Identifiable Phenotyping using Constrained Non–Negative Matrix Factorization

Shalmali Joshi  
Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, TX, USA

Suriya Gunasekar  
Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, TX, USA

David Sontag  
Computer Science  
New York University  
NYC, NY, USA

Joydeep Ghosh  
Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, TX, USA

Abstract

This work proposes a new algorithm for automated and simultaneous phenotyping of multiple co–occurring medical conditions, also referred to as comorbidities, using clinical notes from electronic health records (EHRs). A latent factor estimation technique, non-negative matrix factorization (NMF), is augmented with domain constraints from weak supervision to obtain sparse latent factors that are grounded to a fixed set of chronic conditions. The proposed grounding mechanism ensures a one-to-one identifiable and interpretable mapping between the latent factors and the target comorbidities. Qualitative assessment of the empirical results by clinical experts show that the proposed model learns clinically interpretable phenotypes which are also shown to have competitive performance on 30 day mortality prediction task. The proposed method can be readily adapted to any non-negative EHR data across various healthcare institutions.

1. Introduction

Reliably querying for patients with specific medical conditions across multiple organizations facilitates many large scale healthcare applications such as cohort selection, multi-site clinical trials, epidemiology studies etc. (Richesson et al., 2013, Hripcsak and Albers, 2013, Pathak et al., 2013). However, raw EHR data collected across diverse populations and multiple care-givers can be extremely high dimensional, unstructured, heterogeneous, and noisy. Manually querying such data is a formidable challenge for healthcare professionals.
EHR driven phenotypes are concise representations of medical concepts composed of clinical features, conditions, and other observable traits facilitating accurate querying of individuals from EHRs (NIH Health Care Systems Research Collaboratory, 2014). Efforts like eMerge Network\(^1\), PheKB\(^2\) are well known examples of EHR driven phenotyping. Traditionally used rule–based composing methods for phenotyping require substantial time and expert knowledge and have little scope for exploratory analyses. This motivates automated EHR driven phenotyping using machine learning with limited expert intervention.

We propose a weakly supervised model for jointly phenotyping 30 co–occurring conditions (comorbidities) observed in intensive care unit (ICU) patients. Comorbidities are a set of co-occurring conditions in a patient at the time of admission that are not directly related to the primary diagnosis for hospitalization (Elixhauser et al., 1998). Phenotypes for the 30 comorbidities listed in Table 1 are derived using text-based features from clinical notes in a publicly accessible MIMIC-III EHR database (Saeed et al., 2011). We present a novel constrained non–negative matrix factorization (CNMF) for the EHR matrix that aligns the factors with target comorbidities yielding sparse, interpretable, and identifiable phenotypes.

The following aspects of our model distinguish our work from prior efforts:

1. **Identifiability:** A key shortcoming of standard unsupervised latent factor models such as NMF (Lee and Seung, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for phenotyping is that, the estimated latent factors learnt are interchangeable and unidentifiable as phenotypes for specific conditions of interest. We tackle identifiability by incorporating weak (noisy) but inexpensive supervision as constraints our framework. Specifically, we obtain weak supervision for the target conditions in Table 1 using the Elixhauser Comorbidity Index (ECI) (Elixhauser et al., 1998) computed solely from patient administrative data (without human intervention). We then ground the latent factors to have a one-to-one mapping with conditions of interest by incorporating the comorbidities predicted by ECI as support constraints on the patient loadings along the latent factors.

2. **Simultaneous modeling of comorbidities:** ICU patients studied in this paper are frequently afflicted with multiple co–occurring conditions besides the primary cause for admission. In the proposed NMF model, phenotypes for such co–occurring conditions jointly modeled to capture the resulting correlations.

3. **Interpretability:** For wider applicability of EHR driven phenotyping for advance clinical decision making, it is desirable that these phenotype definitions be clinically interpretable and represented as a concise set of rules. We consider the sparsity in the representations as a proxy for interpretability and explicitly encourage conciseness of phenotypes using tuneable sparsity–inducing soft constraints.

We evaluate the effectiveness of the proposed method towards interpretability, clinical relevance, and prediction performance on EHR data from MIMIC-III. Although we focus on ICU patients using clinical notes, the proposed model and algorithm are general and can be applied on any non-negative EHR data from any population group.

\(^1\) http://emerge.mc.vanderbilt.edu/
\(^2\) http://phekb.org/
2. Data Extraction

The MIMIC-III dataset consists of de-identified EHRs for ~ 38,000 adult ICU patients at the Beth Isreal Deaconess Medical Center, Boston, Massachusetts from 2001–2012. For all ICU stays within each admission, clinical notes including nursing progress reports, physician notes, discharge summaries, ECG, etc. are available. We analyze patients who have stayed in the ICU for at least 48 hours (~ 17000 patients). We derive phenotypes using clinical notes collected within the first 48 hours of patients’ ICU stay to evaluate the quality of phenotypes when limited patient data is available. Further, we evaluate the phenotypes on a 30 day mortality prediction problem. To avoid obvious indicators of mortality and comorbidities, apart from restricting to 48 hour data, we exclude discharge summaries as they explicitly mention patient outcomes (including mortality).

1. Clinically relevant bag-of-words features: Aggregated clinical notes from all sources are represented as a single bag-of-words features. To enhance clinical relevance, we create a custom vocabulary containing clinical terms from two sources (a) the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), and (b) the level-0 terms provided by the Unified Medical Language System (UMLS), consolidated into a standard vocabulary format using Metamorphosys — an application provided by UMLS for custom vocabulary creation.3 To extract clinical terms from the raw text, the notes were tagged for chunking using a conditional random field tagger4. The tags are looked up against the custom vocabulary (generated from Metamorphosys) to obtain the bag-of-words representation. Our final vocabulary has ~3600 clinical terms.

2. Computable weak diagnosis: We incorporate domain constraints from weak supervision to ground the latent factors to have a one-to-one mapping with the conditions of interest. In the model described in Section 3, this is enforced by constraining the non-zero entries on patient loading along the latent factors using a weak diagnosis for comorbidities. The weak diagnoses of target comorbidities in Table 1 are obtained using ECI5, computed solely from patient administrative data without human annotation. We refer to this index as weak diagnoses as it is not a physician’s exact diagnosis and is subject to noise and misspecification. Note that ECI ignores diagnoses code related to the primary diagnoses of admission. Thus, ECI models presence and absence of conditions other than the primary reason for admission (comorbidities). The phenotype candidates from the proposed model can be considered as concise representations of such comorbidities.

---

3. See https://www.nlm.nih.gov/healthit/snomedct/ and https://www.nlm.nih.gov/research/umls/
4. https://taku910.github.io/crfpp/
5. https://git.io/v6e7q
The notation used in the paper are enumerated in Table 2. In summary, for each patient $j \in [N]$, (a) the bag-of-words features from clinical notes is represented as column $x^{(j)}$ of EHR matrix $X \in \mathbb{R}^{d \times N}$, and (b) the list of comorbidities diagnosed using ECI is denoted as $C_j \subseteq [K]$. Let an unknown $W^* \in [0, 1]^{K \times N}$ represent the risk of $N$ patients for $K$ comorbidities of interest; each entry $w^*_{kj}$ lies in the interval $[0, 1]$, with 0 and 1 indicating no-risk and maximum-risk, respectively, of patient $j$ being afflicted with condition $k$. If $C_j^* \subseteq [K]$ denotes an accurate diagnosis for patient $j$, then $w^{*(j)}$ satisfies $\text{supp}(w^{*(j)}) \subseteq C_j^*$.

### Definition 1 (EHR driven phenotype)

**EHR driven phenotypes** for $K$ co–occurring conditions are a set of vectors $\{a^{*(k)} \in \mathbb{R}^d : k \in [K]\}$, such that for a patient $j$ afflicted with conditions $C_j^* \subseteq [K]$, 

$$
\mathbb{E}[x^{(j)}|w^{*(j)}] = \sum_{k \in C_j^*} w^*_{kj} a^{*(k)} + b^*,
$$

(1)

where $b^*$ is a bias representing the feature component observed independent of the $K$ target conditions. $A^* \in \mathbb{R}^{d \times K}$ with $a^{*(k)}$ as columns is referred as the *phenotype factor matrix*.

Note that we explicitly model a feature bias $b^*$ to capture frequently occurring terms that are not discriminative of the target conditions, e.g., temperature, pain, etc.

### Cost Function

The bag-of-words features are represented as counts in the EHR matrix $X$. We consider a factorized approximation of $X$ parametrized by matrices $A \in \mathbb{R}^{d \times K}$, $W \in \mathbb{R}^{K \times N}$ and $b \in \mathbb{R}^d$ as $Y = AW + b\mathbb{1}^\top$, where $\mathbb{1}$ denotes a vector of all ones of appropriate dimension. The approximation error of the estimate is measured using the $I$–divergence defined as follows:

$$
\mathcal{D}(X, Y) = \sum_{ij} y_{ij} - x_{ij} - x_{ij} \log \frac{y_{ij}}{x_{ij}},
$$

(2)

Minimizing the $I$–divergence is equivalent to maximum likelihood estimation under a Poisson distributional assumption on individual entries of the EHR matrix parameterized by $Y = AW + b\mathbb{1}^\top$ (Banerjee et al., 2005).

### Table 2: Notation used in the paper

| Notation | Description |
|----------|-------------|
| $[m]$ for integer $m$ | Set of indices $[m] = \{1, 2, \ldots, m\}$. |
| $\Delta^{d-1}$ | Simplex in dimension $d$, $\Delta^{d-1} = \{x \in \mathbb{R}^d_+ : \sum x_i = 1\}$. |
| $x^{(j)}$ | Column $j$ of a matrix $X$. |
| $\text{supp}(x)$ | Support of a vector $x$, $\text{supp}(x) = \{i : x_i \neq 0\}$. |
| $N, d$ | Number of patients ($\sim 17000$) and features ($\sim 3600$), respectively. |
| $X \in \mathbb{R}^{d \times N}$ | EHR matrix from MIMIC III: Clinically relevant bag-of-words features from notes in first 48 hours of ICU stay for $N$ patients. |
| $k = 1, 2, \ldots, K$ | Indices for $K = 30$ comorbidities in Table 1. |
| $C_j \subseteq [K]$ for $j \in [N]$ | Set of comorbidities patient $j$ is diagnosed with using ECI. |
| $W \in [0, 1]^{K \times N}$ | Estimate of patients’ risk for the $K$ conditions. |
| $A \in \mathbb{R}^{d \times K}$, $b \in \mathbb{R}^d$ | Estimate of phenotype factor matrix and feature bias vector. |

3. Identifiable High–Throughput Phenotyping

The notation used in the paper are enumerated in Table 2. In summary, for each patient $j \in [N]$, (a) the bag-of-words features from clinical notes is represented as column $x^{(j)}$ of EHR matrix $X \in \mathbb{R}^{d \times N}$, and (b) the list of comorbidities diagnosed using ECI is denoted as $C_j \subseteq [K]$. Let an unknown $W^* \in [0, 1]^{K \times N}$ represent the risk of $N$ patients for $K$ comorbidities of interest; each entry $w^*_{kj}$ lies in the interval $[0, 1]$, with 0 and 1 indicating no-risk and maximum-risk, respectively, of patient $j$ being afflicted with condition $k$. If $C_j^* \subseteq [K]$ denotes an accurate diagnosis for patient $j$, then $w^{*(j)}$ satisfies $\text{supp}(w^{*(j)}) \subseteq C_j^*$. 

**Definition 1 (EHR driven phenotype)** EHR driven phenotypes for $K$ co–occurring conditions are a set of vectors $\{a^{*(k)} \in \mathbb{R}^d : k \in [K]\}$, such that for a patient $j$ afflicted with conditions $C_j^* \subseteq [K]$, 

$$
\mathbb{E}[x^{(j)}|w^{*(j)}] = \sum_{k \in C_j^*} w^*_{kj} a^{*(k)} + b^*,
$$

(1)

where $b^*$ is a bias representing the feature component observed independent of the $K$ target conditions. $A^* \in \mathbb{R}^{d \times K}$ with $a^{*(k)}$ as columns is referred as the phenotype factor matrix.

Note that we explicitly model a feature bias $b^*$ to capture frequently occurring terms that are not discriminative of the target conditions, e.g., temperature, pain, etc.

**Cost Function** The bag-of-words features are represented as counts in the EHR matrix $X$. We consider a factorized approximation of $X$ parametrized by matrices $A \in \mathbb{R}^{d \times K}$, $W \in \mathbb{R}^{K \times N}$ and $b \in \mathbb{R}^d$ as $Y = AW + b\mathbb{1}^\top$, where $\mathbb{1}$ denotes a vector of all ones of appropriate dimension. The approximation error of the estimate is measured using the $I$–divergence defined as follows:

$$
\mathcal{D}(X, Y) = \sum_{ij} y_{ij} - x_{ij} - x_{ij} \log \frac{y_{ij}}{x_{ij}},
$$

(2)

Minimizing the $I$–divergence is equivalent to maximum likelihood estimation under a Poisson distributional assumption on individual entries of the EHR matrix parameterized by $Y = AW + b\mathbb{1}^\top$ (Banerjee et al., 2005).
Algorithm 1: Phenotyping using constrained NMF.
Input: \(X, \{C_j : j \in [N]\}\) and parameter \(\lambda\). Initialization: \(A_{(0)}, b_{(0)}\).

\[
\text{while Not converged do }
\begin{align*}
W_{(t)} & \leftarrow \arg \min_W D(X, A_{(t-1)}W + b_{(t-1)}1^\top) \text{ s.t. } W \in [0,1]^{K \times N}, \supp(w^j) = C_j, \forall j \\
A_{(t)}, b_{(t)} & \leftarrow \arg \min_{A,b \geq 0} D(X, AW_{(t)} + b1^\top) \text{ s.t. } a_{(k)}^j \in \lambda \Delta^{d-1}, \forall k
\end{align*}
\]

Phenotypes For the \(K\) comorbidities, columns of \(A, \{a_{(k)}^j\}_{k \in [K]}\) are proposed as candidate phenotypes derived from the EHR \(X\), i.e. approximations to \(\{a^*_k\}_{k \in [K]}\).

Constraints The following constraints are incorporated in learning \(A\) and \(W\).

1. Support Constraints: The non-negative rank–\(K\) factorization of \(X\) is ‘grounded’ to \(K\) target comorbidities by constraining the support of risk \(w^j\) corresponding to patient \(j\) using weak diagnosis \(C_j\) from ECI as an approximation of the conditions in Definition 1.

2. Sparsity Constraints: Scaled simplex constraints are imposed on the columns of \(A\) with a tuneable parameter \(\lambda > 0\) to encourage sparsity of phenotypes. Restricting the patient loadings matrix as \(W \in [0,1]^{K \times N}\) not only allows to interpret the loadings as the patients’ risk, but also makes simplex constraints effective in a bilinear optimization.

Simultaneous phenotyping of comorbidities using constrained NMF is posed as follows:

\[
\tilde{A}, \tilde{W}, \tilde{b} = \arg\min_{A \geq 0, W \geq 0, b \geq 0} D(X, AW + b1^\top) \text{ s.t. } \supp(w^j) = C_j, \forall j \in [N], W \in [0,1]^{K \times N}, a_{(k)}^j \in \lambda \Delta^{d-1}, \forall j \in [K],
\]

4. Empirical Evaluation

The estimated phenotypes are evaluated on various metrics. We denote the model learned using Algorithm 1 with a given parameter \(\lambda > 0\) as \(\lambda\)-CNMF. The following baselines are used for comparison:

1. Labeled LDA (LLDA): LLDA (Ramage et al., 2009) is the supervised counterpart of LDA, a probabilistic model to estimate topic distribution of a corpus. It assumes that word counts of documents arise from multinomial distributions. It incorporates supervision on topics contained in a document and can be naturally adapted for phenotyping from bag-of-words clinical features, where the topic–word distributions form candidate phenotypes. While LLDA assumes that the topic loadings of a document lie on the probability simplex \(\Delta^{K-1}\), \(\lambda\)-CNMF allows each patient–condition \(w^j\) loading to lie in \([0,1]\). In interpreting the patient loading as a disease risk, the latter allows patients to have varying levels of disease prevalence. Also, LLDA can induce sparsity only indirectly via a hyperparameter \(\beta\) of the informative prior on the topic–word distributions. While this does not guarantee...
sparse estimates, we obtain reasonable sparsity on LLDA estimates. We use the Gibbs sampling code from MALLET (McCallum, 2002) for inference. For a fair comparison to CNMF which uses an extra bias factor, we allow LLDA to model an extra topic shared by all documents in the corpus.

2. **NMF with support constraints (NMF+support):** This NMF model incorporates non-negativity and support constraints from weak supervision but not the sparsity inducing constraints on the phenotype matrix. This allows to study the effect of sparsity inducing constraints for interpretability. On the other hand, imposing sparsity without our grounding technique does not yield identifiable topics and hence is not studied as a baseline.

3. **Multi-label Classification (MLC):** This baseline treats weak supervision (from ECI) as accurate labels in a fully supervised model. A sparsity inducing $\ell_1$ regularized logistic regression classifier is learned for each condition independently. The learned weight vector for each condition $k$ determines importance of clinical terms towards discriminating patients with condition $k$ and are treated as candidate phenotypes for condition $k$.

The weak supervision does not account for the primary diagnosis for admission in the ICU population as the ECI ignores primary diagnoses at admission (Elixhauser et al., 1998). However, the learning algorithm can be easily modified to account for the primary diagnoses, if required by using a modified form of supervision or absorbing the effects in an additional additive term appended to the model. Nevertheless, the proposed model generates highly interpretable phenotypes for comorbidities. Finally, to mitigate the effect of local minima, whenever applicable, for each model, the corresponding algorithm was run with 5 random initializations and results providing the lowest divergence were chosen for comparison.

4.1 **Interpretability–accuracy trade–off**

Sparsity of the latent factors is used as a proxy for interpretability of phenotypes. Sparsity is measured as the median of the number of non-zero entries in columns of the phenotype matrix $A$ (lower is better). The $\lambda$ parameter in $\lambda$–CNMF controls the sparsity by imposing scaled simplex constraints on $A$. CNMF was trained on multiple $\lambda$ in the range of 0.1 to 1. Stronger sparsity-inducing constraints results in worse fit to the cost function. This trade–off is indeed observed in all models (see A for details). For all models, we pick estimates with lowest median sparsity while ensuring that the phenotype candidate for every condition is represented by at least 5 non-zero clinical terms.

4.2 **Clinical relevance of phenotypes**

We requested two clinicians to evaluate the candidate phenotypes based on the top 15 terms learned by each model. The ratings were requested on a scale of 1 (poor) to 4 (excellent). The experts were asked to rate based on whether the terms are relevant towards the corresponding condition and whether the terms are jointly discriminative of the condition. Figure 1 shows the summary of qualitative ratings obtained for all models. For each model, we show two columns (corresponding to two experts). The stacked bars show the histogram of the ratings for the models. Nearly 50% of the phenotypes learned from our model were rated ‘good’ or better by both annotators. In contrast, NMF with support constraints but without sparsity inducing constraints hardly learns clinically relevant phenotypes. The proposed model 0.4–CNMF also received significantly higher number of ‘excellent’ and ‘good’
Figure 1: Qualitative Ratings from Annotation: The two bars represent the ratings provided by the two annotators. Each bar is a histogram of the scores for the 30 comorbidities sorted by scores.

![Figure 1: Qualitative Ratings from Annotation](image)

Table 3: Relative Rankings Matrix: Each row of the table is the number of times the model along the row was rated strictly better than the model along the column by clinical experts, e.g., column 3 in row 2 implies that LLDA was rated better than MLC 12 times over all conditions by all experts.

|         | 0.4–CNMF | LLDA | MLC | NMF |
|---------|----------|------|-----|-----|
| LLDA    | 0        | 28   | 20  | 44  |
| MLC     | 7        | 0    | 12  | 35  |
| NMF+support | 1     | 0    | 1   | 0   |

ratings from both experts. Although LLDA and MLC estimate sparse phenotypes, they are not at par with 0.4–CNMF. Table 3 shows a summary of relative rankings for all models. Each cell entry shows the number of times the model along the corresponding row was rated strictly better than that along the column. 0.4–CNMF is better than all three baselines. The supervised baseline MLC outperforms LLDA even though LLDA learns comorbidities jointly suggesting that the simplex constraint imposed by LLDA may be restrictive.

Figure 2 is an example of a phenotype (top 15 terms) learned by all models for psychoses. For this condition, the proposed model was rated “excellent” and strictly better than both LLDA and MLC by both annotators while LLDA and MLC ratings were tied. However, the phenotype for Hypertension (in Figure 3) learned by 0.4–CNMF has more terms related to ‘Renal Failure’ or ‘End Stage Renal Disease’ rather than hypertension. One of our annotators pointed out that “Candidate 1 is a fairly good description of renal disease, which is an end organ complication of hypertension”, where the anonymized Candidate 1 refers to 0.4–CNMF. Exploratory analysis suggests that hypertension and renal failure are the most commonly co-occurring set of conditions. Over 93% of patients that have hypertension (according to ECI) also suffer from Renal Failure. Thus, our model is unable to distinguish between highly co-occurring conditions. Other baselines were also rated poorly for hypertension, while LLDA was rated only slightly better. More examples of phenotypes are provided in B.
4.3 Mortality prediction

To quantitatively evaluate the utility of the learned phenotypes, we consider the 30 day mortality prediction task. We divide the EHR into 5 cross-validation folds of 80% training and 20% test patients. As this is an imbalanced class problem, the training–test splits are stratified by mortality labels. For each split, all models were applied on the training data to obtain phenotype candidates $\tilde{A}$ and feature biases $\tilde{b}$. For each model, the patient loadings $W$ along the respective phenotype space $(\tilde{A}, \tilde{b})$ are used as features to train a logistic regression classifier for mortality prediction. For CNMF and NMF+support, these are obtained as $W_{\text{train/test}} = \arg\min_{W\in[0,1]^K\times N} D(\tilde{A}W + \tilde{b}^T, X_{\text{train/test}})$ for fixed $(\tilde{A}, \tilde{b})$. For LLDA, these are obtained using Gibbs sampling with fixed topic–word distributions. For MLC, the predicted class probabilities of the comorbidities are used as features. Additionally, we train a logistic regression classifier using the full EHR matrix as features.

We clarify the following points on the methodology: (1) $\tilde{A}$ is learned on the patients in the training dataset only, hence there is no information leak from test patients into training. (2) Test patients’ comorbidities from ECI are not used as support constraints on their loadings. (3) Regularized logistic regression classifiers are used to learn models for mortality prediction. The regularization parameters are chosen via grid-search.

The performance of the above baselines trained on $\ell_2$ regularized logistic regression over a 5-fold cross-validation is reported in Table 4 rows 1–5. The classifier trained on the full EHR unsurprisingly outperforms all baselines as it uses richer high dimensional information. All phenotyping baselines, except NMF+support, show comparable performance on mortality prediction which in spite of learning on a small number of 30 features, is only slightly worse than predictive performance of full EHR with $\sim$3500 features.
Augmented features for mortality prediction (CNMF+Full EHR) Unsurprisingly, Table 4 suggests that the high dimensional EHR data has additional information towards mortality prediction which are lacking in the 30 dimensional features generated via phenotyping. To evaluate whether this additional information can be captured by CNMF if augmented with a small number of raw EHR features, we train a mortality prediction classifier using $\ell_1$ regularized logistic regression on CNMF features/loadings combined with raw bag-of-words features, with parameters tuned to match the performance of the full EHR model. The results are reported in the final row of Table 4.

In exploring the weights learned by the classifier for all features, we observe that only 8.3% of the features corresponding to raw EHR based bag-of-words features have non-zero weights. This suggests that comorbidities capture significant amount of predictive information on mortality and achieve comparable performance to full EHR model with a small number of additional terms. See Figure 35 in Appendix showing the weights learned by the classifier for all features. Figure 4 shows comorbidities and EHR terms with top magnitude weights learned by the CNMF+full EHR classifier. For example, it is interesting to note that the top weighted EHR term – dnr or ‘Do Not Resuscitate’ is not indicative of any comorbidity but is predictive of patient mortality.

5. Discussion and Related Work

Supervised learning methods like Carroll et al. (2011); Kawaler et al. (2012); Chen et al. (2013) or deep learning methods (Lipton et al., 2015; Kale et al., 2015; Henao et al., 2015) for EHR driven phenotyping require expert supervision. Although unsupervised methods like NMF (Anderson et al., 2014) and non-negative tensor factorization (Kolda and Bader, 2009; Harshman, 1970) are inexpensive alternatives (Ho et al., 2014a b c; Luo et al. 2015), they pose challenges with respect to identifiability, interpretability and computation, limiting their scalability.

Most closely related to our paper is work by Halpern et al. (2016b) which is a semi-supervised algorithm for learning the joint distribution on conditions, requiring only that a domain expert specify one or more ‘anchor’ features for each condition (no other labeled data). An ‘anchor’ for a condition is a set of clinical features that when present are highly indicative of the target condition, but whose absence is not a strong label for absence of
Figure 4: Top magnitude weights on (a) EHR and (b) CNMF features in CNMF+Full EHR classifier

the target condition (Halpern et al., 2014, 2016a). For example, the presence of insulin medication is highly indicative of diabetes, but the converse is not true. Joshi et al. (2015) use a similar supervision approach for comorbidities prediction. Whereas the conditions in Halpern et al. (2016b) are binary valued, in our work they are real-valued between 0 and 1. Furthermore, we assume that the support of the conditions is known in the training data.

Our approach achieves identifiability using support constraints to ground the latent factors and interpretability using sparsity constraints. The phenotypes learned are clinically interpretable and predictive of mortality when augmented with a sparse set of raw bag-of-words features on unseen patient population. The model outperforms baselines in terms of clinical relevance according to experts and significantly better than the model which includes supervision but no sparsity constraints. The proposed method can be easily extended to other non-negative data to obtain more comprehensive phenotypes. However, it was observed that the algorithm does not discriminate between frequently co-occurring conditions, e.g. renal failure and hypertension. Further, the weak supervision (using ECI) does not account for the primary diagnoses of admission. Additional model flexibility to account for a primary condition in explaining the observations could potentially improve performance. Addressing the above limitations along with quantitative evaluation of risk for disease prediction, and understanding conditions for uniqueness of phenotyping solutions are interesting areas of follow-up work.
Acknowledgements

We thank Dr. Saul Blecker and Dr. Stephanie Kreml for their qualitative evaluation of the computational phenotypes. SJ, SG and JG were supported by NSF: SCH #1418511. DS was supported by NSF CAREER award #1350965. We also thank Yacine Jernite for sharing a code used in preprocessing clinical notes.

References

A. Anderson, P. K. Douglas, W. T. Kerr, V. S. Haynes, A. L. Yuille, J. Xie, Y. N. Wu, J. A. Brown, and M. S. Cohen. Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. Neuroimage, 2014.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. Journal of Machine Learning Research, 2005.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 2003.

R. J. Carroll, A. E. Eyler, and J. C. Denny. Naive electronic health record phenotype identification for rheumatoid arthritis. In AMIA Annual Symposium, 2011.

Y. Chen, R. J. Carroll, E. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association, 2013.

A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. Medical Care, 1998.

Y. Halpern, Y. Choi, S. Horng, and D. Sontag. Using anchors to estimate clinical state without labeled data. In AMIA Annual Symposium, 2014.

Y. Halpern, S. Horng, Y. Choi, and D. Sontag. Electronic medical record phenotyping using the anchor and learn framework. Journal of the American Medical Informatics Association, 2016a.

Y. Halpern, S. Horng, and D. Sontag. Clinical tagging with joint probabilistic models. Conference on Machine Learning for Health Care, 2016b.

R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis. UCLA Working Papers in Phonetics, 1970.

R. Henao, J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin. Electronic health record analysis via deep poisson factor models. Journal of Machine Learning Research, 2015.

J. C Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. Journal of Biomedical Informatics, 2014a.

J. C. Ho, J. Ghosh, and J. Sun. Extracting phenotypes from patient claim records using nonnegative tensor factorization. In International Conference on Brain Informatics and Health, 2014b.

J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014c.
G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 2013.

S. Joshi, O. Koyejo, and J. Ghosh. Simultaneous prognosis of multiple chronic conditions from heterogeneous EHR data. In *International Conference on Healthcare Informatics*, 2015.

D. C. Kale, Z. Che, M. T. Bahadori, W. Li, Y. Liu, and R. Wetzel. Causal Phenotype Discovery via Deep Networks. In *AMIA Annual Symposium*, 2015.

E. Kawaler, A. Cobian, P. Peissig, D. Cross, S. Yale, and M. Craven. Learning to predict post-hospitalization VTE risk from EHR data. In *AMIA Annual Symposium*, 2012.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 2009.

D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

C. J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 2007.

Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Y. Luo, Y. Xin, E. Hochberg, R. Joshi, O. Uzuner, and P. Szolovits. Subgraph augmented non-negative tensor factorization (santf) for modeling clinical narrative text. *Journal of the American Medical Informatics Association*, 2015.

A. K. McCallum. Mallet: A machine learning for language toolkit, 2002. URL http://mallet.cs.umass.edu

NIH Health Care Systems Research Collaboratory. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials, 2014.

N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 2014.

J. Pathak, A. N. Kho, and J. C. Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 2013.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing*, 2009.

R. L. Richesson, W. E. Hammond, M. Nahm, D. Wixted, G. E. Simon, J. G. Robinson, A. E. Bauck, D. Cifelli, M. M. Smerek, J. 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*, 2013.

M. Saied, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical Care Medicine*, 2011.
Appendix A. Phenotype Sparsity

As suggested in Section 4.1, there is an inherent tradeoff between fit to the cost function and desired sparsity. The trade-off is made explicit for $\lambda$–CNMF in Figure 5. The sparsity of LLDA is controlled by tuning the hyperparameter ($\beta$) of the word-topic multinomial parameters (Blei et al., 2003) and for MLC via the $\ell_1$ regularization parameter $\eta$. A smaller value of $\beta$ ensures that the word-topic probabilities are sparse. As the value of $\beta$ is increased, sparsity decreases (i.e., number of non-zero elements increases). For logistic regression (used by MLC), as the $\ell_1$ regularization parameter increases, sparsity increases. Figure 6a demonstrates the sparsity of the estimated phenotypes for LLDA and Figure 6b shows that of logistic regression. We choose phenotypes obtained at $\beta = 1 \times 10^{-8}$ and $\eta = 100$ for qualitative annotation. The parameters were chosen to achieve the lowest median sparsity while ensuring that for each chronic condition, the corresponding phenotype candidate is represented by at least 5 non-zero clinical terms. Our fourth baseline (NMF + support) did not estimate sparse phenotypes and does not have a tuneable sparsity parameter (but were nevertheless annotated for qualitative evaluation). The proposed model provides the best sparsity among all baselines.

![Figure 5: Sparsity–Accuracy Trade–off. Sparsity of the model is measured as the median of the number of non-zero entries in columns of the phenotype matrix $A$. (a) shows a box plots of the median sparsity across the 30 chronic conditions for varying $\lambda$ values. The median and third–quartile values are explicitly noted on the plots. (b) divergence function value of the estimate from Algorithm 1 plotted against $\lambda$ parameter.](image)

Appendix B. Sample Phenotypes for Baseline Models

Figures 7–33 show the top 15 terms learned for all target chronic conditions for the proposed model and baselines. The sparsity level chosen is based on the criterion described in Section 4.1. For all conditions, the terms are ordered in decreasing order of importance as learned by the models.
Figure 6: Phenotype sparsity for baseline models

(a) LLDA

(b) MLC

Figure 7: Learned Phenotypes for Liver Disease

Figure 8: Learned Phenotypes for Solid Tumor
### Figure 9: Learned Phenotypes for Metastatic Cancer

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| metastatic | metastatic_melanoma | metastatic_prostate_cancer | metastasis |
| metastatic_disease | metastatic_renal_cell_carcinoma | melanoma | mets |
| metastatic colon cancer | lung mass | metastatic disease | metastatic prostate cancer |

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| pain | mass | hypotension | metastatic |
| metastatic | malignant neoplasm | tumor | metastatic melanoma |
| stress ulcer | sob | cancer | metastatic renal cell carcinoma |
| metastatic | disease | metastases | mets |
| metastatic prostate cancer | ovarian ca | lung mass | metastatic renal cell cancer |

### Figure 10: Learned Phenotypes for Chronic Pulmonary Disorder

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| copd | asthma | emphysema | agitated |
| chronic obstructive pulmonary disease | bronchitis | bronchiectasis | asthma |
| emphysema | asbestosis | aspergillosis | alcohol abuse |
| copd exacerbation | obstructive lung disease | pulmonary infarct | alcohol_withdrawal |

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| pain | respiratory failure | asthama | pain |
| copd | pneumonia | sob | edema |
| respiratory failure | emphysema | hyperventilation | chest pain |
| sob | emphysema | sob | pneumonia |
| pneumonia | sob | pneumonia | sob |
| asthma | pneumonia | sob | pneumonia |
| emphysema | sob | pneumonia | sob |
| stress ulcer | stress ulcer | pneumonia | sob |
| sob | pneumonia | sob | pneumonia |
| pneumonia | sob | pneumonia | sob |
| sob | pneumonia | sob | pneumonia |

### Figure 11: Learned Phenotypes for Alcohol Abuse

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| etoh_abuse | alcohol_abuse | alcohol_withdrawal | alcohol_abuse |
| alcoholic cirrhosis | alcoholic hepatitis | withdrawal symptoms | alcohol_abuse |
| alcoholism | delirium tremens | alcoholic hepatitis | alcohol_abuse |
| delirium tremens | alcoholic hepatitis | withdrawal symptoms | alcohol_abuse |
| alcoholic hepatitis | withdrawal | cirrhosis | alcohol_abuse |
| withdrawal | cirrhosis | alcoholic hepatitis | alcohol_abuse |
| cirrhosis | alcoholic hepatitis | fracture | alcohol_abuse |
| alcoholic hepatitis | fracture | altered mental status | alcohol_abuse |
| fracture | altered mental status | htn | alcohol_abuse |

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| pain | edema | pneumonia | alcohol_abuse |
| pneumonia | hemorrhage | agitation | alcohol_abuse |
| alcohol_abuse | stress ulcer | agitated | alcohol_abuse |
| alcohol_abuse | cough | fall | alcohol_abuse |
| alcohol_abuse | fever | stroke | alcohol_abuse |
| alcohol_abuse | seizure | subarachnoid hemorrhage | alcohol_abuse |
| subarachnoid hemorrhage | hematoma | alcohol_abuse | alcohol_abuse |

15
| 0.4-CNMF          | LLDA   | MLC    | NMF+support |
|------------------|--------|--------|-------------|
| **dm**           | **dm** | **dm2**| **niddm**   |
| **diabetes**     | **type_2_diabetes** | **type_ii_diabetes** | **diabetes**   |
| **niddm**        | **type_2_diabetes_mellitus** | **type_ii_diabetes** | **diabetes**   |
| **convulsive_status_epilepticus** | **diabetes_type_2** | **type_2_diabetes_mellitus** | **diabetes**   |
| **diabetes_type_2** | **diabetes_mellitus_type_2** | **skin_ulcers** | **hypoglycemia** |
| **chest_pains**  | **pain** | **dm** | **htn**     |
|                  |        |        | **edema**   |
|                  |        |        | **cad**     |
|                  |        |        | **diaibetes_mellitus** |
|                  |        |        | **chest_pain** |
|                  |        |        | **hypertension** |
|                  |        |        | **dm2**     |
|                  |        |        | **chf**     |
|                  |        |        | **diaibetes** |
|                  |        |        | **hypertension** |
|                  |        |        | **sob**     |
|                  |        |        | **bleeding** |

Figure 12: Learned Phenotypes for Diabetes Uncomplicated

| 0.4-CNMF          | LLDA   | MLC    | NMF+support |
|------------------|--------|--------|-------------|
| **dm**           | **dm** | **htn**| **hypoglycemia** |
| **hypo**         | **glycemia** | **diaibetes_mellitus** | **cad**   |
| **pain**         | **pain** | **dm** | **htn**     |
|                  |        |        | **edema**   |
|                  |        |        | **cad**     |
|                  |        |        | **diaibetes_mellitus** |
|                  |        |        | **chest_pain** |
|                  |        |        | **hypertension** |
|                  |        |        | **dm2**     |
|                  |        |        | **chf**     |
|                  |        |        | **diaibetes** |
|                  |        |        | **hypertension** |
|                  |        |        | **sob**     |
|                  |        |        | **bleeding** |

Figure 13: Learned Phenotypes for Diabetes Complicated

| 0.4-CNMF          | LLDA   | MLC    | NMF+support |
|------------------|--------|--------|-------------|
| **pvd**          | **pain** | **pvd** | **pvd**     |
| **peripheral_vascular_disease** | **pvd** | **edema** | **aaa**   |
| **aaa**          | **aortic** | **aneurysm** | **rupture**   |
|                  | **claudication** | **induration** | **heel_ulcer** |
|                  | **type_a_aortic_dissection** | **leg_ulcer** | **carotid_stenosis** |
|                  | **dural_tear** | **endoleak** | **vascular_disease** |
|                  | **eschar** |        | **pain**     |
|                  |          |        | **pvd**     |
|                  |          |        | **edema**   |
|                  |          |        | **aaa**     |
|                  |          |        | **htn**     |
|                  |          |        | **hematoma** |
|                  |          |        | **nausea**  |
|                  |          |        | **peripheral_vascular_disease** |
|                  |          |        | **ischemia** |
|                  |          |        | **atelectasis** |
|                  |          |        | **coronary_artery_disease** |
|                  |          |        | **stress_ulcer** |
|                  |          |        | **aflb**    |
|                  |          |        | **hypertension** |
|                  |          |        | **pvd**     |
|                  |          |        | **peripheral_vascular_disease** |
|                  |          |        | **pseudoaneurysm** |
|                  |          |        | **aaa**     |
|                  |          |        | **coronary_artery_disease** |
|                  |          |        | **carotid_stenosis** |
|                  |          |        | **aortic_dissection** |
|                  |          |        | **ptx**     |
|                  |          |        | **cardiomegaly** |
|                  |          |        | **aortic_aeurysm** |
|                  |          |        | **renal_artery_stenosis** |
|                  |          |        | **mesenteric_ischemia** |
|                  |          |        | **complaints** |
|                  |          |        | **vegetation** |
|                  |          |        | **calcifications** |

Figure 14: Learned Phenotypes for Peripheral Vascular Disorder
| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| ESRD | Hypotension | CRI | Pain |
| Chronic Kidney Disease | ESRD | AVF | CP |
| Chronic Renal Failure | Renal Failure | ESRD | Nausea |
| CKD | Seizure Disorder | Acute Renal Failure | Chronic Pain |
| End Stage Renal Disease | Fever | Cardiomegaly | Hypertension |
| Acute on Chronic Renal Failure | Unresponsive | Atelectasis | Edema |
| Thrill | Stress Ulcer | Pneumothorax | Cough |
| Atrophic Kidneys | Infection | Seizures | Nausea |
| CRF | Hypertension | Stroke | Hemorrhage |
| Pulmonary Artery Hypertension | Hypotension | Acute Pain | Stress Ulcer |
| Diverticular Disease | Status Epilepticus | Intracranial Hemorrhage | Shock |
| Non Reactive | Mental Status | Shortness of Breath | Nausea |

Figure 15: Learned Phenotypes for Renal Failure

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| Seizure | Seizure Disorder | Restless Leg Syndrome | Seizure |
| Seizure Disorder | Seizures | Seizures | Seizures |
| Status Epilepticus | Unresponsive | Mental Status | Seizures |
| Mental Retardation | Febrile | Dementia | Seizures |
| Seizures | Stress Ulcer | Dementia | Seizures |
| Restless Leg Syndrome | Stress Ulcer | Hemorrhax | Seizures |
| Epilepsy | Infection | Retropulsion | Seizures |
| Multiple Sclerosis | Pneumonia | Multiple Sclerosis | Seizures |
| Tonic Clonic Seizure | Hypertension | Epilepsy | Seizures |
| CNS Infection | Agitated | Hydrocephalus | Seizures |
| Trigeminal Neuralgia | Status Epilepticus | Lethargic | Seizures |
| Parkinsons Disease | Seizure Disorder | Hypoxemia | Seizures |
| Grand Mal Seizure | Mental Status | Overdose | Seizures |
| Generalized Seizure | Dementia | Shortness of Breath | Seizures |
| Facial Twitching | Seizure Disorder | Infarction | Seizures |

Figure 16: Learned Phenotypes for Other Neurological Disorders

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| A Fib | Atrial Fibrillation | Rapid Ventricular Response | Pain |
| Atrial Fibrillation | Atrial Fibrillation | Atrial Fibrillation | A Fib |
| RVR | Pain | Cardiomegaly | A Fib |
| A Fib | Stress Ulcer | Atrial Fibrillation | Hemorrhage |
| Pain | Stroke | Acute Cholecystitis | Atelectasis |
| Stress Ulcer | Edema | Calculations | Atrial Fibrillation |
| HTN | Bleeding | Acute Coronary Syndrome | Stroke |
| Stroke | Hypertension | Subdural Hematoma | Pneumothorax |
| Stroke | AVA | Acute on Chronic Renal Failure | HTN |
| Stroke | Ga | Ischemic Heart Disease | Cough |
| Stroke | GI Bleed | Stroke | Stress Ulcer |
| Bleed | Altered Mental Status | Atrial Flutter | Pleural Effusion |
| Aspiration | Bleed | Tachycardia | Intracranial Hemorrhage |
| Narrowing | | Hip Fracture | Shock |

Figure 17: Learned Phenotypes for Cardiac Arrhythmias
| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| polysubstance_abuse | pain | substance_abuse | pain |
| substance_abuse | stress_ulcer | polypsustance_abuse | edema |
| cocaine_abuse | polypsustance_abuse | overdose | pneumothorax |
| overdose | agitated | substance_abuse | headache |
| addiction | asthma | chest_pressure | aneurysm |
| poisoning | chronic_pain | cocaine_abuse | cough |
| rhabdomyolysis | pneumonia | withdrawal | subarachnoid_hemorrhage |
| assault | anxiety | skin_warm | hemorrhage |
| heroin_abuse | fever | fracture | dyspnea |
| hep_c | agitation | epidural_abscess | sob |
| multiple_stab_wounds | substance_abuse | tamponade | hiv |
| bile_leak | respiratory_distress | hep_c | fracture |
| bipolar_disorder | aspiration | chronic_renal_failure | afebrile |
| esophageal_injury | movement | intracranial_hemorrhage | alectasis |
| hep | infection | lower_extremity_weakness | subdural_hematoma |
| | | quadriplegia | headache |
| | | right_hemiplegia | |
| | | | |

Figure 18: Learned Phenotypes for Drug Abuse

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| hemiparesis | stroke | movement | pain |
| stroke | edema | hemiparesis | hemorrhage |
| paraplegia | cva | paraplegia | edema |
| cerebral_palsy | hemorrhage | cebra_palsy | seizure |
| decubitus_ulcers | seizure | cva | fever |
| ischemic_attack | weakness | infarction | syncope |
| lower_extremity_weakness | intracranial_hemorrhage | quadrplegia | bleeding |
| quadriplegia | infarct | brain | epidural_hematoma |
| expressive_aphasia | movement | expressive_aphasia | varix |
| right_hemiplegia | aspiration | pneumonia | sinus_tachycardia |
| quadrplegia | stress_ulcer | hypotension | necrosis |
| contractures | cerebral_infarction | fever | loose_stool |
| thalamic_hemorrhage | infection | syncope | subcutaneous_air |
| mca_infarct | | | afebrile |
| | | | lower_gl_bleed |
| | | | abd |
| | | | ascites |
| | | | lung_cancer |
| | | | aneurysm |

Figure 19: Learned Phenotypes for Paralysis

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| hiv | aids | hiv | pain |
| bacterial_meningitis | pneumonia | scalp_laceration | pneumothorax |
| epidural_hematoma | hypotension | nsf | subarachnoid_hemorrhage |
| cryptocogenic_cirrhosis | fever | posturing | ascites |
| occipital_fracture | syncope | varix | hiv |
| orthostasis | fall | sinus_tachycardia | appendicitis |
| human_immunodeficiency_virus | edema | necrosis | afebrile |
| aids | respiratory_distress | loose_stool | chf |
| acquired_immunodeficiency | bleeding | subcutaneous_air | nausea |
| temporal_bone_fracture | epidural_hematoma | afebrile | bm |
| syncope | braquycardia | lower_gl_bleed | aneurysm |
| hlv_positive | aspiration | abd | opactities |
| memory_loss | cough | ascites | sepis |
| acute_liver_failure | | | abdominal_distension |
| conjunctiva | | | |
| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| **hypotension** | **hypotension** | **metabolic_acidosis** | **pain** |
| **lactic_acidosis** | **hypotension** | **hypodermophosis** | **edema** |
| **hyperkalemia** | **sepsis** | **hypernatremia** | **pneumothorax** |
| **hypertremia** | **acute_renal_failure** | **hyperkalemia** | **hypotension** |
| **renal_failure** | **stress_ulcer** | **hyperkalemia** | **stress_ulcer** |
| **hypoatremia** | **arfd** | **hypopotassium** | **nausea** |
| **respiratory_failure** | **renal_failure** | **acidosis** | **aspiration** |
| **renal_failure** | **infection** | **respiratory_acidosis** | **atelectasis** |
| **hyponatremia** | **ARDS** | **complications** | **cough** |
| **hyperkalemia** | **ARDS** | **obstruction** | **pleural_fffusion** |
| **hyponatremia** | **ARDS** | **ARDS** | **hematoma** |
| **lactic_acidosis** | **ARDS** | **ARDS** | **bleeding** |
| **acute_renal_failure** | **ARDS** | **ARDS** | **pneumonia** |
| **hyposmolality** | **ARDS** | **ARDS** | **subarachnoid_hemorrhage** |

**Figure 21: Learned Phenotypes for Fluid Electrolyte Disorders**

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| **rheumatoid_arthritis** | **pain** | **rheumatoid_arthritis** | **fever** |
| **lupus** | **fever** | **lupus** | **cad** |
| **scleroderma** | **hypotension** | **polyamygia_rheumatica** | **pain** |
| **polymyagia_rheumatica** | **infection** | **hip_fraacture** | **pna** |
| **hip_fracture** | **sepsis** | **scleroderma** | **chf** |
| **absent_bowel_sounds** | **chronic_pain** | **chronic_renal_insufficiency** | **sob** |
| **ankylosing_spondylitis** | **card** | **ankylosing_spondylitis** | **coronary_artery_disease** |
| **imi** | **afibrile** | **interstitial_lung_disease** | **bleeding** |
| **myelosplastic_flagonrome** | **pna** | **dvt** | **ml** |
| **exertional_dyspnea** | **hip_fraacture** | **reflux** | **cp** |
| **eye_pain** | **stress_ulcer** | **feeling_weak** | **crackles** |
| **interstitial_lung_disease** | **hypotensive** | **primary_biliary_cirrhosis** | **pulmonary_edema** |
| **amyloid_angiopathy** | **crackles** | **occlusion** | **edema** |
| **femoral_neck_fraacture** | **femoral_neck_fraacture** | **exertional_dyspnea** | **dementia** |
| **liver_hematoma** | **liver_hematoma** | **tamponade** | **ischemic_heart_disease** |

**Figure 22: Learned Phenotypes for Rheumatoid Arthritis**

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| **multiple_myeloma** | **lymphoma** | **lymphoma** | **lesion** |
| **myeloma** | **multiple_myeloma** | **hodgkins_lymphoma** | **pain** |
| **lymphoma** | **fever** | **multiple_myeloma** | **afib** |
| **hodgkins_lymphoma** | **hypotension** | **myeloma** | **dementia** |
| **achalasia** | **fevers** | **myeloma** | **edema** |
| **amyloidosis** | **pneumonia** | **esophagitis** | **atrial_fibrillation** |
| **remission** | **sob** | **opacities** | **proptosis** |
| **hemochromatosis** | **myeloma** | **edematous** | **peripheral_swelling** |
| **foot_pain** | **hypercalcemia** | **remission** | **infection** |
| **barotrauma** | **hypoxia** | **sah** | **htn** |
| **neutropenic_fever** | **chest_pain** | **orthopnea** | **seizure** |
| **mmn** | **anemia** | **discomfort** | **pneumothorax** |
| **shingles** | **pna** | **hypercalcemia** | **aibcess** |
| **fungemia** | **renal_failure** | **febrile_neutropenia** | **laceration** |
| **hypoxic_brain_injury** | **stress_ulcer** | **subcutaneous_emphysema** | **subdural_hematoma** |

**Figure 23: Learned Phenotypes for Lymphoma**
| 0.4-CNMF | LLDA | MLC | NMF+support |
|---------|------|-----|-------------|
| thrombocytopenia | hit | hypotension | pain |
| hit | coagulopathy | liver_failure | pneumothorax |
| hepatic_encephalopathy | bleeding | ascites | edema |
| hepatorenal_syndrome | fever | edematous | pleural_education |
| cirrhosis_of_liver | acute_renal_failure | generalized_edema | bleeding |
| schistocytes | renal_failure | fatigue | atelectasis |
| low_fibrinogen | arf | cirrhosis | fever |
| splenic_sequestration | infection | cirrhosis | hypotensive |
| hepatic_dysfunction | stress_ulcer | transcription | stress_ulcer |
| polysubstance_abuse | coagulopathy | splenomegaly | fevers |
| liver_cirrhosis | fevers | pulmonary_emia | cough |
| dic | ards | hepatitis_c | sepsis |
| kidney_failure | sinus_tachycardia | hemorrage | hematoma |

**Figure 24: Learned Phenotypes for Coagulopathy**

| 0.4-CNMF | LLDA | MLC | NMF+support |
|---------|------|-----|-------------|
| morbid_obesity | obese | obesity | pain |
| obesity | pain | respiratory_failure | edema |
| osa | obesity | respiratory_failure | edema |
| bronchomalacia | respiratory_failure | respiratory_failure | edema |
| obesity_hypoventilation_syndrome | respiratory_failure | respiratory_failure | edema |
| obstructive_sleep_aphrenia | respiratory_failure | respiratory_failure | edema |
| bronchomalacia | respiratory_failure | respiratory_failure | edema |
| tracheomalacia | respiratory_failure | respiratory_failure | edema |
| pannus | respiratory_failure | respiratory_failure | edema |
| obese | respiratory_failure | respiratory_failure | edema |
| pancreatic_pseudocyst | respiratory_failure | respiratory_failure | edema |
| venous_stasis_ulcers | respiratory_failure | respiratory_failure | edema |
| eeg | respiratory_failure | respiratory_failure | edema |
| daytime_somnolence | respiratory_failure | respiratory_failure | edema |
| group_a_strep | respiratory_failure | respiratory_failure | edema |

**Figure 25: Learned Phenotypes for Obesity**

| 0.4-CNMF | LLDA | MLC | NMF+support |
|---------|------|-----|-------------|
| hip_fracture | pe | dyspnea | pain |
| pulmonary_hypertension | pain | hypoxia | pneumothorax |
| polycythemia | pain | hypoxia | pneumothorax |
| femoral_neck_fracture | pain | hypoxia | pneumothorax |
| pulmonary_infarct | pain | hypoxia | pneumothorax |
| mediastinal_mass | pain | hypoxia | pneumothorax |
| pseudocyst | pain | hypoxia | pneumothorax |
| mucositis | pain | hypoxia | pneumothorax |
| stasis | pain | hypoxia | pneumothorax |
| pulmonary_emia | pain | hypoxia | pneumothorax |
| chest_tightness | pain | hypoxia | pneumothorax |
| pe | pain | hypoxia | pneumothorax |
| pca_infarct | pain | hypoxia | pneumothorax |
| acute_pulmonary_emia | pain | hypoxia | pneumothorax |
| myeloma | pain | hypoxia | pneumothorax |

**Figure 26: Learned Phenotypes for Pulmonary Circulation Disorder**
Figure 27: Learned Phenotypes for Valvular Disease

Figure 28: Learned Phenotypes for Peptic Ulcer

Figure 29: Learned Phenotypes for Congestive Heart Failure
| 0.4-CNMF                  | LLDA          | MLC           | NMF+support   |
|--------------------------|---------------|---------------|---------------|
| hypothyroidism           | pain          | hypothyroidism| pain          |
| hypothyroid              | hypothyroid   | hypothyroid   | pneumothorax  |
| sick_sinus_syndrome      | hypertension  | endometrial_ca| edema         |
| thyroid_ca               | stress_ulcer  | infection     | ateleclasis   |
| respiratory_infection    | edema         | hypoglycemia  | hypothyroidism|
| essential_tremor         | pneumonia     | hypoxia       | stress_ulcer  |
| pancreatic_duct          | hypothyroid   | hip_fracture  | nausea        |
| first_degree_heart_block | bleeding      | cardiomegaly  | hypotension   |
| straining                | nausea        | aortic_stenosis| htn           |
| aplastic_anaemia         | sob           | encephalopathy| sob           |
| acute_delirium           | chronic_pain  | ateleclasis   | pleural_effusion|
| pulm_hypertension        | acute_pain    | hypovolemic   | bleeding      |
| stress                   | pericardial_effusion | meningo ... | afebrile      |

Figure 30: Learned Phenotypes for Hypothyroidism

| 0.4-CNMF                  | LLDA          | MLC           | NMF+support   |
|--------------------------|---------------|---------------|---------------|
| malnutrition             | respiratory_failure | weight_loss  | pain          |
| ulcerative_colitis       | pneumonia     | failure_to_thrive | edema         |
| failure_to_thrive        | wound         | calcifications | hemorrhage    |
| hepatic_cirrhosis        | ascites       | anasarca      | stroke        |
| hydrothorax              | aspiration    | ulcerative_colitis | fever       |
| pancreatic_pseudocyst    | bleeding      | pneumonia      | pneumothorax  |
| volvulus                 | pleural_effusion | volvulus       | subdural_hemorrhage |
| esophageal_varices       | fever         | neutropenic_fever | stress_ulcer |
| gastroparesis            | stress_ulcer  | upper_gastrointestinal_bleed | facial_fra ... |
| bloody_diarrhea          | hypoxia       | glaucoma       | cough         |
| hemochromatosis          | sepsis        | subdural_hemorrhage | ateleclasis |
| necrotizing_fascitis     | pna           | pneumonia      | pneumonia     |
| malnourished             | dvt           | fracture       | fracture      |
| diverticulum             | atelectasis   | intracranial_hemorrhage | intracranial_hemorrhage |
| gastric_cancer           | malnutrition  | epidural_abscess | necrotizing_fascitis |

Figure 31: Learned Phenotypes for Weight loss

| 0.4-CNMF                  | LLDA          | MLC           | NMF+support   |
|--------------------------|---------------|---------------|---------------|
| hypotension              | pain          | iron_deficiency_anemia | pain |
| pain                     | fever         | sinus_rhythm  | pneumothorax  |
| anemia_of_chronic_disease | hypotension   | esrd           | edema         |
| end_stage_renal_disease  | pneumonia     | chronic_renal_failure | nausea     |
| iron_deficiency_anemia   | anemia        | hydronephrosis | sob           |
| hypercalcemia            | sepsis        | mitral_regurgitation | fever       |
| anemia                   | sob           | endocarditis  | pneumothorax  |
| chronic_anemia           | stress_ulcer  | hip_fracture  | subdural_hemorrhage |
| esrd                     | nausea        | vomiting      | stress_ulcer  |
| pancolitis               | cough         | pulmonary_edema | ateleclasis   |
| babesiosis               | infection     | shortness_of_breath | hypothyroid ... |
| microcytic_anemia        | edema         | pyelonephritis | cough         |
| guaiac_stools            | fevers        | gerd           | pneumonia     |
| dry_gangrene             | chest_pain    | uti            | afebrile      |
|                         | pna           |                | chest_pain    |

Figure 32: Learned Phenotypes for Deficiency Anemias
Figure 33: Learned Phenotypes for Blood Loss Anemia

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| cryptogenic_cirrhosis | pain | fulminant_hepatic_failure | pain |
| squamous_cell_carcinoma | bleeding | tired | bleeding |
| heel_ulcer | gi | hocm | chf |
| diverticular_disease | anemia | hit | pneumothorax |
| lactate_levels | stress_ulcer | restless | hit |
| anastomotic_leak | hives | lower_gi_bleed | lower_gi_bleed |
| dark_stools | hypotension | effusions | effusions |
| gangrenous_cholecystis | gastrointestinal_bleed | calcifications | calcifications |
| gastropathy | abdominal_pain | peripheral_neuropathy | peripheral_neuropathy |
| bowel_perforation | chest_pain | blood_loss | blood_loss |
| portal_hypertensive_gastropathy | melena | unresponsiveness | unresponsiveness |
| syncopal_episodes | wound | sinus_tachycardia | sinus_tachycardia |
| angioedema | chf | bacteremia | bacteremia |
| neutropenic_fever | diarrhea | upper_gi_bleed | upper_gi_bleed |
| irritable_bowel_syndrome | | duodenal_perforation | duodenal_perforation |

Figure 34: Learned Phenotypes for Depression

| 0.4-CNMF | LLDA | MLC | NMF+support |
|----------|------|-----|-------------|
| depression | pain | depression | pain |
| overdose | depression | systolic_dysfunction | hypotension |
| serotonin_synrome | stress_ulcer | overdose | bleeding |
| od | anxiety | chronic_pain | sob |
| fibromyalgia | nausea | osteoarthritis | edema |
| clonus | chest_pain | ha | depression |
| blurred_vision | hypotension | blurred_vision | stress_ulcer |
| elevated_ammonia | aspiration | chest_pressure | nausea |
| type_1_diabetes | fever | cerebral_edema | bleed |
| crohns_disease | | back_pain | pneumothorax |
| fulminant_hepatic_failure | sob | lightheaded | atelectasis |
| liver_injury | chronic_pain | pulmonary_edema | aspiration |
| toxic_ingestion | bleeding | obesity | hematoma |
| vp_shunt | abdominal_pain | osa | hematoma |
| bronchopleural_fistula | vomiting | hypothyroidism | pleural_effusion |
| | htn | | anxiety |
Appendix C. Augmented Mortality Prediction

Figure 35 shows weights learned by the classifier for all features. The weights shaded red correspond to phenotypes and are relatively high compared to raw notes based features (shaded blue), indicating that comorbidities capture significant amount of predictive information on mortality and achieve comparable performance to full EHR model when augmented with additional raw clinical terms.

Figure 35: Weights learned by the CNMF+Full EHR classifier for all features. The weights shaded red correspond to phenotypes.