Phenotyping through Semi-Supervised Tensor Factorization (PSST)

Jette Henderson, MS[1], Huan He, MS[2], Bradley A. Malin, PhD[3], Joshua C. Denny, MD MS[3], Abel N. Kho, MD[4], Joydeep Ghosh, PhD[1], Joyce C. Ho, PhD[2]
[1]The University of Texas at Austin, Austin, TX; [2]Emory University, Atlanta, GA; [3]Vanderbilt University, Nashville, TN; [4]Northwestern University, Evanston, IL

Abstract

A computational phenotype is a set of clinically relevant and interesting characteristics that describe patients with a given condition. Various machine learning methods have been proposed to derive phenotypes in an automatic, high-throughput manner. Among these methods, computational phenotyping through tensor factorization has been shown to produce clinically interesting phenotypes. However, few of these methods incorporate auxiliary patient information into the phenotype derivation process. In this work, we introduce Phenotyping through Semi-Supervised Tensor Factorization (PSST), a method that leverages disease status knowledge about subsets of patients to generate computational phenotypes from tensors constructed from the electronic health records of patients. We demonstrate the potential of PSST to uncover predictive and clinically interesting computational phenotypes through case studies focusing on type-2 diabetes and resistant hypertension. PSST yields more discriminative phenotypes compared to the unsupervised methods and more meaningful phenotypes compared to a supervised method.

Introduction

Computational phenotypes are meaningful and actionable disease- or condition-specific patient characterizations that can be derived from electronic health records (EHRs). It has been shown that such derived knowledge can provide healthcare practitioners with a better understanding of their underlying populations[1,2,3]. Additionally, they can support the practice of precision medicine via clinical predictive modeling and improve comparative effectiveness research, as well as advance our understanding of disease risk and drug responses[4]. In the past, computational phenotyping approaches were based predominantly on rules and supervised machine learning, which required domain expertise and only identified phenotypes that were essentially already known[5,6]. Rule-based methods replicate clinical knowledge, which may be the goal in some situations (e.g., the phenotyping efforts of the eMERGE network)[7]. Recently, unsupervised machine learning approaches have been proposed to extract meaningful computational phenotypes that minimize domain experts' efforts. These methods offer a viable alternative to directly mining electronic health records, which requires expert-defined labels and other domain expertise. Techniques such as frequent pattern mining and deep learning have been introduced for computational phenotyping[8,9,10], yet they are limited in that they 1) fail to find the underlying latent characteristics or 2) are difficult to comprehend due to nonlinear combinations of multiple layers.

Dimensionality-reduction methods for computational phenotyping have gained in popularity due to their robustness to sparse and noisy data and to their interpretability because they allow patients to be probabilistically assigned to latent subgroups. In particular, various tensor factorization approaches have been proposed to encapsulate the interaction between multiple information sources (e.g., diagnosis, medications, and procedures)[11,12,13,14]. Tensors, which are generalizations of vectors and matrices to higher dimensions, are ideal for capturing the multidimensional relationships inherent in EHR count and continuous data[15]. Unlike a matrix, a tensor can readily express the relationship between a medication and the different diseases it is prescribed to manage. For example, metformin, commonly prescribed to manage diabetes, has also shown promise in treating the symptoms of polycystic ovary syndrome[16]. Ho et al showed that tensor factorization could be applied to tensors constructed from EHRs to derive phenotypes that map to clinically meaningful concepts[11]. Subsequent methods have improved the resulting phenotypes to fit clinicians' expectations of sparsity and diversity[12,13,14].

While tensor-based approaches can extract interpretable and meaningful phenotypes and potentially uncover novel characterizations, nearly all existing methods neglect auxiliary patient information (e.g., disease status, gender, age, etc). Such auxiliary patient information can be useful in guiding the derivation process and potentially yield more discriminative, interpretable, and meaningful computational phenotypes. Furthermore, one of the potential goals of computational phenotyping is to use derived phenotypes to identify case and control patients for future studies,[17] which makes it problematic if patients with different disease statuses appear in the same derived phenotypes. To overcome this challenge, we introduce Phenotyping through Semi-Supervised Tensor Factorization (PSST), a novel method that uses auxiliary patient information to discourage patients with different disease statuses from appearing in the same phenotypes. We posit that the use of a semi-supervised based approach to leverage *known information available for a subset of the patients* will lead to phenotypes that are descriptive of the interplay between different diseases. We demonstrate the potential of PSST to extract clinically interesting and discriminative phenotypes by focusing on a dataset of 1,622 patients gathered at Vanderbilt University Medical Center (VUMC) where the disease status is known for a subset of patients. Specifically, we construct a tensor that consists of the following four types of patients: cases and controls of resistant hypertension patients and cases and controls of type-2 diabetes patients. We compare PSST with three other tensor-based computational phenotyping methods, two of which are unsupervised and one of which is supervised. This investigation demonstrates that using disease status for a specific diagnosis (e.g., resistant hypertension or type-2 diabetes) can reveal discriminative phenotypes–even for other diagnoses–that may not be realized in fully supervised or unsupervised approaches.

Methods

Phenotyping via tensor factorization. A tensor is a generalization of a matrix to a multidimensional array where each element of a tensor represents an n-way interaction. Tensors have the capability to capture complex relationships that exist in healthcare. In this paper, we consider tensors with three dimensions (or *modes*) – patients, diagnoses, and medications. Each element in the tensor represents the number of times each patient was prescribed a medication within a specified amount of time of receiving a diagnosis.

Prior to describing our approach, we introduce notation and concepts used throughout this paper. The number of dimensions (or modes) in a tensor is called the *order* of the tensor. A vector is a tensor of order 1, and a matrix is a tensor of order 2. A tensor can be decomposed using information in the multidimensional structure to extract succinct components that are more likely to be interpretable than the raw data contained in the original tensor[15]. A common tensor decomposition model is the CANDECOMP/PARAFAC (CP) decomposition. CP decomposition can be thought of as an extension of Singular Value Decomposition or Principal Component Analysis from matrices to higher-ordered tensors. CP decomposition factorizes the original tensor $\boldsymbol{\mathcal{X}}$ as a sum of R rank-one tensors where an N-way *rank-one tensor* can be expressed as the outer product of N factor vectors[18,19]. The CP decomposition is denoted as:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{Z}} = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C}]\!] \tag{1}$$

We will use $\boldsymbol{\mathcal{Z}} = [\![\boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C}]\!]$ as shorthand notation where the λ_r weights are organized into the vector $\boldsymbol{\lambda} = [\lambda_1 \cdots \lambda_R]$ and the factor column vectors are stacked into factor matrices (e.g., $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_R]$). The most popular algorithm for fitting the CP model is CP-Alternating Least Squares (CP-ALS), which assumes the underlying distribution of the data is Gaussian[18,19]. For count (nonnegative integer) data, a Poisson-based distribution has been proposed to yield non-negative factor vectors consistent with the observed data[20].

Computational phenotypes can be derived by constructing a tensor from patient-level EHR data and factoring it via the CP model, illustrated in Figure 1. Here, the original tensor $\boldsymbol{\mathcal{X}}$ is a 3-mode tensor (i.e., patients, diagnoses, and medications) that is decomposed into a sum of R rank-one tensors. Each rank-one tensor is formed by taking the outer product of a patient factor vector $\mathbf{a}_r$, a diagnosis factor vector $\mathbf{b}_r$, and a medication factor vector $\mathbf{c}_r$. The rank-one tensor can then be interpreted as a phenotype where the non-zero elements (denoted as colored blocks) of each factor vector form the clinical characteristics. The weight λ_r denotes the importance of the r^{th} rank-one tensor (i.e., computational phenotype) in terms of explaining the observed EHR tensor. For EHR tensors where the elements contain the co-occurrences of patients, diagnoses, and medications (counts of the 3-way interaction), the Poisson-based loss function has been shown to be robust and to yield interpretable, clinically relevant patterns for practitioners[11,12].

Marble is a sparse, nonnegative tensor factorization method that simultaneously derives multiple phenotype candidates with virtually no domain expert supervision[12]. The algorithm decomposes the observed EHR tensor into two terms, a rank-one bias tensor and an interaction tensor. The bias tensor represents the baseline characteristics common among the overall population and the interaction tensor defines the phenotype candidates, as shown in Figure 1. Granite, an extension of Marble, was introduced by Henderson et al. to yield diverse phenotypes[14]. Granite incorporates an angular penalty into the loss function to produce phenotypes that are both sparse and diverse, which are characteristics valuable to clinicians. Both Marble and Granite are purely unsupervised methods (no domain expertise required).

Other tensor factorization models have been proposed to extract computational phenotypes and incorporate some form of supervision. Rubik, a model that uses the same assumptions as CP-ALS, introduced a combination of pairwise constraints on the vectors in the factor matrices and guidance matrices to improve the meaningfulness of the factors[13]. Rubik's guidance matrices, which encode information that is already known, attempt to induce classes that have minimal overlap by guiding the non-patient modes using domain knowledge. However by focusing on guiding non-patient modes, Rubik's approach may leave out clinically interesting phenotypes. Kim et al. also proposed a supervised tensor factorization method where patient outcome information guides the tensor decomposition to discover phenotypes that are good predictors of patient outcomes for unseen patients as well as generate distinct phenotypes[21]. However, this work, hereinafter referred to as DDP (Discriminative and Distinct Phenotyping), requires complete knowledge of the outcomes for each cohort patient. Furthermore, they use preprocessing methods to ensure all terms in a phenotype are similar and cohesive. Like the guidance provided in Rubik, this approach could smooth over novel phenotypes important to understanding a condition.

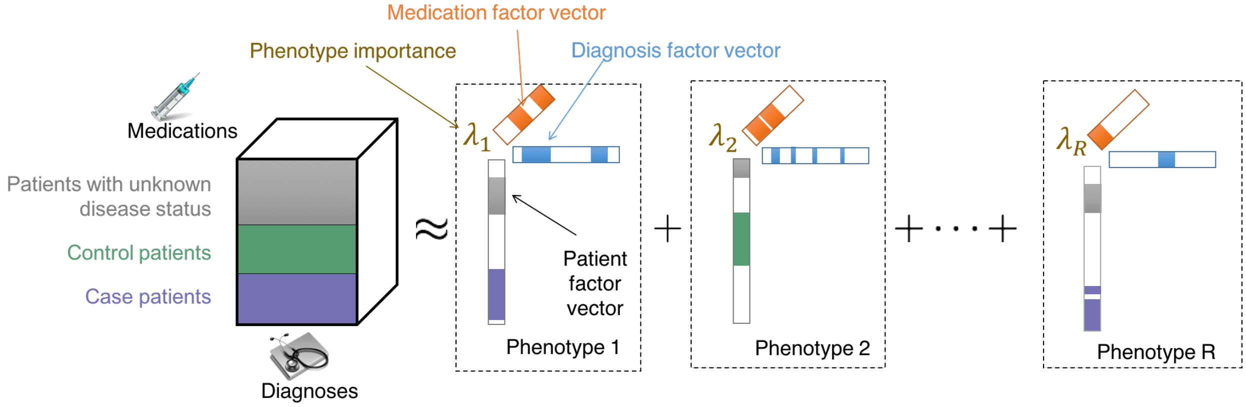


Figure 1: An example of phenotyping via tensor factorization. The tensor containing the observed data is pictured as the cube on the left. Each element of the observed tensor corresponds to the number of times a patient received a medication prescription and diagnosis in a set amount of time. A set of rank-one components, formed by taking the outer product of a patient, a diagnosis, and a medication factor vector, is found by minimizing a loss function. The non-zero elements in each component are indicated by colored bars in the factor vectors and consist of the clinical characteristics in that phenotype. The goal of PSST is to use information about the disease status of just a few of the patients within the tensor to encourage patients with different statuses to be in different components. This is indicated by the various colored blocks in the patient factor vectors.

Semi-supervised tensor factorization. Semi-supervised learning (SSL) is a hybrid of supervised and unsupervised learning where there is a (small) portion of labeled data and unlabeled data. The assumption in SSL is that the unlabeled data provides information about the distribution of the examples that are useful. One class of approaches, transductive SSL, is useful in situations where we know something about the relationships between observations and wish to incorporate that information into the learning process[22]. In particular, semi-supervised clustering introduces the notion that there are pairs of data points that must be clustered together, or *must-link*, and pairs that must not be clustered together in the same cluster, or *cannot-link*. While tensor factorization is similar to clustering, relatively few tensor decomposition methods incorporate semi-supervision. Peng introduced must-link and cannot-link constraints for the least squares objective function (data follows Gaussian distribution)[23]. We use the cannot-link constraints approach but formulate our method for count data.

Through cannot-link constraints, PSST encourages patients with different disease statuses to be affiliated with different phenotypes. For example, when trying to identify interesting groups of patients, we may have prior information (e.g., case or control for a specific disease) that a subset of patients should be grouped together while some others should not[24]. We propose the use of soft cannot-link constraints on the patient phenotype membership matrix ($\mathbf{A}$) to encourage separation between the known classes of patients. Here soft constraints refer to restrictions that will result in a penalty if they are not met. By using partial patient status information, PSST does not dictate which phenotypes should be extracted. Rather, it encourages patients with different disease statuses to map to different phenotypes. Moreover, it does not require all patients have a known label.

PSST Mathematical Formulation. Our work is developed for EHR count tensors, where each element represents an n-way co-occurrence count. PSST is built on existing nonnegative CP decomposition algorithms that model the observed data using the Poisson distribution[11,12,14]. For simplicity, we focus on a 3-mode tensor where the three dimensions are (1) patients, (2) diagnoses, and (3) medications. However, our approach can easily generalize to an N-mode tensor. An observed tensor, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is approximated as the sum of R 3-way rank-one tensors $\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{Z}} = [\![\boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C}]\!]$. We place several constraints on the resulting factor matrices to encourage phenotypes with certain characteristics. For the diagnoses and medication factor matrices ($\mathbf{B}$ and $\mathbf{C}$ respectively), we incorporate angular penalty matrices to diversify the phenotypes whenever possible. PSST also introduces a cannot-link matrix ($\mathbf{M}$) on the patient factor matrix ($\mathbf{A}$) to encourage separation in the patients, where different disease statuses are in different phenotypes (e.g., hypertension case patients and hypertension control patients). This notion is illustrated in Figure 1. The optimization problem for the observed tensor $\boldsymbol{\mathcal{X}}$ is:

$$f(\boldsymbol{\mathcal{X}}) = \min(\underbrace{\sum_{\vec{i}}(z_{\vec{i}} - x_{\vec{i}} \log z_{\vec{i}})}_{\text{KL divergence}} + \underbrace{\beta_1 \text{trace}(\mathbf{A}^\intercal \mathbf{M} \mathbf{A})}_{\text{cannot-link constraints}} + \underbrace{\frac{\beta_2}{2} \sum_{r=1}^{R} (||\mathbf{a}_r||_2^2 + ||\mathbf{b}_r||_2^2 + ||\mathbf{c}_r||_2^2)}_{\ell_2 \text{regularization}} \tag{2}$$

$$+ \frac{\beta_3}{2} \sum_{r=1}^{R} \sum_{p=1}^{r} \left((\max\{0, \frac{(\mathbf{b}_p)^\intercal \mathbf{b}_r}{||\mathbf{b}_p||_2 ||\mathbf{b}_r||_2} - \theta\})^2 + (\max\{0, \frac{(\mathbf{c}_p)^\intercal \mathbf{c}_r}{||\mathbf{c}_p||_2 ||\mathbf{c}_r||_2} - \theta\})^2 \right) \tag{3}$$

$$\text{s.t } \boldsymbol{\mathcal{Z}} = [\![\sigma; \mathbf{u}_a; \mathbf{u}_b; \mathbf{u}_c]\!] + [\![\boldsymbol{\lambda}; \mathbf{A}; \mathbf{B}; \mathbf{C}]\!] \tag{4}$$

$$||\mathbf{a}_r||_1 = ||\mathbf{b}_r||_1 = ||\mathbf{c}_r||_1 = ||\mathbf{u}_a||_1 = ||\mathbf{u}_b||_1 = ||\mathbf{u}_c||_1 = 1, \mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r \geq 0, \mathbf{u}_a, \mathbf{u}_b, \mathbf{u}_c > 0 \tag{5}$$

For count data, the loss function is the negative log-likelihood between the observed data $\mathbf{x}$ and the model $\mathbf{z}$ parameters (i.e., the term labeled "KL divergence" in (2)). As introduced in Granite, an angular penalty term (3) discourages any factors that are too similar, where similarity is defined as the cosine angle between two factor vectors. Additionally to control the growth of the size of the factors and for computational stability, we include an l_2 penalty term (labeled in (2)).

Unlike Granite and Marble, PSST incorporates partial class knowledge to encourage patients with different disease statuses to appear in different phenotypes using a cannot-link semi-supervised penalty term. In the term labeled "cannot-link constraints" in (2), the cannot-link matrix $\mathbf{M} \in \mathbb{R}^{I_1 \times I_1}$ is constructed such that $m_{i,j} = 1$ only if patients i and j have different disease statuses and is otherwise 0. If patients i and j are in different classes but both belong to phenotype r, then the penalty $a_{ir} \cdot a_{jr}$ is added to the objective function. Thus the cannot-link constraint term will only contribute to the objective function if two patients have two different disease statuses (e.g., one patient is a case and one is a control) and will be 0 otherwise (e.g., both patients are case, both are control, or one of them is unknown). Figure 1 illustrates the impact of this cannot-link term, phenotype 1 and R consists of cases and patients with unknown disease status and phenotype 2 consists of controls and patients with unknown disease status. Since this is a soft penalty, some case and control patients can be in the same phenotype–provided they are highly similar. We use gradient descent to solve the optimization problem.

Experiment Design

Dataset and preprocessing. We constructed a tensor from the diagnosis and medication counts of 1,622 patients from the Synthetic Derivative (SD), a de-identified EHR database of VUMC patients[25]. The SD contains clinical and billing code information for over 2 million inpatient and outpatient interactions. Previously, a panel of domain experts identified sets of characteristics in the form of billing and medical codes of patients as case and control for a set of diseases[26]. In this paper, we focus on resistant hypertension case and control patients and type-2 diabetes case and control patients. A small subset of these patients are both resistant hypertension and type-2 diabetes cases (see Table 1 for the number of patients in each class).

Table 1: Patient disease status (supplied by domain experts) in the VUMC SD dataset used in this study.

Disease Class	**Number of Patients**
Resistant hypertension case	304
Resistant hypertension control	399
Type 2 diabetes case	373
Type 2 diabetes control	452
Type 2 diabetes and resistant hypertension case	94

For each case patient, we counted the medication and diagnosis interactions that occurred two years before they received the diagnosis of the disease (i.e., hypertension or type-2 diabetes). For each control patient, we counted the medication and diagnosis interactions that occurred two years before their last interaction with the VUMC. The diagnosis codes follow the International Classification of Diseases (ICD-9) system and capture information at a high level of detail for insurance purposes. We use PheWAS coding to aggregate the diagnosis codes into broader categories[24]. Additionally, we use MeSH pharmacological terms provided by the RxClass RESTful API, a service of the US National Library of Medicine, to group the medications into more general categories (https://rxnav.nlm.nih.gov/RxClassAPIs.html). These groupings resulted in a tensor with the following dimensions: 1622 patients by 1325 diagnoses by 148 medications.

Evaluation Metrics. We evaluate PSST with respect to three criteria: (1) the efficacy of the cannot-link constraint in encouraging case and control patients to belong to different phenotypes, (2) the discriminative quality of the resulting phenotypes on an unrelated classification task, and (3) the clinical meaningfulness of the resulting phenotypes.

For the second evaluation metric, we use a cannot-link matrix on resistant hypertension case and control patients to perform the factorization and then use the resulting patient factor matrix to predict which are the type-2 diabetes cases and controls. Likewise, we reversed the two, where the tensor factorization is carried out with a cannot-link matrix on the type-2 diabetes case and control patients, and then the resulting patient factor matrix was used to predict resistant hypertension. For each classification task, we row-normalize the patient factor matrix ($\mathbf{A}$) to obtain a phenotype membership (probability that a patient belongs to each phenotype). Then, using a logistic regression model, we perform a 5-fold cross-validation to evaluate the lift and the area under the receiver operating curve (AUC). Lift is the ratio between the results obtained through the predictive model and results obtained without a model. Our hypothesis is that the resistant hypertension cannot-link constraints in PSST will result in phenotypes that uncover latent factors pertinent to type-2 diabetes patients and that type-2 diabetes cannot-link constraints will have a similar effect for identifying hypertension case patients.

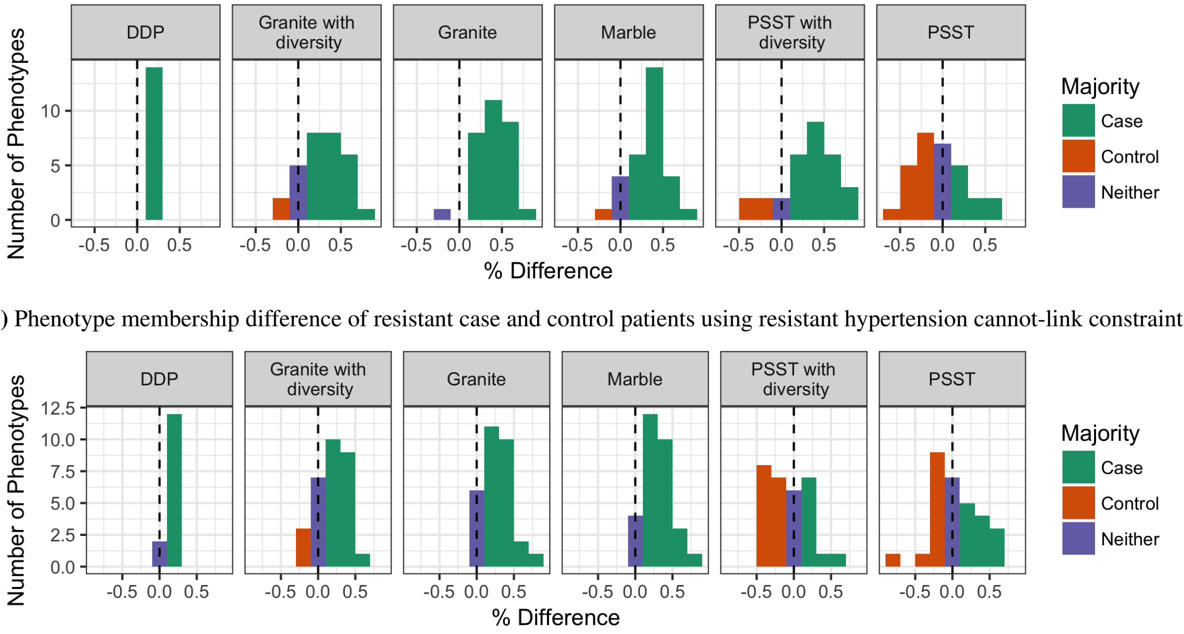
To evaluate the clinical meaningfulness, we enlisted two clinicians to annotate the phenotypes as clinically relevant or not clinically relevant. To reduce the annotation burden, the classification task results were used to identify highly predictive phenotypes and these were randomly shuffled to avoid biasing the experts.

Unsupervised and Supervised Comparison Models. We compared PSST with three other tensor factorization methods: Marble[12], Granite[14], and DDP[21]. Marble has two sets of parameters relating to the strength of the underlying characteristics (bias term) and the sparsity of the resulting factors. These parameters are tuned to achieve comparable results with respect to the number of non-zero elements per computational phenotype. Granite has both a sparsity-inducing and a diversity-inducing regularization term to yield a sparse set of diverse phenotypes. The Granite parameters (excluding the diversity-inducing term) are tuned to yield the best predictive accuracy. DDP incorporates a logistic regression term, as well as a similarity-based cluster structure, to encourage distinctness. Since this cluster structure requires existing knowledge, we excluded it from our analysis.

Results

We chose $R = 30$ phenotypes for PSST, Marble, and Granite through a grid search on R. Since DDP was restricted to case and control patients and resulted in a smaller tensor, we found 15 phenotypes resulted in a reasonably good fit.

Efficacy of Cannot-Link Constraints: Class Separation in Patient Factor Matrix. After fitting the PSST decomposition, we analyzed how well it encouraged class separation within the phenotypes, and we compared it to the performance of DDP, Granite, and Marble. For these experiments, we show results for two formulations of PSST and Granite, one with the angular penalty, denoted as "with diversity," and without the angular penalty. In each phenotype, we calculated the percentage of patients who were case and the percentage of patients that were control and then took the difference. For example, a difference of .2 in phenotype k means that the control class consisted of 40% of the phenotype while the case class consisted of 60% of the phenotype. Figures 2a and 2b depict histograms of the difference between the percentages within each phenotype for PSST (with and without diversity constraints), Marble, Granite (with and without diversity constraints), and DDP. The bin color was set to orange (majority control) if the difference was $< -.1$, to teal if the difference was $> .1$ (majority case), and to purple (no majority) otherwise. Ideally, there should be bars on each side of the dotted line. This would indicate that there are phenotypes that are distinct to case patients and to control patients (i.e., they mostly contain case patients or mostly contain control patients). Figure 2a shows that PSST with and without diversity resulted in phenotypes where the majority was either hypertension case (teal bins) or hypertension control (orange bins). Marble and Granite (with and without diversity) resulted in phenotypes that most often consisted of case patients, and DDP resulted in phenotypes that consisted only of case patients. Thus, the competing methods fail to separate the case and control patients and fail to discover phenotypes distinct to the disease statuses.



(a) Phenotype membership difference of resistant case and control patients using resistant hypertension cannot-link constraints.

(b) Phenotype membership difference for type-2 diabetes case and control patients using type-2 diabetes cannot-link constraints.

Figure 2: Histograms of difference between the percent membership by class for patients using disease-specific cannot-link constraints. The x-axis ("% Difference") is the difference between the percentage of case patients and the percentage of control patients in each phenotype, while the y-axis is the number of phenotypes. A positive difference refers to more case patients (green), a negative difference refers to more controls (orange), and approximately 0 means neither population dominates (purple).

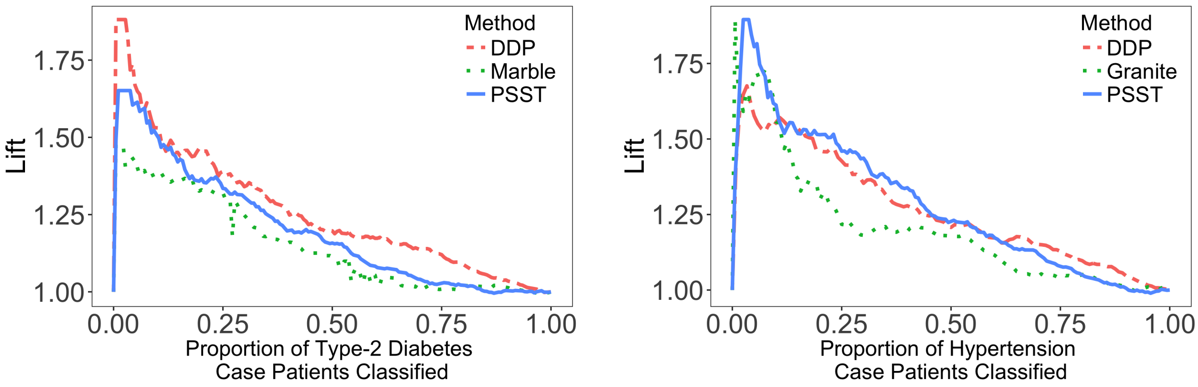
Similarly in Figure 2b, PSST with and without diversity constraints results in phenotypes that are either primarily type-2 diabetes case or control patients. Granite with diversity was the only decomposition aside from PSST to derive any phenotypes consisting of a majority control patients. DDP's lack of separation between patient classes is surprising given that it incorporates a logistic regression loss term in its fitting process. In both case studies, the cannot-link constraints in PSST encourage class separation within the phenotypes.

Discriminative Evaluation. Using logistic regression, we compared how well each method discriminates between case and control patients. For PSST, we predict case and control patients that were not used in the cannot-link constraints. Specifically, if a fit used the cannot-link constraints on type-2 diabetes case and control patients, we then predict the resistant hypertension case and control patients, and vice versa for cannot-link constraints on resistant hypertension. The features for the logistic regression are the row-normalized patient factor matrix and restricted to the rows corresponding to case and control patients. Table 2 shows the AUC values averaged across the five runs for each method for predicting resistant hypertension and type-2 diabetes. As expected, the supervised method DDP outperformed all methods, but PSST had the second highest AUC for each condition. Secondly, there is a tradeoff between diversity constraints (e.g., in PSST and Granite) and the predictive quality of the phenotypes, which was previously noted by Henderson et al[14]. Furthermore, the relatively low AUC values indicate that these are difficult classification problems, but the performance of PSST implies that incorporating knowledge about a subset of patients can be beneficial.

Table 2: AUC for predicting case and control patients using decompositions with cannot-link constraints on the other case and control patients. For example, "Hypertension" below refers to the AUC for predicting hypertension patients when the cannot-link constraints were applied to type-2 diabetes case and control patients.

	Condition	
Method	**Hypertension**	**Type-2 Diabetes**
PSST	0.6618	0.6074
PSST with diversity	0.6275	0.5830
DDP	0.6928	0.6614
Granite	0.6074	0.5528
Granite with diversity	0.5939	0.5745
Marble	0.5919	0.5928

Figures 3a and 3b show the lift of the three methods with the highest AUCs in each classification task. When predicting who is a type-2 diabetes case patient (Figure 3a), DDP has a higher lift than Marble and Granite. On the other hand, when predicting which patients are resistant hypertension case and control in this particular instance (Figure 3b), PSST consistently has the highest lift. This is surprising given DDP incorporates the resistant hypertension case and control status into fitting the decomposition and has the highest AUC. This indicates that semi-supervision in PSST could be guiding the decomposition toward phenotypes that are meaningful for resistant hypertension patients.



(a) Type-2 diabetes prediction task.

(b) Resistant hypertension prediction task.

Figure 3: Lift curves for the two prediction tasks

Clinical Relevance Evaluation. As a final step in our analysis, two clinicians annotated the clinical relevance of the phenotypes generated by PSST, Marble, and DDP that were most predictive of being a resistant hypertension case patient. The clinicians assigned each phenotype one of the following labels: 1) clinically meaningful, 2) possibly clinically meaningful, and 3) not clinically meaningful. In total, the clinicians annotated 5 PSST-, 5 Marble-, and 3 DDP-generated phenotypes (DDP had only three positive coefficients). In cases where the annotator's disagreed, we used the label with the lowest clinical relevance score. Using Cohen's Kappa, the inter-rater reliability score was $\kappa = .45$, suggesting the inter-rater agreement was moderate.

Figure 4 shows the distribution of the annotations by method. For DDP, 66% of the phenotypes were possibly or not clinically meaningful, suggesting there may be a trade-off between seemingly good predictive quality and clinical relevance. PSST and Marble had the same number of clinically relevant phenotypes, with only 20% deemed not significant. By incorporating semi-supervision through soft constraints, PSST maintains predictive power and interpretative value in this case study.

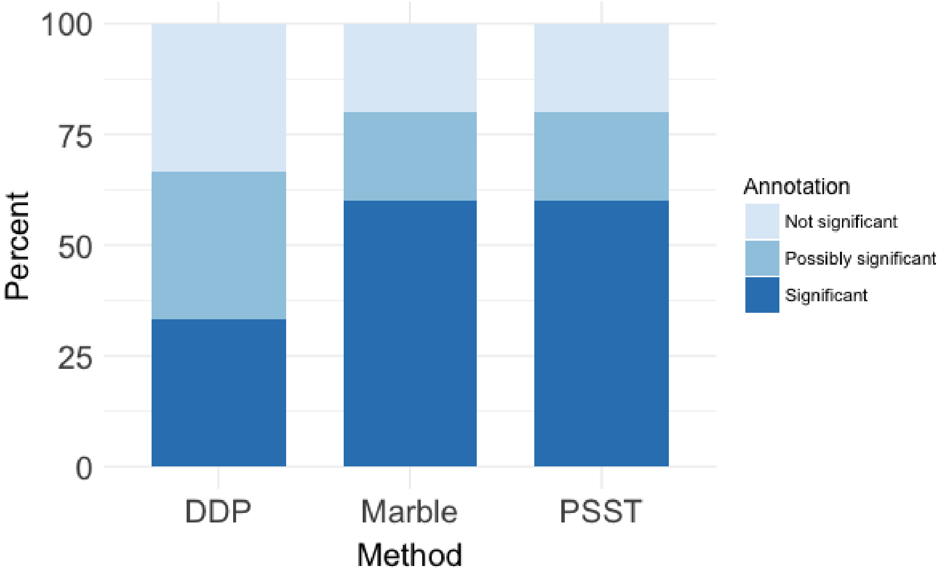


Figure 4: The percentage of most predictive phenotypes generated by PSST, Marble, and DDP phenotypes that were clinically significant, possibly clinically significant, not clinically significant.

Discussion and Conclusions

PSST, which only incorporates partial patient information, resulted in phenotypes that exhibited a high degree of separation between case and control patients. The phenotypes extracted by PSST were more predictive of case and control for the two conditions hypertension and type-2 diabetes than two unsupervised methods. It did not perform quite as well on the prediction task as the supervised method, DDP, but DDP requires complete knowledge of patient disease status while PSST only needs information about a subset of patients. Additionally, in terms of clinical relevance, the phenotypes produced by DDP were not as clinically relevant overall as compared to PSST. This implies that for DDP there may be a trade-off between clinical relevance and predictive power. Furthermore, DDP requires labels for all patients, and the cost of obtaining labels in medical informatics can be high in terms of time and expertise required. Therefore, a semi-supervised method like PSST could help researchers use information available to them without restricting their work to labeled observations.

One major challenge in extracting phenotypes through automatic, machine learning methods is verifying the phenotypes are clinically interesting and meaningful. This validation step is a task that requires domain expertise and time. Furthermore, the phenotypes themselves should be annotated by a panel of experts, and the analysis in the previous section showed that annotators do not agree on the clinical significance of a phenotype at all times. Therefore, it may be beneficial to use a third-party annotator. For this purpose, we developed PheKnow–Cloud, a tool that uses co-occurrence analysis on a publicly available repository of medical articles to calculate a clinical validity score for

Table 3: Example of phenotype labelled "possibly clinically significant."

Diagnoses	**Medications**
Hyperlipidemia	Angiotensin converting enzyme inhibitors
GERD	Antihyperlipidemic agents
	Antiadrenergic agents, centrally acting

a supplied phenotype[27]. PheKnow–Cloud could prove useful for situations where annotators labeled a phenotype as "possibly clinically significant," as they did for a PSST phenotype show in Table 3. According to PheKnow–Cloud, this phenotype is likely clinically meaningful, which may lead to further discussion between the annotators.

In conclusion, we presented Phenotyping through Semi-Supervised Tensor factorization, or PSST, a method that incorporates information from subsets of patients to encourage class separation in patient phenotype membership. Using two case studies, we demonstrated the benefits of integrating partial information into the tensor factorization process to derive phenotypes. We showed the semi-supervised constraints induce considerable class separation between patients with different disease statuses (i.e., case and control) whereas a supervised and two unsupervised methods resulted in little to no class separation. Additionally, PSST may help extract phenotypes that are more descriptive and predictive of patients' disease statuses than purely unsupervised methods, and while PSST did not outperform a supervised method on a prediction task, it did result in phenotypes that were more interpretable than those of the supervised method. We note that this is a pilot study and more study is necessary to provide sufficient evidence of PSST's viability, but the early results are promising.

There are opportunities to extend PSST to a larger sets of conditions and outcomes. For example, if within a set of hypertension patients, we knew a subset had heart attacks, it would be useful to put cannot-link constraints between those who had heart attacks and those who did not and examine the resulting phenotypes. Furthermore, it would be useful to analyze which phenotypes are highlighting indicators of disease progression or outcome.

Acknowledgements

This work was supported by NSF grants 1418511, 1417819, and 1418504.

References

1. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013;20(e2):e206–11.

2. Xu J, Rasmussen LV, Shaw PL, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc. 2015;22(6):1251–60.

3. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. J Am Med Inform Assoc. 2018;25(3):289–94.

4. Richesson RL, Sun J, Pathak J, et al. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med. 2016;71:57–61.

5. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):13.

6. Carroll RJ, Eyler AE, Denny JC. Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. In: AMIA Annu Symp Proc; 2011. 189–96.

7. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inf Assoc. 2013;20(e1):e147–54.

8. Moskovitch R, Shahar Y. Medical temporal-knowledge discovery via temporal abstraction. In: AMIA Annu Symp Proc; 2009. 452–456.

9. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PloS one. 2013;8(6):1–13.

10. Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. In: Proc of ACM Conference SIGKDD Conference on Knowledge Discovery and Data Mining. ACM; 2016. 1495–1504.

11. Ho JC, Ghosh J, Steinhubl SR, et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inf. 2014 Dec;52:199–211.

12. Ho JC, Ghosh J, Sun J. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In: Proc of ACM Conference SIGKDD Conference on Knowledge Discovery and Data Mining; 2014. 115–24.

13. Wang Y, Chen R, Ghosh J, et al. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In: Proc of ACM Conference SIGKDD Conference on Knowledge Discovery and Data Mining; 2015. 1265–74.

14. Henderson J, Ho JC, Kho AN, et al. Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping. In: Proc of IEEE International Conference on Healthcare Informatics; 2017. 214–23.

15. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Review. 2009;51(3):455–500.

16. Lashen H. Role of metformin in the management of polycystic ovary syndrome. Ther Adv Endocrinol Metab. 2010;1(3):117–28.

17. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117–21.

18. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. Psychometrika. 1970;35(3):283–319.

19. Harshman RA. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis. UCLA Working Papers in Phonetics. 1970;16:1–84.

20. Chi EC, Kolda TG. On tensors, sparsity, and nonnegative factorizations. SIAM J Matrix Anal Appl. 2012;33(4):1272–99.

21. Kim Y, El-Kareh R, Sun J, et al. Discriminative and distinct phenotyping by constrained tensor factorization. Scientific reports. 2017;7(1):1114.

22. Zhu X. Semi-Supervised Learning. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010. 892–97.

23. Peng W. Constrained Nonnegative Tensor Factorization for Clustering. In: 2010 Ninth International Conference on Machine Learning and Applications; 2010. 954–957.

24. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31(12):1102.

25. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008;84(3):362–69.

26. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010;86(4):560–572.

27. Henderson J, Bridges R, Ho JC, et al. PheKnow-Cloud: A Tool for Evaluating High-Throughput Phenotype Candidates using Online Medical Literature. In: Proc of AMIA Jt Summits Transl Sci Proc. vol. 2017. AMIA; 2017. 149–57.