large majority of SS3M semi-supervised phenotypes communicated the characteristics of the conditions described by their ground truth labels. Over all evaluative tasks, our expert evaluators demonstrate a fair degree of interrater reliability. Figure 4 displays a sample of phenotypes which received strong qualitative evaluations from both expert reviewers. Figure 5 shows the full set of semi-supervised phenotypes employed in the qualitative evaluation.



Acute and unspecified renal failure

Acute cerebrovascular disease

Acute myocardial infarction

Acute posthemorrhagic anemia

Alcohol-related disorders

Aortic, peripheral, and visceral artery aneurysms

Aspiration pneumonitis, food/vomitus

Asthma

Bacterial infection, unspecified site

Cardiac arrest and ventricular fibrillation

Cardiac dysrhythmias

Chronic kidney disease

Chronic obstructive pulmonary disease and bronchiectasis

Coagulation and hemorrhagic disorders

Congestive heart failure, nonhypertensive

Crushing injury or internal injury

Delirium, dementia, and amnestic and other cognitive disorders

Diabetes mellitus with complications

Epilepsy, convulsions

Gastrointestinal hemorrhage

Heart valve disorders

Hepatitis

Intracranial injury

Mood disorders

Mycoses

Open wounds of head, neck, and trunk

Pancreatic disorders (not diabetes)

Paralysis

Phlebitis, thrombophlebitis and thromboembolism

Pleurisy, pneumothorax, pulmonary collapse

Pneumonia

Pulmonary heart disease

Respiratory failure, insufficiency, arrest (adult)

Secondary malignancies

Septicemia (except in labor)

Shock

Spondylosis, intervertebral disc disorders, other back problems

Substance-related disorders

Thyroid disorders

Urinary tract infections

Figure 5: SS3M semi-supervised phenotypes. Phenotypes learned with P=160, and 75% labels retained for training. Token size is proportional to token likelihood within a phenotype. Red - words from clinical notes; Green - clinical lab names; Blue - medication names.

Quantitative Evaluation Results Table 2 summarizes the results of our quantitative evaluation. In general, SS3M's phenotype prediction performance, as measured by macro and micro averaged AUC-ROC and AUC-PR, grows for all values of P as the percentage of labeled patients increases from 1% to 100% of the training cohort. Moreover, performance appears to increase as P increases, particularly for larger amounts of labeled training data.

SS3M demonstrated competitive predictive performance relative to our baselines. In nearly all cases, SS3M with $P \geq 80$ outperforms our multilabel classification baselines (RF and KNN) once 25% of total labels are made available for training. In all cases, the set of 40 one-versus-rest L1-regularized logistic regression (LR) models outperformed all competitors. However, SS3M was the only multilabel classifier that approached LR's performance in at least a subset of cases (e.g. micro averaged AUC-ROC for $P = 160$ and 100% training labels).

Table 3 illustrates SS3M's per-label predictive performance for various proportions of labeled training data. As with the averaged predictive performance, per-label predictive performance tends to increase as more labels are made available for training. However, this trend is not entirely consistent. For some labels, performance increases for a time with the percentage of training labels, but then suddenly suffers a steep drop, possibly followed by a similarly steep rise. This volatility may be due in part to SS3M's inference algorithm getting caught in similar but distinct posterior modes.

		AUC-ROC (Training label %)						AUC-PR (Training label %)					
Average	Model	1%	5%	25%	50%	75%	100%	1%	5%	25%	50%	75%	100%
Macro	SS3M (P=40)	0.557	0.653	0.723	0.73	0.723	0.737	0.156	0.22	0.29	0.305	0.286	0.302
	SS3M (P=80)	0.48	0.622	0.717	0.766	0.787	0.802	0.117	0.226	0.313	0.381	0.389	0.401
	SS3M (P=160)	0.445	0.54	0.702	0.781	0.798	0.813	0.0955	0.162	0.331	0.412	0.444	0.464
	RF (ML)	0.643	0.687	0.721	0.734	0.744	0.75	0.171	0.206	0.246	0.269	0.281	0.291
	KNN (ML)	0.605	0.641	0.679	0.695	0.701	0.704	0.15	0.184	0.228	0.246	0.256	0.263
	LR (OVR)	0.711	0.812	0.843	0.844	0.846	0.846	0.336	0.471	0.526	0.531	0.53	0.53
Micro	SS3M (P=40)	0.627	0.676	0.76	0.783	0.787	0.804	0.143	0.188	0.233	0.24	0.266	0.304
	SS3M (P=80)	0.629	0.699	0.786	0.837	0.847	0.858	0.18	0.241	0.329	0.431	0.442	0.465
	SS3M (P=160)	0.621	0.658	0.787	0.841	0.858	0.866	0.157	0.187	0.341	0.441	0.478	0.521
	RF (ML)	0.716	0.751	0.779	0.788	0.796	0.8	0.239	0.296	0.345	0.369	0.38	0.389
	KNN (ML)	0.657	0.693	0.725	0.74	0.743	0.746	0.193	0.239	0.288	0.309	0.319	0.325
	LR (OVR)	0.766	0.842	0.864	0.865	0.867	0.867	0.407	0.533	0.576	0.578	0.576	0.572

Table 2: Quantitative evaluation summary. Macro and micro averages are calculated for each model over all label targets. ML - multilabel classifier; OVR - one-versus-rest classifier.

	Prevalence		AUC-ROC (Training label %)						AUC-PR (Training label %)					
Label	Train (full)	Test	1%	5%	25%	50%	75%	100%	1%	5%	25%	50%	75%	100%
Acute and unspecified renal failure	0.205	0.207	0.505	0.483	0.808	0.569	0.841	0.856	0.242	0.246	0.64	0.362	0.681	0.671
Acute cerebrovascular disease	0.0841	0.0859	0.474	0.567	0.928	0.941	0.943	0.939	0.0834	0.123	0.599	0.711	0.714	0.744
Acute myocardial infarction	0.117	0.119	0.555	0.53	0.859	0.846	0.87	0.907	0.141	0.127	0.642	0.587	0.656	0.701
Acute posthemorrhagic anemia	0.0869	0.0939	0.616	0.484	0.499	0.729	0.746	0.766	0.141	0.108	0.107	0.311	0.367	0.379
Alcohol-related disorders	0.0866	0.0892	0.445	0.836	0.884	0.902	0.899	0.902	0.0811	0.563	0.636	0.652	0.68	0.665
Aortic, peripheral, and visceral artery aneurysms	0.0441	0.0419	0.502	0.434	0.888	0.907	0.906	0.871	0.0427	0.0344	0.52	0.459	0.545	0.5
Aspiration pneumonitis, food/vomitus	0.0706	0.0687	0.458	0.81	0.811	0.82	0.729	0.45	0.0639	0.268	0.39	0.271	0.219	0.0722
Asthma	0.0634	0.0659	0.478	0.475	0.906	0.891	0.894	0.898	0.0624	0.0634	0.383	0.349	0.371	0.354
Bacterial infection, unspecified site	0.0863	0.09	0.346	0.446	0.406	0.689	0.631	0.505	0.0684	0.0998	0.0822	0.254	0.249	0.142
Cardiac arrest and ventricular fibrillation	0.0339	0.0362	0.456	0.545	0.666	0.916	0.908	0.909	0.0389	0.055	0.104	0.396	0.323	0.298
Cardiac dysrhythmias	0.321	0.32	0.605	0.826	0.541	0.802	0.853	0.86	0.486	0.785	0.401	0.763	0.821	0.825
Chronic kidney disease	0.104	0.103	0.389	0.649	0.652	0.721	0.747	0.663	0.084	0.302	0.303	0.36	0.391	0.329
Chronic obstructive pulmonary disease and bronchiectasis	0.117	0.116	0.5	0.445	0.861	0.857	0.858	0.816	0.122	0.104	0.511	0.509	0.523	0.444
Coagulation and hemorrhagic disorders	0.109	0.102	0.332	0.671	0.447	0.732	0.756	0.757	0.0739	0.271	0.106	0.368	0.39	0.384
Congestive heart failure, nonhypertensive	0.241	0.233	0.545	0.45	0.832	0.79	0.805	0.85	0.328	0.214	0.727	0.655	0.688	0.745
Crushing injury or internal injury	0.0409	0.0424	0.513	0.604	0.903	0.873	0.877	0.895	0.0475	0.0633	0.456	0.457	0.481	0.538
Delirium, dementia, and amnestic and other cognitive disorders	0.0703	0.0687	0.394	0.419	0.83	0.855	0.871	0.874	0.0538	0.0575	0.448	0.434	0.369	0.419
Diabetes mellitus with complications	0.0757	0.08	0.444	0.401	0.932	0.942	0.935	0.918	0.068	0.0614	0.567	0.597	0.599	0.619
Epilepsy, convulsions	0.0643	0.063	0.448	0.917	0.933	0.92	0.931	0.919	0.0557	0.556	0.575	0.582	0.645	0.59
Gastrointestinal hemorrhage	0.0686	0.0732	0.407	0.472	0.921	0.914	0.902	0.897	0.0583	0.0775	0.579	0.559	0.612	0.611
Heart valve disorders	0.151	0.153	0.528	0.614	0.543	0.678	0.833	0.834	0.159	0.237	0.195	0.33	0.618	0.668
Hepatitis	0.0467	0.0482	0.417	0.746	0.858	0.794	0.842	0.801	0.0418	0.289	0.316	0.297	0.338	0.335
Intracranial injury	0.0633	0.0601	0.572	0.766	0.784	0.931	0.938	0.936	0.0811	0.185	0.257	0.552	0.595	0.565
Mood disorders	0.101	0.106	0.424	0.433	0.694	0.788	0.432	0.857	0.0896	0.0955	0.298	0.418	0.0972	0.486
Mycoses	0.0327	0.0358	0.301	0.425	0.552	0.481	0.572	0.559	0.0254	0.0378	0.073	0.0533	0.0906	0.0905
Open wounds of head, neck, and trunk	0.0283	0.0246	0.522	0.605	0.569	0.912	0.919	0.916	0.0294	0.0381	0.0323	0.236	0.252	0.235
Pancreatic disorders (not diabetes)	0.0283	0.0283	0.308	0.464	0.69	0.945	0.961	0.952	0.0192	0.032	0.216	0.416	0.502	0.426
Paralysis	0.0248	0.0298	0.39	0.44	0.533	0.555	0.655	0.585	0.0232	0.0267	0.0423	0.06	0.117	0.0756
Phlebitis, thrombophlebitis and thromboembolism	0.0584	0.0585	0.391	0.385	0.45	0.468	0.478	0.804	0.0494	0.0508	0.0581	0.0676	0.0727	0.35
Pleurisy, pneumothorax, pulmonary collapse	0.0949	0.101	0.432	0.495	0.622	0.69	0.679	0.616	0.0966	0.12	0.285	0.371	0.35	0.278
Pneumonia	0.133	0.142	0.348	0.48	0.754	0.778	0.791	0.81	0.111	0.171	0.498	0.509	0.506	0.541
Pulmonary heart disease	0.0635	0.0629	0.415	0.439	0.669	0.618	0.691	0.676	0.0519	0.0569	0.186	0.221	0.294	0.262
Respiratory failure, insufficiency, arrest (adult)	0.219	0.211	0.369	0.466	0.733	0.773	0.587	0.792	0.179	0.246	0.496	0.547	0.421	0.642
Secondary malignancies	0.0611	0.0611	0.463	0.482	0.958	0.959	0.959	0.958	0.0617	0.0616	0.617	0.647	0.625	0.615
Septicemia (except in labor)	0.136	0.133	0.362	0.549	0.534	0.705	0.506	0.778	0.111	0.216	0.204	0.422	0.183	0.52
Shock	0.0801	0.0791	0.406	0.583	0.531	0.852	0.876	0.859	0.0783	0.134	0.15	0.447	0.498	0.458
Spondylosis, intervertebral disc disorders, other back problems	0.0448	0.0468	0.439	0.478	0.721	0.748	0.761	0.804	0.0432	0.0505	0.211	0.226	0.228	0.254
Substance-related disorders	0.0419	0.0409	0.455	0.441	0.466	0.609	0.848	0.792	0.0375	0.0376	0.0383	0.101	0.288	0.3
Thyroid disorders	0.103	0.105	0.431	0.497	0.47	0.952	0.947	0.953	0.0895	0.11	0.105	0.664	0.691	0.68
Urinary tract infections	0.125	0.123	0.472	0.436	0.483	0.435	0.833	0.843	0.129	0.114	0.134	0.116	0.558	0.544
Macro Average			0.445	0.54	0.702	0.781	0.798	0.813	0.0955	0.162	0.331	0.412	0.444	0.464
Micro Average			0.622	0.658	0.788	0.842	0.858	0.866	0.158	0.188	0.341	0.44	0.476	0.518

Table 3: SS3M per-label predictive performance. Shown are results for $P = 160$, and 75% labels retained for training.

5. Conclusion

SS3M is a model for semi-supervised learning of disease phenotypes from clinical data. We exposed SS3M to data for a simulated cohort of patients as well as data for a cohort of patients from the MIMIC-III clinical database. Our simulated results demonstrate the effectiveness of the semi-supervised mechanism we built into the model. With only a small set of labels retained for training, SS3M is able to fully recover the ground truth phentoypes used to generate our dataset.

Encouraged by these results, we applied SS3M to clinical data using HCUP CCS ICD9 code groups as our label set. Here again our semi-supervised mechanism demonstrated its utility. Our clinical expert evaluators judged that a signifcant proportion of the phenotypes inferred by our model did indeed recover the clinical characteristics of their associated disease labels. Furthermore, relative to several commonly used baselines, SS3M showed competetive performance in phenotype label prediction on a held-out patient cohort.

The labels we employ in our experiments with clinical data were derived from readily available ICD9 diagnosis codes. Though we reduce the noisiness of our labels by utilizing HCUP CCS code groups, we still lack a true gold-standard to train and test our model against. This limitation complicates our evaluation of the model particularly in comparison to our baselines. Specifically, because we do not completely trust that our labels accurately represent the disease status of the patients in our cohort, there is some doubt regarding the accuracy of the predictive performance assessments of all our models.

In future work we will obtaina small set of expert-generated, gold-standard disease labels for use in training and testing our semi-supervised model and our baselines. Using a true gold-standard will give us more confidence in the performance of our models, and will allow us to carry out error analysis to better identify cases for which SS3M struggles to recover a patient's ground truth disease status.

Acknowledgments

Many thanks to Ruijun Chen, MD and Allen Do Won Seol, MD for their assistance in evaluating model outputs.
Research reported in this publication was supported by The National Library of Medicine of the National Institutes of Health under award number F31LM012894

References

Savannah L Bergquist, Gabriel A Brooks, Nancy L Keating, Mary Beth Landrum, and Sherri Rose. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proceedings of machine learning research*, 68:25, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Santiago Esteban, Manuel Rodríguez Tablado, Francisco E Peper, Yamila S Mahumud, Ricardo I Ricci, Karin S Kopitowski, and Sergio A Terrasa. Development and validation of various phenotyping algorithms for diabetes mellitus using data from electronic health records. *Computer methods and programs in biomedicine*, 152:53–70, 2017.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, 52:199–211, 2014a.

Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124. ACM, 2014b.

George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Shalmali Joshi, Suriya Gunasekar, David Sontag, and Joydeep Ghosh. Identifiable phenotyping using constrained non-negative matrix factorization. *arXiv preprint arXiv:1608.00704*, 2016.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

Andreas Mueller. WordCloud. https://github.com/amueller/word_cloud, 2019.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58:156–165, 2015.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

Rachel L Richesson, W Ed Hammond, Meredith Nahm, Douglas Wixted, Gregory E Simon, Jennifer G Robinson, Alan E Bauck, Denise Cifelli, Michelle M Smerek, John Dickerson, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the nih health care systems collaboratory. *Journal of the American Medical Informatics Association*, 20(e2):e226–e231, 2013.

Rachel L Richesson, Jimeng Sun, Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial intelligence in medicine*, 71:57–61, 2016.

Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274. ACM, 2015.

Appendix A.

Here we provide additional details regarding the derivation of SS3M's collapsed Gibbs sampler. The implementation is available at https://github.com/victorarodri/SS3M.

Collapsed Joint Distribution

The joint distribution for SS3M is

$$\begin{aligned} p(\boldsymbol{A}, \boldsymbol{B}, &B^*, \boldsymbol{C}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta, \beta^*, \gamma) = p(B^*; \beta^*) \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \\ &\times p(\boldsymbol{C}_p; \alpha) \prod_{d=1}^{D} p(\boldsymbol{\Theta}_d | \boldsymbol{A}_{d:}, \boldsymbol{B}, B^*) p(\boldsymbol{A}_{dp} | \boldsymbol{C}_p) \prod_{s=1}^{S} p(\boldsymbol{\Phi}_{sp}; \gamma_s) \prod_{n=1}^{N_{sd}} p(\boldsymbol{W}_{sdn} | \boldsymbol{Z}_{sdn}, \boldsymbol{\Phi}_{s:}) p(\boldsymbol{Z}_{sdn} | \boldsymbol{\Theta}_d). \end{aligned} \tag{2}$$

The distribution for each factor on the RHS is given in the generative process described in Algorithm 1.

We integrate $\boldsymbol{C}$,$\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ out of SS3M's joint distribution to obtain the collapsed joint:

$$
\begin{aligned}
p(\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta, \beta^*, \gamma) = p(B^*; \beta^*) & \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \\
& \times \prod_{s=1}^{S} \prod_{d=1}^{D} \prod_{n=1}^{N_{sd}} \int_{\boldsymbol{C}_p} p(\boldsymbol{A}_{dp} | \boldsymbol{C}_p) p(\boldsymbol{C}_p; \alpha) d\boldsymbol{C}_p \qquad (3) \\
& \times \int_{\boldsymbol{\Theta}_d} p(\boldsymbol{Z}_{sdn} | \boldsymbol{\Theta}_d) p(\boldsymbol{\Theta}_d | \boldsymbol{A}_{d:}, \boldsymbol{B}, B^*) d\boldsymbol{\Theta}_d \\
& \times \int_{\boldsymbol{\Phi}_{sp}} p(\boldsymbol{\Phi}_{sp}; \gamma_s) p(\boldsymbol{W}_{sdn} | \boldsymbol{Z}_{sdn}, \boldsymbol{\Phi}_{s:}) d\boldsymbol{\Phi}_{sp} \\
= p(B^*; \beta^*) & \prod_{p=1}^{P} p(\boldsymbol{B}_p; \beta) \\
& \times \prod_{d=1}^{D} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \sum_d \boldsymbol{A}_{dp}) \Gamma(\alpha_2 + D - \sum_d \boldsymbol{A}_{dp})}{\Gamma(\alpha_1 + \alpha_2 + D)} \\
& \qquad (4) \\
& \times \frac{\Gamma(\sum_p \boldsymbol{r}_{dp})}{\prod_p \Gamma(\boldsymbol{r}_{dp})} \frac{\prod_p \Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_p \boldsymbol{r}_{dp} + n_{dp})} \\
& \times \prod_{s=1}^{S} \frac{\Gamma(\sum_v \gamma_{sv})}{\prod_v \Gamma(\gamma_{sv})} \frac{\prod_v \Gamma(\gamma_{sv} + n_{spv})}{\Gamma(\sum_v \gamma_{sv} + n_{spv})},
\end{aligned}
$$

where $\Gamma(\cdot)$ indicates the Gamma function, $\boldsymbol{r}_{dp} = \boldsymbol{A}_{dp} \odot \boldsymbol{B}_p + (\mathbb{1} - \boldsymbol{A}_{dp}) B^*$, n_{dp} is the number of patient d's observations assigned to phenotype p, and n_{spv} is the number of times token v from data source s has been assigned to phenotype p.

Complete Conditional Distributions

Here we obtain proportionalities for the complete conditional distributions of each latent variable in our collapsed joint. Note we use "$-$" to indicate all variables in the joint *excluding* that which appears on the left side of the conditioning bar.

$$
p(\boldsymbol{A}_{dp} | -) \propto \Gamma(\alpha_1 + \sum_{d'} \boldsymbol{A}_{d'p}) \Gamma(\alpha_1 + D - \sum_{d'} \boldsymbol{A}_{d'p}) \frac{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'})}{\Gamma(\boldsymbol{r}_{dp})} \frac{\Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'} + n_{dp'})} \qquad (5)
$$

$$
p(\boldsymbol{B}_p | -) \propto \text{Gamma}(\boldsymbol{B}_p; \beta) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'})}{\Gamma(\boldsymbol{r}_{dp})} \frac{\Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \boldsymbol{r}_{dp'} + n_{dp'})} \qquad (6)
$$

$$
p(B^* | -) \propto \text{Gamma}(B^*; \beta^*) \prod_{d=1}^{D} \frac{\Gamma(\sum_p \boldsymbol{r}_{dp})}{\prod_p \Gamma(\boldsymbol{r}_{dp})} \frac{\prod_p \Gamma(\boldsymbol{r}_{dp} + n_{dp})}{\Gamma(\sum_p \boldsymbol{r}_{dp} + n_{dp})} \qquad (7)
$$

$$p(\boldsymbol{Z}_{sdn}|-) \propto (\boldsymbol{r}_{dp} + n_{dp}^{-sdn}) \frac{\gamma_{sv} + n_{spv}^{-sdn}}{\sum_{v'} \gamma_{sv'} + n_{spv'}^{-sdn}} \tag{8}$$

In the proportionality for $p(\boldsymbol{Z}_{sdn}|-)$, the $n_{\cdot}^{-sdn}$ terms indicate total token assignment counts excluding the current assignment, $\boldsymbol{Z}_{sdn}$. The index v refers to the observed value of $\boldsymbol{W}_{sdn}$.

The proportionalities for $p(\boldsymbol{A}_{dp}|-)$ and $p(\boldsymbol{Z}_{sdn}|-)$ are simple to normalize, and can be sampled from directly afterward. This is not the case for $p(\boldsymbol{B}_p|-)$ and $p(B^*|-)$, which we sample from using Hamiltonian Monte Carlo (HMC).

Hamiltonian Monte Carlo

To use HMC must calculate a potential energy function proportional to our target distribution and calculate its gradient with respect to the corresponding random variable. Note that the $\boldsymbol{B}$ and B^* are constrained to $\mathbb{R}^+$. We remove this constraint by applying a change of variables to sample in log space.

$$p(\hat{\boldsymbol{B}}_p|-) \propto \exp(\hat{\boldsymbol{B}}_p \beta_1 - \exp(\hat{\boldsymbol{B}}_p)/\beta_2) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'})}{\Gamma(\hat{\boldsymbol{r}}_{dp})} \frac{\Gamma(\hat{\boldsymbol{r}}_{dp} + n_{dp})}{\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'} + n_{dp'})} \tag{9}$$

$$p(\hat{B}^*|-) \propto \exp(\hat{B}^* \beta_1^* - \exp(\hat{B}^*)/\beta_1^*) \prod_{d=1}^{D} \frac{\Gamma(\sum_{p} \hat{\boldsymbol{r}}_{dp}^*)}{\prod_p \Gamma(\hat{\boldsymbol{r}}_{dp}^*)} \frac{\prod_p \Gamma(\hat{\boldsymbol{r}}_{dp}^* + n_{dp})}{\Gamma(\sum_p \hat{\boldsymbol{r}}_{dp}^* + n_{dp})}, \tag{10}$$

where $\hat{\boldsymbol{B}}_p = \log \boldsymbol{B}_p$, $\hat{B}^* = \log B^*$, $\hat{\boldsymbol{r}}_{dp} = \boldsymbol{A}_{dp} \odot \exp(\hat{\boldsymbol{B}}_p) + (\mathbb{1} - \boldsymbol{A}_{dp})B^*$, and $\hat{\boldsymbol{r}}_{dp}^* = \boldsymbol{A}_{dp} \odot \boldsymbol{B}_p + (\mathbb{1} - \boldsymbol{A}_{dp}) \exp(\hat{B}^*)$.

The potentials we require are obtained by taking the negative log of our transformed target distributions.

$$\begin{aligned} U(\hat{\boldsymbol{B}}_p) &= -\log p(\hat{\boldsymbol{B}}_p|-) \qquad (11)\\ &\propto \exp(\hat{\boldsymbol{B}}_p)/\beta_2 - \hat{\boldsymbol{B}}_p \beta_1 \qquad (12)\\ &\quad - \sum_{d=1}^{D} \log\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'}) - \log\Gamma(\hat{\boldsymbol{r}}_{dp}) + \log\Gamma(\hat{\boldsymbol{r}}_{dp} + n_{dp}) - \log\Gamma(\sum_{p'} \hat{\boldsymbol{r}}_{dp'} + n_{dp'}) \end{aligned}$$

$$\begin{aligned} U(\hat{B}^*) &= -\log p(\hat{B}^*|-) \qquad (13)\\ &\propto \exp(\hat{B}^*)/\beta_1^* - \hat{B}^* \beta_1^* \qquad (14)\\ &\quad - \sum_{d=1}^{D} \log\Gamma(\sum_{p} \hat{\boldsymbol{r}}_{dp}^*) - \log\Gamma(\sum_{p} \hat{\boldsymbol{r}}_{dp}^* + n_{dp}) + \sum_{d=1}^{D}\sum_{p=1}^{P} \log\Gamma(\hat{\boldsymbol{r}}_{dp}^*) - \log\Gamma(\hat{\boldsymbol{r}}_{dp}^* + n_{dp}) \end{aligned}$$

Their gradients are as follows.

$$\begin{aligned}\frac{\partial U(\hat{\boldsymbol{B}}_p)}{\partial \hat{\boldsymbol{B}}_p} &= -\beta_1 + \exp(\hat{\boldsymbol{B}}_p)[\frac{1}{\beta_2} \\ &+ \sum_{d=1}^{D}\{\Psi(\hat{\boldsymbol{r}}_{dp}) - \Psi(\sum_{p'} \hat{\boldsymbol{r}}_{dp'}) + \Psi(\sum_{p'} \hat{\boldsymbol{r}}_{dp'} + n_{dp'}) - \Psi(\hat{\boldsymbol{r}}_{dp} + n_{dp})\}\boldsymbol{A}_{dp}]\end{aligned} \tag{15}$$

$$\begin{aligned}&\frac{\partial U(\hat{B}^*)}{\partial \hat{B}^*} = -\beta_1^* + \exp(\hat{B}^*)[\frac{1}{\beta_2^*} \\ &+ \sum_{d=1}^{D}\sum_{p=1}^{P}\{\Psi(\hat{\boldsymbol{r}}^*_{dp}) - \Psi(\sum_{p'} \hat{\boldsymbol{r}}^*_{dp'}) + \Psi(\sum_{p'} \hat{\boldsymbol{r}}^*_{dp'} + n_{dp'}) - \Psi(\hat{\boldsymbol{r}}^*_{dp} + n_{dp})\}(1 - \boldsymbol{A}_{dp})],\end{aligned} \tag{16}$$

where $\Psi(\cdot)$ indicates the Digamma function.

As detailed in Neal et al., given a step size, ϵ, and path length, L, these gradients allow us to integrate trajectories in log space to arrive at new candidate states for our random variables. We then evaluate the total energy change using our potential energy functions to decide whether to accept or reject our candidate states.

Collapsed Gibbs Sampler

We now have all the necessary elements to construct a collapsed Gibbs sampler for SS3M. The procedure is described below in Algorithm 2.

Algorithm 2 Collapsed Gibbs Sampler for SS3M

```
Intialize: $\alpha$, $\beta$, $\beta^*$, $\{\gamma_s\}_{s=1}^S$
Sample: $\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}$ from their priors in the complete joint
Load: tokenized observations into $\boldsymbol{W}$ & labels into $\boldsymbol{A}$
for each iteration do
    for each patient $d = 1$ to $D$ do
        for each data source $s = 1$ to $S$ do
            for each observation $n = 1$ to $N_{sd}$ do
                Sample $\boldsymbol{Z}_{sdn} \sim p(\boldsymbol{Z}_{sdn}|-)$
            end
        end
        for each phenotype $p = 1$ to $P$ do
            if $\boldsymbol{A}_{dp}$ does not have a fixed label then
                Sample $\boldsymbol{A}_{dp} \sim p(\boldsymbol{A}_{dp}|-)$
            end
        end
    end
    for each phenotype $p = 1$ to $P$ do
        $\boldsymbol{B}_p \leftarrow \exp(HMC(\hat{\boldsymbol{B}}_p, U(\hat{\boldsymbol{B}}_p), \nabla_{\hat{\boldsymbol{B}}_p} U(\hat{\boldsymbol{B}}_p), \epsilon, L))$
    end
    $B^* \leftarrow \exp(HMC(\hat{B}^*, U(\hat{B}^*), \nabla_{\hat{B}^*} U(\hat{B}^*), \epsilon, L))$
end
Return: Samples of $\boldsymbol{A}, \boldsymbol{B}, B^*, \boldsymbol{Z}$
```

Note, in Algorithm 2 we refer to the gradients in Equations 15 and 16 using the symbol ∇.