Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Limestone: High-throughput candidate phenotype generation via tensor factorization

Joyce C. Ho [a,*], Joydeep Ghosh [a], Steve R. Steinhubl [b], Walter F. Stewart [c], Joshua C. Denny [d,e], Bradley A. Malin [d,f], Jimeng Sun [g]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, United States
[b] Scripps Translational Science Institute, Scripps Health, La Jolla, CA 92037, United States
[c] Sutter Health Research, Development, and Dissemination Team, Sutter Health, Walnut Creek, CA 94598, United States
[d] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, United States
[e] Department of Medicine, Vanderbilt University, Nashville, TN 37232, United States
[f] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37232, United States
[g] School of Computational Science and Engineering at College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, United States

ARTICLE INFO

Article history:
Received 13 November 2013
Accepted 2 July 2014
Available online 16 July 2014

Keywords:
Dimensionality reduction
Nonnegative tensor factorization
EHR phenotyping

ABSTRACT

The rapidly increasing availability of electronic health records (EHRs) from multiple heterogeneous sources has spearheaded the adoption of data-driven approaches for improved clinical research, decision making, prognosis, and patient management. Unfortunately, EHR data do not always directly and reliably map to medical concepts that clinical researchers need or use. Some recent studies have focused on EHR-derived phenotyping, which aims at mapping the EHR data to specific medical concepts; however, most of these approaches require labor intensive supervision from experienced clinical professionals. Furthermore, existing approaches are often disease-centric and specialized to the idiosyncrasies of the information technology and/or business practices of a single healthcare organization.

In this paper, we propose Limestone, a nonnegative tensor factorization method to derive phenotype candidates with virtually no human supervision. Limestone represents the data source interactions naturally using tensors (a generalization of matrices). In particular, we investigate the interaction of diagnoses and medications among patients. The resulting tensor factors are reported as phenotype candidates that automatically reveal patient clusters on specific diagnoses and medications. Using the proposed method, multiple phenotypes can be identified simultaneously from data.

We demonstrate the capability of Limestone on a cohort of 31,815 patient records from the Geisinger Health System. The dataset spans 7 years of longitudinal patient records and was initially constructed for a heart failure onset prediction study. Our experiments demonstrate the robustness, stability, and the conciseness of Limestone-derived phenotypes. Our results show that using only 40 phenotypes, we can outperform the original 640 features (169 diagnosis categories and 471 medication types) to achieve an area under the receiver operator characteristic curve (AUC) of 0.720 (95% CI 0.715 to 0.725). Moreover, in consultation with a medical expert, we confirmed 82% of the top 50 candidates automatically extracted by Limestone are clinically meaningful.

1. Introduction

The rapidly increasing availability of electronic health records (EHRs) from multiple heterogeneous sources has spearheaded the adoption of data-driven approaches for improved clinical decision making [1–4], prognosis [5–8], and patient management [9–12]. While knowledge discovery in EHRs show great promise towards providing better quality of care at lower costs [13–15], the vast information captured pose difficulties for both medical practitioners and data analysts [16]. EHR data offers many formidable challenges that has limited their utility for clinical research thus far. These include diverse patient populations from providers who may be using different, potentially incompatible EHR systems; heterogeneous information covering a variety of inter-related aspects of patients such as diagnoses, medication orders, and laboratory test findings [17,18]; sparsely sampled medical event sequences with different time scales across patients [19–21]; and noisy,

[*] Corresponding author.
E-mail address: joyceho@utexas.edu (J.C. Ho).

http://dx.doi.org/10.1016/j.jbi.2014.07.001
1532-0464/© 2014 Elsevier Inc. All rights reserved.

incomplete, and inaccurate representation of patients [22,23]. Clinical research requires precise and concise medical concepts about patients. The process of mapping raw EHR data into meaningful medical concepts, or the task of learning the medically relevant characteristics of the data [24,25] is referred to as EHR-based phenotyping. The phenotyping process can not only be used to identify specific clinical characteristics important in the process of research subject selection [26,27], but also improve the discovery process such as optimizing interventions and predicting response to therapy [24]. While the term EHR-based phenotyping has various meanings [28], this paper focuses primarily on the process of extracting medical concepts, or phenotypes.

Phenotypes encompass the entire spectrum of EHR data, using both structured information (e.g. billing codes, laboratory reports, and medication orders) and unstructured documents (e.g. clinical notes, pathology and radiology reports) [27,29]. Significant progress has been made in the generation and sharing of phenotypes [29–33]. Examples of such large-scale phenotyping efforts are typified by the Electronic Medical Records and Genomics (eMERGE) Network [34] and the Observational Medical Outcomes Partnership (OMOP) [35]. Furthermore, the eMERGE process supports portability via a process that iteratively tests and refines the phenotype at different institutions [29].

The development of EHR-derived phenotypes currently relies primarily on rule-based, heuristic and iterative based approaches, which take significant time and expert knowledge to develop [24,36,37]. Often, the phenotyping process requires a team effort from clinicians, domain experts, and IT experts [24,37,38]. However, phenotypes are often disease-centric and the development of a phenotype for a single disease can take months [39]. Furthermore, phenoytyping requires significant interaction between the domain experts and informaticians [37] and each team member may bring his/her own biases, ignoring potentially useful information [24]. Thus, high-throughput phenotyping, or efficient and automated phenotype extractions to reduce manual development, has gained recent attention [24,36,37,26]. Data mining and machine learning tools have been utilized to automate the phenotype generation process [24,36,37,26]. Yet, the current state of the art high-throughput phenotyping cannot generate large amounts of candidate phenotypes that simultaneously achieve good performance without human annotated samples [37]. Thus, the limitations of existing phenotyping efforts can be summarized as follows:

- A requirement for human annotation of case and control samples, taking substantial time, effort, and expert knowledge to develop.
- A lack of formalized methodology for deriving novel phenotypes such as disease subtypes.
- A failure to incorporate an automated process to support portability across institutions.

To create a high-throughput phenotyping environment, the phenotyping process needs to shift towards a more data-driven, high-throughput approach, where multiple candidate phenotypes are generated while minimizing human intervention [24]. Our paper directly addresses all but the last limitation by focusing on dimensionality reduction to automate the generation of phenotypes.

One possible approach to automatically discover phenotypes from EHR data is to use dimensionality reduction techniques [24], which represent the original data using lower dimensional latent space. Phenotyping takes high-dimensional EHR data and maps it to medical concepts, where an "ideal" phenotype (i) is concise and easily understood by a medical professional, (ii) represents complex interactions between several sources (e.g. diagnosis and medication), and (iii) maps to domain knowledge. Each phenotype can be viewed as the definition of a particular latent space along the multiple sources. Matrix factorization is a common dimensionality reduction approach in high-dimensional settings, but it may not concisely capture structured source interactions, such as multiple medications prescribed to treat a single disease. Thus, a more natural transformation is tensor factorization which utilizes the multiway structure to produce concise and more interpretable results.

This paper presents *Limestone*, a nonnegative tensor factorization method to generate phenotype candidates without expert supervision. Our algorithm is named after a sedimentary rock obtained via geology mining, the extraction of valuable resources from earth. Limestone (rock) has a wide diversity of uses and is an excellent building stone. We view our nonnegative tensor factorization model as a building block for high-throughput phenotyping from EHR data. Our proposed model:

1. Achieves high-throughput phenotyping by deriving multiple candidate phenotypes simultaneously from EHR data without any user supervision or domain expertise.
2. Captures data source interaction, such as the diagnosis and medication interaction from the same medical visit.
3. Generates concise and clinically meaningful phenotypes.
4. Produces stable phenotype definitions across multiple factorizations and small perturbations of the data.

We apply Limestone on real EHR data from Geisinger Health System. The case-control dataset contains 31,815 patients. We use our method to automatically derive multiple candidate phenotypes from the dataset and analyze the factors for stability, conciseness, predictive power, and clinical relevance. We also show that only 40 candidate phenotypes are needed as features to obtain better predictive accuracy of patients at risk of heart failure than the original set of medical features (640), achieving an area under the receiver operator characteristic curve (AUC) of 0.720 with a 95% confidence interval of (0.715, 0.725). Furthermore, 82% percent of the first 50 Limestone-derived phenotypes from the control population are confirmed by a medical expert to be clinically meaningful.

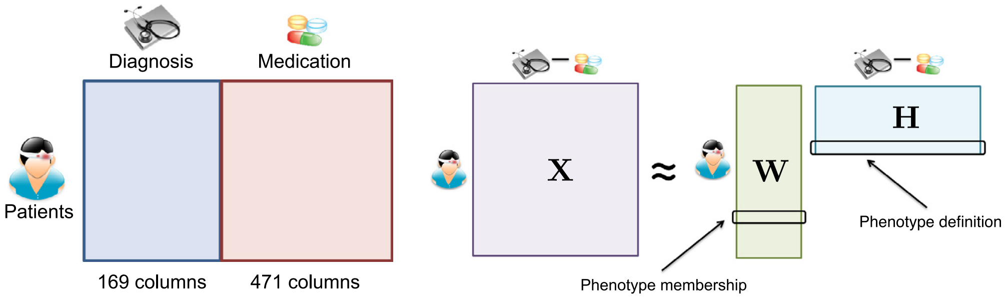
The remainder of the paper is structured as follows. Section 2 presents existing work on matrix factorization and summarizes relevant existing tensor factorization approaches. Next, we detail Limestone in Section 3. Section 4 demonstrates and evaluates our proposed method on real EHR data. This is followed by a discussion of the limitations and proposed future work in Section 5. Finally, we summarize our work in the Section 6.

2. Background and related work

Notation details. Table 1 provides a key to the symbols used in this paper. We adopt the notation from [40] to maintain consistency with the referenced tensor decomposition papers.

Table 1
List of notations used in this paper.

Symbol	Definition
$\mathbf{A}, \mathbf{B}, \boldsymbol{\Lambda}, \boldsymbol{\Pi}$	Matrix
$\mathbf{a}_r$	The rth column in matrix **A**
$\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}$	Tensor
$\vec{i}$	Tensor element index $(i_1, i_2, \ldots, i_N)$
$x_{\vec{i}}$	Tensor element at index $\vec{i}$
$\mathbf{X}_{(n)}$	Mode-n matricization of tensor $\boldsymbol{\mathcal{X}}$
$*$	Element-wise multiplication
$\circ$	Outer product
$\odot$	Khatri-Rao product
$\mathbf{A}^{\top}$	Transpose of **A**



(a) Source independent matrix (b) Standard matrix factorization

Fig. 1. EHR matrix representations and matrix factorization.

2.1. Matrix factorization: PCA and NMF

Structured EHR data can be represented using a feature matrix. The simplest representation for the data is a source independent feature matrix, where each row denotes a patient and each column represents a feature from a single source. Fig. 1a shows two matrices from a diagnosis source and a medication source. For context, in the Geisinger dataset, the diagnosis feature matrix contains 169 columns, where each column represents a single diagnosis such as asthma. However, the source independent feature matrix ignores potential interactions between the various sources, such as medications prescribed to treat a specific diagnosis. To incorporate "same visit" interactions,[1] a matrix whose column contains the combinations between the sources can be used. Fig. 1b illustrates a source interaction matrix for all diagnosis-medication combinations. This matrix introduces two problems: (1) the data is sparse because patients generally only experience a fraction of the diagnosis-medication combinations and (2) the data is high-dimensional (e.g. 149×471 possible combinations in the Geisinger case). Thus, dimensionality reduction can assist both interpretability and scalability of the data.

Matrix factorization (MF) is a common dimensionality reduction approach, which represents the original data using a lower dimensional latent space. Standard MF approaches, which focus primarily on numeric data, find two lower dimensional matrices such that when multiplied together approximately produce the original matrix. The mathematical formulation is as follows, given an $N \times M$ matrix $\mathbf{X}$, find matrices $\mathbf{W}$ and $\mathbf{H}$ of size $N \times R$ and $R \times M$ such that:

$$\mathbf{X} \approx \mathbf{WH}. \tag{1}$$

Fig. 1b illustrates the use of matrix factorization to derive phenotypes using the source interaction matrix. Although many matrix decomposition techniques exist [41], principal component analysis (PCA) and nonnegative matrix factorization (NMF) are two common algorithms used to reduce the feature dimension.

PCA calculates a set of basis vectors, or principal components, that minimizes the loss of information (i.e., the optimal approximation of the data in terms of least squared error). Generally, the number of principal components (R) is much smaller than the number of dimensions, which enables an encoding of the data as linear combinations of the basis vectors. Thus, PCA transforms the original, high-dimensional data to a lower-dimensional space defined by the principal components. One pitfall with PCA is the loss of "interpretability" which stems from several issues: (i) the principal components can have negative elements and (ii) the observed data can be approximated using both positive and negative combinations of the principal components. These are problematic because, in certain domains, negative elements and/or negative combinations are not easily interpretable [42,43]. For example, imagine the EHR feature matrix where each element represents the number of times a diagnosis or medication is recorded. Performing PCA on such a matrix results in a set of phenotypes, where each principal component defines the phenotype. A positive value in the principal component indicates the presence of a feature (diagnosis/medication) and a zero value denotes the absence. However, a negative entry does not readily map to some understanding about the feature's relationship to the phenotype.

The desire to prevent negative components motivated NMF [43]. Given a nonnegative matrix $\mathbf{X}$, the NMF finds two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ that approximate $\mathbf{X}$. Furthermore, the nonnegative constraint often leads to a sparse representation [43]. The enhanced semantic interpretability of NMF has led to its use across various fields such as mathematics, data mining, computer vision, and chemometrics [44]. Applications of NMF to biomedical data include discriminative feature selection from time–frequency representation of EEG data [45], feature extraction from brain CT images [46], and microarray gene data reduction for visualization and clustering purposes [47].

2.2. Tensor factorization

A tensor, or multiway array, is a generalization of a matrix (and a vector and a scalar) to higher dimensions. A *mode* of a tensor refers to a dimension, or way, of the tensor. The number of modes in a tensor is also known as the *order* of the tensor. Tensor representations are powerful because they can capture relationships for high-dimensional data. An overview of tensors can be found in [48–50].

A *rank-one* tensor can be written as the outer product of N vectors, where the outer product is defined as follows:

Definition 1. The outer product of N vectors, $\mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}$, produces a Nth order tensor $\boldsymbol{\mathcal{X}}$ where each element $x_{\bar{i}} = x_{i_1,i_2,\ldots,i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$.

Tensor factorization or decomposition is a natural extension of matrix factorization and utilizes information from the multiway structure that is lost when modes are collapsed to use matrix factorization algorithms [48,49,51,52]. One of the common tensor decompositions, CANDECOMP/PARAFAC (CP) [53,54], can be considered a higher-order generalization of singular value decomposition [48]. The CP model approximates the original tensor $\boldsymbol{\mathcal{X}}$ as a sum of R rank-one tensors:

$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}^{(1)}_r \circ \cdots \circ \mathbf{a}^{(N)}_r = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!].$$

[1] Note that we do not explicitly define "same visit", as what constitutes a same visit (e.g. a doctor visit, a hospital stay, etc) depends on the particular application.

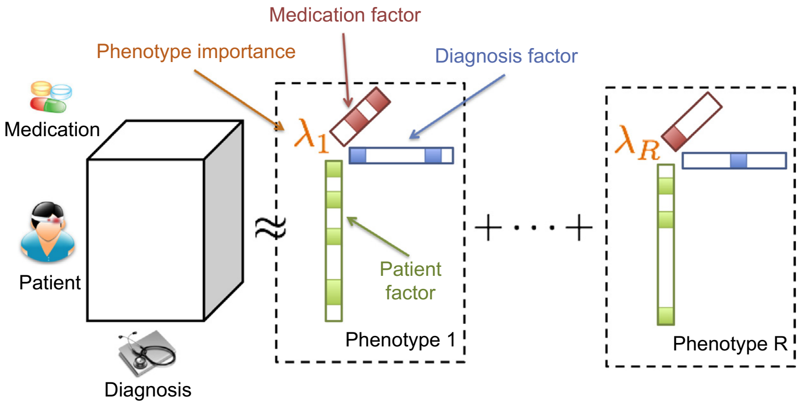


Fig. 2. Generating candidate phenotypes using CP tensor factorization.

Note that $[\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!]$ is shorthand notation to describe the CP decomposition, where $\boldsymbol{\lambda}$ is a vector of the weights λ_r and $\mathbf{a}_r^{(n)}$ is the rth column of $\mathbf{A}^{(n)}$. Fig. 2 conceptually illustrates the process of generating phenotypes via a CP decomposition. The details of our algorithm to generate concise phenotypes are presented in Section 3.2.

While several other tensor decomposition methods exist (Kolda and Bader provide a survey of existing models and example applications in their paper [48]), we focus on the CP decomposition for two primary reasons: (i) it is a well-known and commonly applied tensor factorization model [55], and (ii) the resulting structure (R rank-one tensors) is well-suited for capturing medical concepts in a concise and interpretable manner. The CP decomposition has been used to complete missing data in medical questionnaires [56], localize and extract artifacts from EEG data to analyze epileptic seizures [57,58], and as an exploratory decomposition tool for wavelet-transformed multi-channel EEG data [59].

Nonnegative tensor factorization (NTF) models have been proposed for CP decompositions. Analogous to NMF, NTF requires the elements of the factor matrices and the weights to be nonnegative. Some examples of NTF models in the medical and bioinformatics domain include the extraction of features from EEG data [60,61] and gene-sample-time microarray data [62]. Cichocki et al. provides a broad survey of practical and useful NMF and NTF algorithms [42].

The standard CP model is well-suited for continuous data, where the random variation follows a Gaussian distribution. However count data, which is nonnegative and discrete, is better described using a Poisson distribution [63]. The nonnegative CP alternating Poisson regression (CP-APR) model has been developed to fit count data [40]. We provide the CP-APR optimization problem formulation from [40] for convenience:

$$\begin{aligned} &\min f(\mathcal{M}) \equiv \underbrace{\sum\nolimits_{\vec{i}} m_{\vec{i}} - x_{\vec{i}} \log m_{\vec{i}}}_{\text{Kullback-Leibler(KL)divergence}} \\ &\text{subject to } \mathcal{M} = [\![\lambda; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!] \in \Omega \leftarrow \text{ sample space of } \mathcal{M} \\ &\qquad \Omega = \Omega_\lambda \times \Omega_1 \times \cdots \times \Omega_N \\ &\qquad \Omega_\lambda = [0, +\infty)^R \leftarrow \text{ weights are nonnegative} \\ &\qquad \Omega_n = \underbrace{\{\mathbf{A} \in [0,1]^{I_n \times R} \mid ||\mathbf{a}_r||_1 = 1 \forall r\}}_{\text{stochastic constraints on columns}}, \end{aligned} \tag{2}$$

where $\vec{i}$ represents the tensor element index $(i_1, i_2, \ldots, i_N)$, $\mathcal{X}$ is the observed tensor, and $\mathcal{M}$ is the CP tensor factorization that approximates $\mathcal{X}$.

The CP-APR algorithm solves the optimization problem via an alternating minimization approach, where each subproblem computes the solution for an individual mode while fixing all the other modes. CP-APR specifies the mode-n matricization as $\mathcal{X}$ as $\mathbf{X}_{(n)} = \mathbf{B}^{(n)}\mathbf{\Pi}^{(n)}$ [40], where

$$\begin{aligned} \text{Let}: \quad & \mathbf{B}^{(n)} = \mathbf{A}^{(n)} \\ & \boldsymbol{\lambda}\Pi^{(n)} = (\mathbf{A}^{(1)} \odot \ldots \odot \mathbf{A}^{(n-1)} \odot \mathbf{A}^{(n+1)} \odot \ldots \odot \mathbf{A}^{(N)})^\intercal. \end{aligned}$$

$\mathbf{B}^{(n)}$ represents the weighed nth mode factor matrix and $\Pi^{(n)}$ denotes the fixed part.[2] The CP-APR optimization subproblem (repeated from [40]) for the nth factor matrix is:

$$\mathbf{B}^{(n)} = \underset{\mathbf{B} \geqslant 0}{\arg\min}\, \mathbf{1}^\intercal [\mathbf{B}\Pi^{(n)} - \mathbf{X}_{(n)} * \log(\mathbf{B}\Pi^{(n)})]\mathbf{1}. \tag{3}$$

In Eq. (3), $\mathbf{1}$ corresponds to a vector of ones and captures the summation of the tensor elements shown in Eq. (2). The details of the subproblem solver and the overall CP-APR algorithm can be found in the paper [40].

NTF generally results in sparse representations. However, additional sparsity may be desired, for example, to improve factor interpretability. Various techniques have been used to induce sparsity, such as extending an NMF sparseness measure [64], enforcing L_1 penalties on the factor matrices and/or the core matrix for Tucker models [65–67], or regularizing the factors with both ℓ_1 and ℓ_2 norms [52].

3. Limestone: phenotype tensor factorization

Limestone is a tensor factorization model to achieve high-throughput phenotyping from EHR data. Our model extends the CP-APR work to (i) produce concise phenotype definitions for better interpretability and (ii) calculate a new patient's phenotype membership given the learned phenotypes. Fig. 3 illustrates the conceptual diagram for the Limestone process. This section details the tensor construction from raw EHR data, formally defines the candidate phenotypes obtained via tensor factorization, and the process to obtain the phenotype membership matrix for new patients.

3.1. EHR tensor construction

The first step in Limestone is to construct a count tensor from the raw EHR data. In this paper, we focus on diagnoses and medications due to their prominence in existing phenotype definitions [29,68]. However, our tensor construction can be generalized to other EHR data. We use medication orders from the raw EHR data that details the interaction between diagnoses and medications. Each medication order contains the prescribed medication, the

[2] The definition of the Khatri-Rao product is provided in the supplemental material.

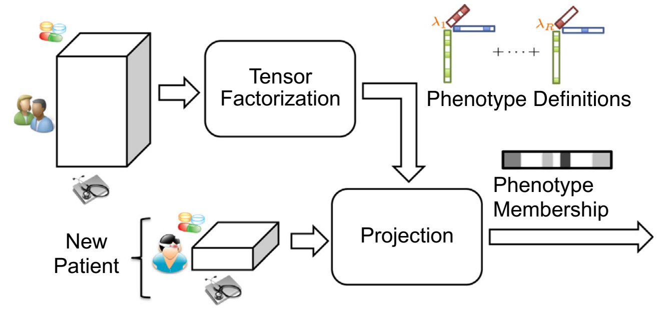


Fig. 3. A high-level depiction of the Limestone process by which candidate phenotypes are generated and patients are projected onto the candidates.

diagnosis (such as an ICD-9 billing code) associated with the prescription, and the date of the prescription.

Each patient is anchored using an index date (e.g. heart failure diagnosis date). The observation window is defined as a fixed time window of 2 years prior to the index date, as illustrated in Fig. 4. Only data occurring during the observation window is used for the raw EHR construction. The tensor is constructed using the count of the co-occurrences between medications and diagnoses. For Fig. 4, the patient has the following counts in the 2-year observation window encompassing 3 visits:

- 2 counts of loop diuretics to treat coronary atherosclerosis;
- 1 count of cardio-selective beta blockers to treat coronary atherosclerosis;
- 2 counts of sulfonylureas to treat diabetes;
- 1 count of nitrates to treat coronary atherosclerosis; and
- 1 count of ACE inhibitors to treat hypertension.

Note that the medication orders of sulfonylureas to treat diabetes at time t_0 and loop diuretics to treat congestive heart failure at time t_4 are outside the window and omitted from the tensor construction.

The result is a third-order tensor with a patient mode, diagnosis mode, and medication mode. Each tensor element denotes the number of times medication m is prescribed to treat diagnosis d for patient p. Slicing the tensor along the three different modes yields the following views:

1. Patient mode: a matrix of the patient's diagnoses and associated medication treatment.
2. Diagnosis mode: a matrix of the prescribed medications to treat the disease for all patients.
3. Medication mode: a matrix of all the patients and the diseases treated with this medication.

The count tensor is a more natural representation of the interactions between diagnoses and medications as it succinctly captures hierarchical information such as the set of medications that are used to treat a disease. In addition, the Limestone implementation only stores the non-zero elements of the tensor for efficient memory storage.

3.2. Phenotype definition via tensor factorization

Limestone extends the CP-APR model to derive phenotype candidates without supervision. The third-order count tensor is approximated using the CP decomposition $\mathcal{M} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}]\!]$, shown in Fig. 2. The factor matrix for the nth mode, $\mathbf{A}^{(n)}$, defines the elements from the mode that comprise the candidate phenotypes. Thus the jth candidate phenotype is defined using the jth column from the three factor matrices. Note that the stochasticity constraint (i.e., the last line in Eq. 2) on the factor matrix yields a conditional probability of the element's membership to the phenotype. Given the jth phenotype, $a_{ij}^{(k)}$ represents the probability of seeing the ith element in the kth mode. Thus, the sum of the entries for a mode element ($\sum_j a_{ij}^{(k)}$) across all the phenotypes may not equal 1. Furthermore, λ allows us to automatically rank the candidate phenotypes in order of significance, or the candidate phenotype's ability to capture the tensor data. Fig. 2 illustrates the tensor factorization of a patient by diagnosis by medication tensor into R phenotypes.

We provide an illustrative example of a candidate phenotype resulting from Limestone in Fig. 5. The percentage of patients with the phenotype is calculated using the percentage of non-zero elements in the kth column of the patient factor matrix. The phenotype is defined as patients diagnosed with hypertension and taking three medications: (1) beta blockers cardio-selective, (2) thiazides and thiazide-like diuretics, and (3) HMG CoA reductase inhibitors. Limestone produced a single non-zero element along the diagnosis factor and three non-zero components along the medication factor.

Our proposed model incorporates a sparsity constraint to minimize the presence of "minuscule and unnecessary" factor components. We extend the original CP-APR model by employing a hard-thresholding operator [69] to further reduce the phenotype

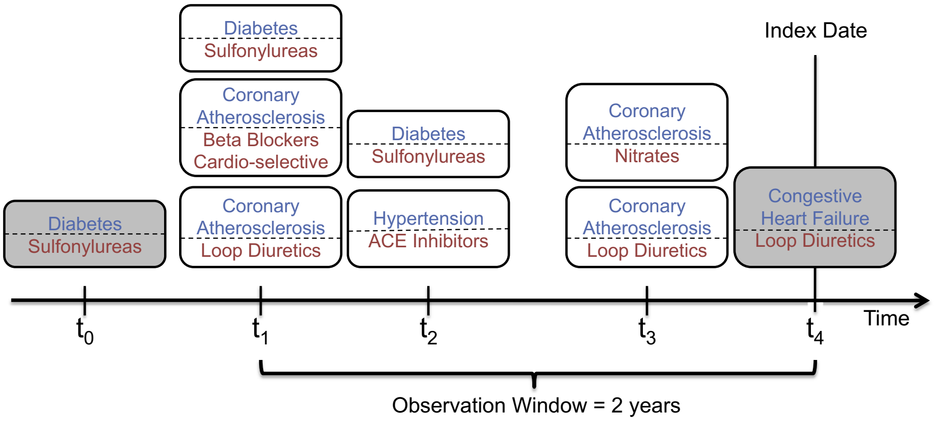


Fig. 4. The observation window is defined as a fixed time window prior to the index date (e.g. diagnosis date) and is used to determine the data used for tensor construction. The medication orders in gray are excluded during feature construction because they are outside the observation window.

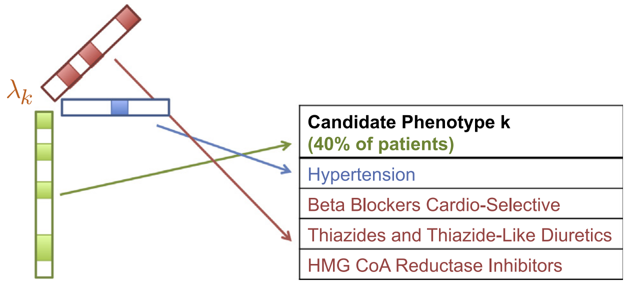

λ_k
Candidate Phenotype k
(40% of patients)
Hypertension
Beta Blockers Cardio-Selective
Thiazides and Thiazide-Like Diuretics
HMG CoA Reductase Inhibitors

Fig. 5. An example of the *k*th candidate phenotype produced from the tensor factorization, and the interpretation of the tensor factorization result. The green text, blue, and red text correspond to non-zero elements in the patient, diagnosis, and medication factors, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

factors by removing small factor components. Thus, Limestone minimizes KL divergence with a hard thresholding constraint, replacing Eq. (2) with the following objective:

$$\min \underbrace{\sum\nolimits_{\mathbf{i}}[m_{\mathbf{i}} - x_{\mathbf{i}} \log m_{\mathbf{i}}]}_{\text{CP-APR objective}} + \underbrace{\gamma \sum\nolimits_{j,n,r} \mathbb{1}_{\left\{a_{jr}^{(n)} > 0\right\}}}_{\text{hard-thresholding operator}}, \tag{4}$$

where $a_{jr}^{(n)}$ denotes the *j*th component of the factor vector $\mathbf{a}_r^{(n)}$. Individual components $a_{jr}^{(n)}$ that are below the threshold $\sqrt{2\gamma}$ are set to zero. Thus, the candidate phenotypes are concise, which should offer better interpretability.

3.3. Projection on candidate phenotypes

Limestone also computes a new patient's phenotype membership vector by projecting their observed features onto the space of existing candidate phenotypes. The phenotype membership vector $(\hat{\mathbf{a}})^1$ is defined as the convex combination of the candidate phenotypes, where the *r*th element of the vector, $\hat{a}_r^{(1)}$, is the probability the patient belongs to *r*th phenotype. For example, a new patient's vector may indicate probabilities of 0.6, 0.3, and 0.1 for the phenotypes of diabetes type 2, severe hypertension, and asthma, respectively. Note that the phenotype membership vector is not equivalent to the patient factor matrix, as the *r*th column of the patient factor matrix $\mathbf{A}^{(1)}$ represents a probabilistic interpretation over the entire patient population for a single phenotype.

Our method uses the diagnosis factor matrix $\mathbf{A}^{(2)}$ and the medication factor matrix $\mathbf{A}^{(3)}$ from the existing candidate phenotypes to calculate the phenotype membership vector. Thus, given a new patient's data, $\widehat{\mathcal{X}}$, we wish to find $\hat{\lambda}$ and $\hat{\mathbf{a}}^{(1)}$ that best approximates the new patient's tensor:

$$\widehat{\mathcal{X}} \approx \sum_r \underbrace{\hat{\lambda}_r \hat{a}_r^{(1)}}_{\text{membership}} \circ \overbrace{\mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)}}^{\text{phenotype definition}}$$

$$\text{s.t} \sum_r \hat{a}_r^{(1)} = 1.$$

The projection onto the candidate phenotypes is illustrated in Fig. 6. Therefore, the optimization for calculating the phenotype membership vector is

$$\hat{\mathbf{b}}^{(1)} = \underset{\mathbf{b} \geqslant 0}{\arg\min}\, \mathbf{1}^\intercal [\mathbf{b}\boldsymbol{\Pi}^{(1)} - \widehat{\mathbf{X}}_{(1)} * \log(\mathbf{b}\boldsymbol{\Pi}^{(1)})]\mathbf{1}$$

$$\text{s.t} \sum_r \hat{a}_r^{(1)} = 1,$$

where $\hat{\mathbf{b}}^{(1)} = (\hat{\mathbf{a}})^1 \hat{\boldsymbol{\Lambda}}$. The objective function of this problem is equivalent to the optimization subproblem for mode 1. Therefore, we can utilize the same iterative MM approach to solve for the optimal $\hat{\mathbf{b}}^{(1)}$.

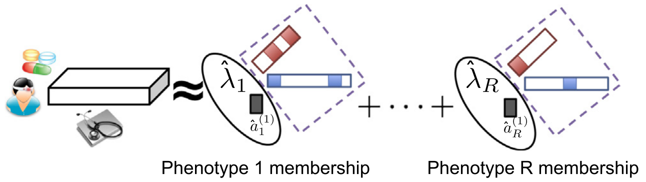


Fig. 6. A new patient's phenotype membership vector is computed by projecting the new patient's data onto the *R* candidate phenotypes in the purple dashed line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The new patient's phenotype membership vector $\hat{\mathbf{a}}^{(1)}$ is the entries of $\hat{\mathbf{b}}^{(1)}$ normalized by the weights $\hat{\boldsymbol{\lambda}}$.

Software implementation. We developed a Python package based on the Matlab Tensor Toolbox[3] and Pytensor,[4] a partial Python implementation of the Matlab Tensor Toolbox. Our software package implements the CP-APR algorithm described in [40] and provides the functions to post-process the tensor decomposition to obtain concise phenotypes and project new patients onto learned phenotypes.

4. Heart failure case study

Our case study focuses on heart failure (HF), a leading cause of healthcare use with a projected medical cost in 2015 of $32.5 billion [70]. Heart failure (HF) affects roughly 5.7 million people in the US and is mentioned as the contributing cause for 1 out of every 9 deaths [71]. Nearly a quarter of the patients hospitalized with heart failure are readmitted within 30 days [72]. Thus far, heart failure research has focused on epidemiology results, lifetime risk assessments from the Framingham study [73,74], predictions of hospital readmissions [75] or survival [76], and data-driven feature selection to complement known risk factors [77]. We demonstrate Limestone on a dataset primarily used for heart failure onset prediction studies and illustrate the potential of tensor factorization to derive candidate phenotypes without the supervision of domain experts. For this section, we will refer to candidate phenotypes (discovered clusters) as phenotypes for simplicity.

Evaluation Metric Details. Our case study focuses on algorithmic evaluation and qualitative analysis of Limestone-derived phenotypes. We will evaluate the results in terms of similarity, conciseness, predictive power, and clinically meaningfulness. The metrics we will use are the following:

1. Similarity$(\mathbf{a}_r, \mathbf{b}_r) = \frac{\mathbf{a}_r{}^\intercal \mathbf{b}_r}{\|\mathbf{a}_r\| \|\mathbf{b}_r\|}$.
2. Conciseness = number of non-zero elements per mode.
3. Predictive power = area under receiver operator characteristic curve (AUC) on a classification task.
4. Clinical meaningfulness = domain expert's opinion of whether or not a Limestone-derived phenotype mapped to a medical concept.

The similarity calculation is the cosine similarity between two vectors, a component of the factor match score (FMS). FMS is [40,78,79] commonly used to compare two tensor factorization results, quantifying the closeness via a single number between $[0, 1]$. However, FMS is an aggregate measure and can mask the mode-specific similarity results. Therefore, we compare the cosine along each mode, where the ideal value with two equivalent vectors is 1. Phenotypes from the two tensor factorization results are paired using an existing greedy FMS algorithm [40].

[3] http://www.sandia.gov/tgkolda/TensorToolbox/index-2.5.html.
[4] https://code.google.com/p/pytensor/.

4.1. Data description

The data for this study is based on real EHR data from the Geisinger Health system, which contains over 7 years of longitudinal patient records. The dataset has a diverse set of clinical information that includes diagnoses comprised of ICD-9 billing codes and medication records with generic drug names, pharmacy class and subclass information. For this study, we analyze the following sets of patients:

1. 4626 case patients, where each patient has at least 2 outpatient HF diagnoses or 1 outpatient HF diagnoses with 2 or more HF medications.
2. 27,189 group-matched control patients, where each case is matched with 10 controls with the same gender, age, and clinic information of the case patients.[5] The control patients did not meet the HF diagnosis criteria described above.

In the study, the heart failure index date for control patients is the date of the matched case patient (e.g. if the case patient was diagnosed on January 4, 2014, then the matched control patient would use January 4, 2014 as the index date). Further details of the cohort construction can be found in [23].

The Geisinger dataset recorded the interaction between diagnoses and medications in the medication orders table. Each medication order contains the prescribed medication, the diagnosis (ICD-9 billing code) associated with the prescription, and the date of the prescription. Any medication that was used to treat several diagnoses has multiple entries corresponding to each diagnosis code. The raw diagnosis code and medication captures information at a fairly fine-grained level, which is not ideal for analysis because similar diagnoses and medications are considered independently. To avoid this problem, we consolidated the individual diagnosis codes and medications to higher level concepts using existing medical hierarchies. Specifically, diagnosis codes are aggregated using the Centers for Medicare and Medicaid (CMS) Hierarchical Condition Categories (HCC) and medications defined as pharmacy subclass (e.g. ACE inhibitors, calcium channel blockers, etc.).[6] This resulted in 169 distinct HCC categories and 471 pharmacy subclasses. Therefore, the constructed tensor size for the control patients population is $27,189$ patients by 169 diseases by 471 medications, where $< 1\%$ of the tensor are non-zero.

4.2. Algorithmic evaluation

The first series of experiments focuses on evaluating the convergence, stability, computation time, and sparsity of Limestone. The following questions will be answered:

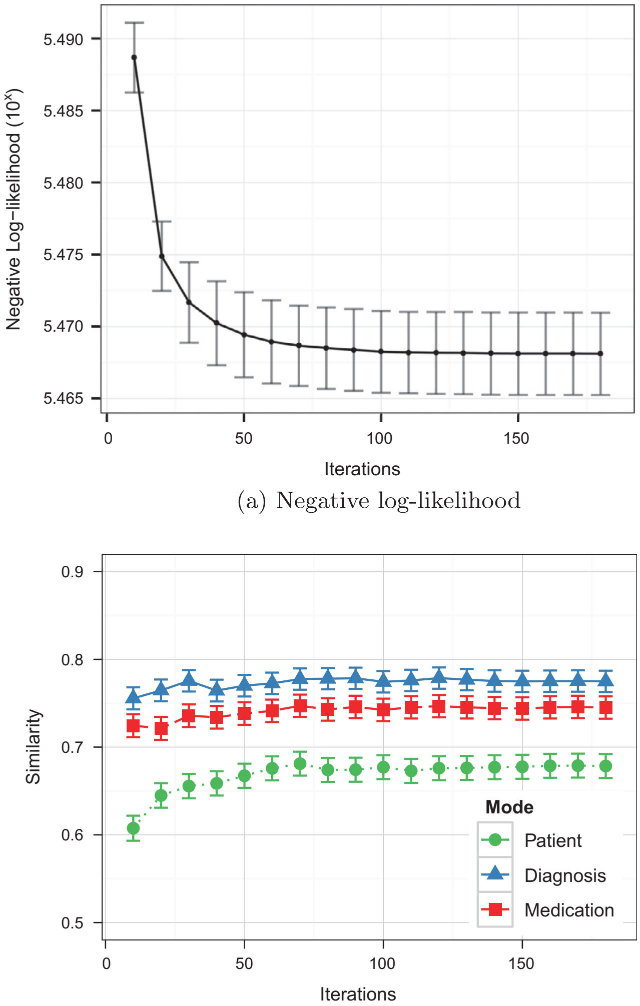
1. How many alternating minimization iterations are necessary to converge to a stable solution?
2. Are the generated phenotypes stable towards perturbation and different initializations?
3. How concise are the generated phenotypes?

4.2.1. Convergence

Given a fixed number of phenotypes (R), we examine the KL divergence (or the objective function values) as a function of the number of alternating minimization iterations across 10 randomly initialized factorizations of the case patients' tensor. The KL divergence is defined as $\sum_{\vec{i}} m_{\vec{i}} - x_{\vec{i}} \log m_{\vec{i}}$. Fig. 7a shows the mean and confidence interval of the objective function values as the number of iterations are increased. The first 30 iterations result in a significant decrease in the negative log-likelihood. Above 80 iterations, there are only slight changes in the objective function values with the values flattening around 120 iterations. The results suggest that less than 80 iterations are needed for convergence in this dataset.

4.2.2. Stability

Our algorithm uses random matrices for the initial factor matrices $\mathbf{A}^{(n)}$ which can have an impact on the solution of the tensor factorization. Thus, we study the effect of 10 random initializations of Limestone, factorizing the case patients tensor with a fixed number of phenotypes and varying number of maximum iterations. Fig. 7b illustrates the similarity score for each mode. The results show that the similarity scores are high across all three modes beyond 70 iterations. In particular, the diagnosis and medication modes have scores above 0.70. Note that the score of two random factors will tend towards 0. As such, in this case study, we can conclude that phenotype definitions are generally similar regardless of the initial factor matrices.



(a) Negative log-likelihood

(b) Similarity scores

Fig. 7. Objective function and similarity scores as a function of the number of total iterations for the case patients tensor. The error bars indicate the 95% confidence interval.

[5] Note that the same control patient may be matched by multiple cases. Thus, we post-process the controls to make sure each control patient is only matched with one case. The 19,071 duplicate controls are removed from the dataset.

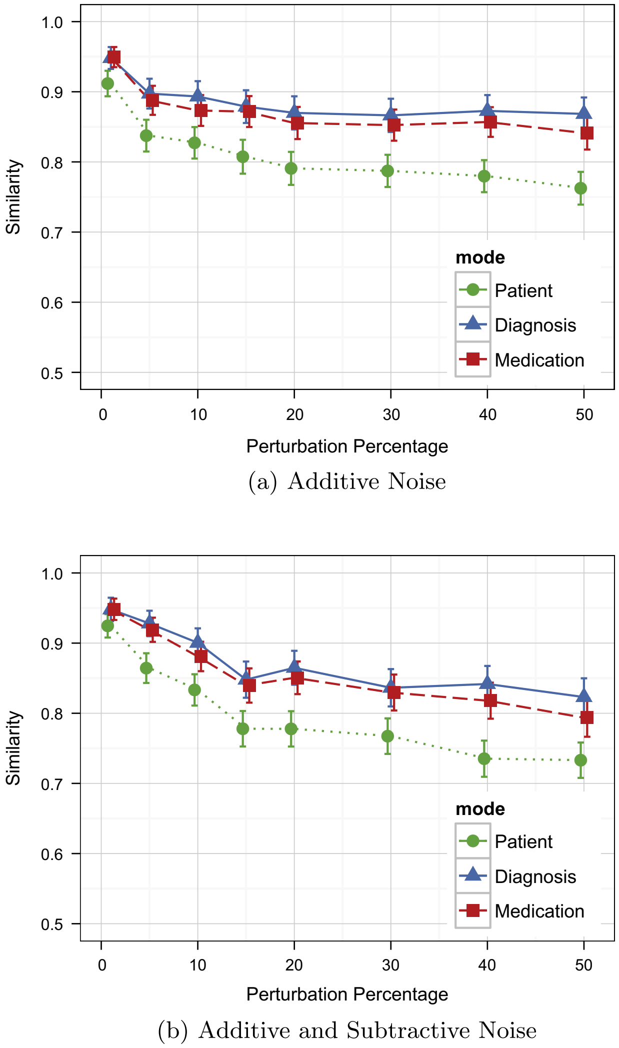
[6] Note that other hierarchies, such as the PheWAS code groups [80] could have been employed.

We also study the effect of noise, or perturbation, on the tensor factorization results. Two experiments were performed:

1. Additive noise: Poisson noise ($\epsilon \sim \text{Poisson}(2)$) is added to randomly selected non-zero elements of original tensor, increasing the overall mean of the tensor.
2. Additive and subtractive noise: Random addition or subtraction of Poisson noise ($\epsilon \sim \text{Poisson}(2)$) to randomly selected non-zero elements of the original tensor. If subtraction results in a negative value, the value is set to zero and a random zero element of the original tensor is selected for added noise to maintain the overall mean and sparsity pattern of the original tensor.

The resulting "noised" tensor is then factorized and compared to the original factorization using the similarity score.

Figs. 8a and b illustrates the average similarity scores for 10 random noisy tensors as a function of the percentage of noised elements. The results show a decay in the similarity score as the percentage of perturbed elements increases, where the effect is more prominent in the additive and subtract noise results. However, even when half of the non-zero elements are perturbed for both experiments, the diagnosis and medication mode similarities remain above 0.75, an impressive number given the high dimensionality of our dataset. This observation suggests that phenotype definitions are stable with regards to perturbation.



(a) Additive Noise

(b) Additive and Subtractive Noise

Fig. 8. Similarity scores to the original tensor factorization results for perturbed versions of the case patient tensor.

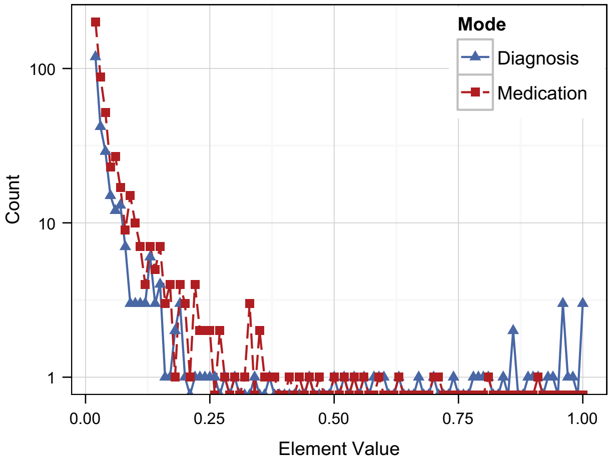


Fig. 9. The distribution of non-zero element values for 50 Limestone-derived phenotypes.

4.3. Sparsity

Limestone uses a hard thresholding operator which enables a tunable parameter to adjust the sparsity of the phenotypes. Fig. 9 shows a graph of the individual mode component values for the diagnosis and medication modes for a case patients tensor factorization. A majority of the nonzero elements in the diagnosis and medication factor matrices are below 0.05 (the two points furthest left in the plot). However, a reasonable number of the components along the diagnosis factor have values above 0.75, while the medication factors tend to have several medications (centered closer to 0.20). Thus, individual components less than a threshold of 0.05 contribute minimally to the phenotype definition in comparison with the other non-zero elements and can be triaged to produce concise phenotypes.

Fig. 10 shows the number of non-zero entries for the diagnosis and medication factors using the suggested threshold from above. Twelve of the phenotype were defined using a single diagnosis. A majority of the phenotype definitions contained less than five medications. Thus, in this case study, Limestone produces concise phenotypes at the threshold of 0.05, where all phenotypes contain less than eight non-zero elements per factor.

4.4. Performance evaluation

The next series of experiments evaluate the Limestone-derived phenotypes compared against the traditional dimensionality

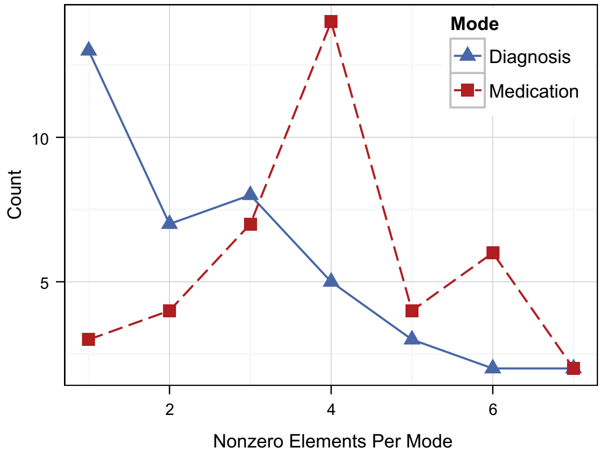


Fig. 10. The number of non-zero entries per factor using a threshold of 0.05.

Limestone Phenotype	
Hypertension	0.94
Hypertensive Heart Disease	0.06
Beta Blockers Cardio-Selective	0.51
Calcium Channel Blockers	0.32
Diuretic Combinations	0.06
Nitrates	0.06
HMG CoA Reductase Inhibitors	0.06
Vasodilators	0.05

NMF Phenotype	
Hypertension – Sympathomimetics	0.0032
Hypertension – Insulin	0.0027
Hypertension – Potassium	0.0018
Hypertension – Beta Blockers Cardio-Selective	0.0004
Hypertension – HMG CoA Reductase Inhibitors	0.0003
Major Symptoms, Abnormalities – Sympathomimetics	0.0167
Major Symptoms, Abnormalities – Insulin	0.0143
Major Symptoms, Abnormalities – Sodium	0.0133
Major Symptoms, Abnormalities – Potassium	0.0097
Major Symptoms, Abnormalities – Coumarin Anticoagulants	0.0092
Vascular Disease – Sympathomimetics	0.0068
Other Gastrointestinal Disorders – Sympathomimetics	0.0065
Other Endocrine/Metabolic/Nutritional Disorders – Sympathomimetics	0.0062
History of Disease – Sympathomimetics	0.0041
Other Dermatological Disorders – Sympathomimetics	0.0040
Other Infectious Diseases – Sympathomimetics	0.0039
… 1,549 total combinations	

Fig. 11. The most significant Limestone-derived phenotype and a "similar" NMF-derived phenotype with several matching diagnosis and medications. The Limestone features are listed in descending order of the probabilistic values. The similar NMF features are listed first, before listing the features in descending order based on element value. The NMF threshold was adjusted to 0.001 to maintain similarities with the Limestone-derived phenotype.

reduction approaches. In particular, we focus on the following two questions:

1. How do the Limestone-derived phenotypes compare to the phenotypes generated using nonnegative matrix factorization?
2. Do the phenotypes contain as much predictive power as traditional dimensionality reduction approaches?

4.4.1. NMF comparison

First, we compare the Limestone-derived phenotypes against the traditional NMF approach.[7] NMF is performed on the mode-1 matricization of the case patients tensor ($\mathbf{X}_{(1)}$), a diagnosis-medication source interaction matrix.

Fig. 11 shows an example of the highest weighted (largest λ) Limestone-derived phenotype and the most similar NMF-derived phenotype according to the cosine similarity score. For interpretability purposes, only the non-zero diagnosis-medication combinations with the largest weights are presented for the NMF-derived phenotype. The phenotype definition is comprised of 1,549 diagnosis-medication combinations. The Limestone-derived phenotype is concise and easier to interpret thanks to the structure of the definition. The NMF-derived phenotype also illustrates the benefit of tensor factorization, in that as several medications are shared across various diseases (e.g. sympathomimetics prescribed to treat hypertension and vascular disease).

4.4.2. Predictive power

Limestone-derived phenotypes are evaluated on a classification task of predicting heart failure patients and compared against three other feature sets. We use 10 random splits of the data, where each split divides the patient population evenly (50% train – 50% test) and maintains the same disease prevalence (otherwise known as stratified sampling). The feature sets are then generated for each split of the dataset:

1. Baseline: matrix with 640 columns (features), where 169 correspond to the different diagnoses and the remainder represent medications. This feature set ignores any potential interaction between diagnoses and medications.
2. PCA membership matrix: PCA is performed on the mode-1 matricization of the observed tensor (the source interaction matrix 169 × 471 columns representing each diagnosis medication combination) with only the training population to obtain the phenotype definitions, or $\mathbf{H}$ in Eq. (1). The PCA membership matrix ($\mathbf{W}$) is then computed for the entire population using the phenotype definitions from the training population.
3. NMF membership matrix: Similar to the calculation for the PCA membership matrix, with the exception that $\mathbf{H}$ and $\mathbf{W}$ are nonnegative.
4. Limestone membership matrix: Limestone generates phenotypes from the observed tensor (only patients in the training set) and then projects all the patients onto the learned phenotypes.

Note that for PCA, NMF, and Limestone, only the patient phenotype matrix (R columns) is used as the features to the predictive model while the baseline uses all 640 columns. A ℓ_1 regularized (Lasso) logistic regression predictive model is trained separately on each of the four feature sets and the model's predictive performance is evaluated on the test set.[8]

Fig. 12 displays a plot of the area under the receiver operating characteristic curve (AUC) while varying the number of phenotypes in the data. All three methods have a significant improvement over the baseline. Moreover, the phenotyping methods provide 20X feature reduction by only using 30 phenotyping features to outperform the baseline using 640 features. Limestone and NMF-derived phenotypes consistently achieve the highest predictive performance compared to PCA, especially above 30 phenotypes. The results show that using only 40 phenotypes, we achieve an AUC of 0.720 with a 95% confidence interval of (0.715, 0.725).

Table 2a shows the mean and median number of non-zero elements per phenotype for PCA, NMF, and Limestone. For comparison purposes, Limestone phenotypes have been converted to the diagnosis-medication representation ($\mathbf{A}^{(2)} \odot \mathbf{A}^{(3)}$). Thus, the results for Limestone is equivalent to the number of non-zero diagnoses elements multiplied by the number of non-zero medication components. Limestone yields concise phenotypes compared to the other two methods and provides a 94.7% reduction on the number of non-zero elements in comparison to NMF.

[7] PCA-derived phenotypes are not considered because negative elements lack a clear clinical interpretation.

[8] The ℓ_1 regularization term performs phenotype selection.

Table 2b summarizes the average computation time (for the 10 random samples) for the three methods. PCA has the smallest computation time because it has a closed form solution, while our algorithm takes the longest. Thus, these results suggest that Limestone can produce concise phenotypes while maintaining similar computational complexity to NMF.

4.5. Domain expert evaluation

We now evaluate the clinical quality of the phenotypes. Our experiment is meant to be a pilot study (a proof of concept) rather than a formal survey designed to detect significance. An experienced cardiologist was provided with 50 candidate phenotypes derived from the control patients and 50 candidate phenotypes derived from the case patients. The phenotypes were ordered by decreasing phenotype importance (λ) and the diagnosis and medication factors were thresholded at 0.01.[9] In addition, we presented the expert with the percentage of patients that belonged to this phenotype. The percentage value was calculated by counting the number of patients in the patient factor that had a value greater than 1^{-10} and dividing by the total number of control patients. The medical expert answered the following questions with a yes, no, or possible for each phenotype: “Are the combinations of diagnosis and medications clinically meaningful?” We also asked the expert to annotate each individual diagnosis and medication regarding its meaningfulness to the phenotype and assign a short label for the meaningful phenotypes.

Although the medical expert provided an overall analysis of each phenotype, in several cases the response was different compared to the individual annotations. Thus, the individual medication and diagnosis annotations were combined to produce an overall score for the phenotype. We used a “lowest common denominator” approach, such that if any element was tagged with a “no” the phenotype would have an overall score of No. Generally, if a phenotype contained a mixture of “possible” and ”yes” annotations, the phenotype was deemed possibly meaningful. The exception occurs when a single medication from the list is marked as possibly meaningful while the remaining diagnoses and medications are marked as yes, the phenotype was tagged as a yes. Table 3 summarizes the aggregated answers for Limestone-derived phenotypes from both the control patients’ tensor and the case patients’ tensor. A high percentage of the control phenotypes, 41 of the 50 (82%), were deemed clinically meaningful. Furthermore, only 3 of the phenotypes were not clinically meaningful. The clinical meaningfulness of the case patients derived phenotypes was not as high as the control set, however only 14 of the 50 case phenotypes (28%) were not clinically meaningful.

We first focus on the phenotypes derived from the control patient tensor. The five most significant (largest λ) control phenotypes are shown in Fig. 13. All but the second phenotype were annotated as clinically meaningful (second was annotated as possibly meaningful) and the expert-assigned short label is displayed in the figure. From the figure, four of the first five Limestone-derived phenotypes consist of a single diagnosis and a handful of medications.

The experimental results also suggest the potential ability to capture disease subtypes. Fig. 14 shows the meaningful phenotypes relating to hypertension derived from the control patients tensor. All three of the phenotypes share the same disease, but have different combinations of medications which may indicate disease severity. The domain expert assigned the following labels for the three candidates: the fourth phenotype corresponds to patients with mild hypertension; the second phenotype is moderate hypertension; and the sixth phenotype is the most severe of the phenotypes.

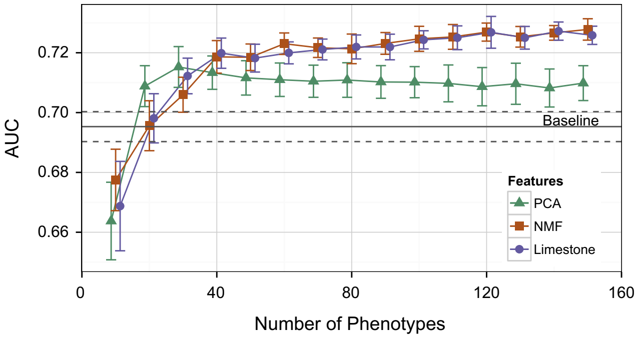


Fig. 12. Area under the receiver operating characteristic curve for the four feature sets while varying the number of phenotypes. The error bars denote the 95% confidence interval and the dashed lines illustrated the confidence interval using the baseline feature set.

Table 2
Average statistics from the 10 random splits using 50 phenotypes.

Type	Mean	Median
(a) Non-zero elements per phenotype		
PCA	10917.50	10921.50
NMF	1533.62	1504.50
Limestone	34.79	32.00
Method		Time (h)
(b) Computation time		
PCA		0.001
NMF		1.648
Limestone		2.366

Table 3
Expert annotation of 50 Limestone-derived phenotypes from the case and control tensors.

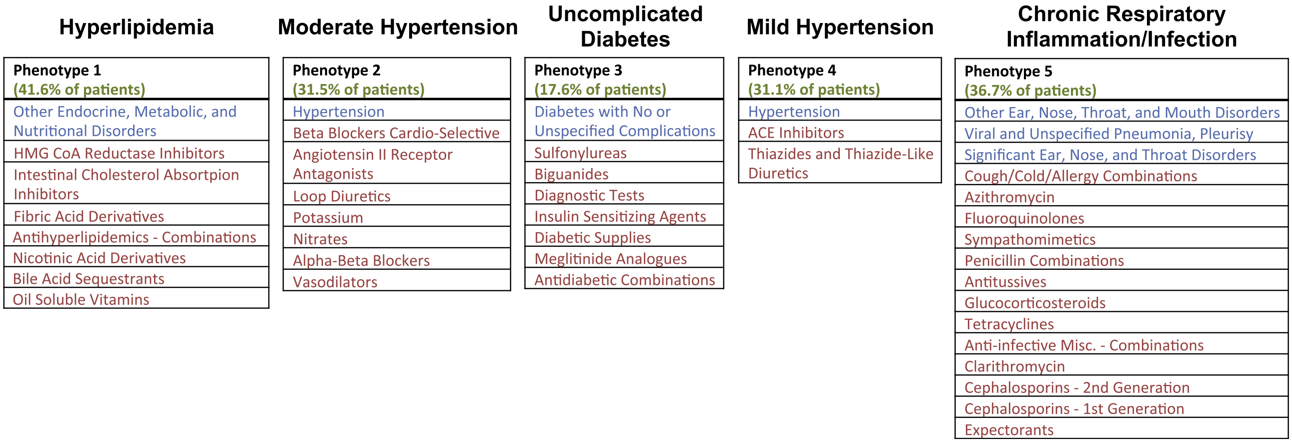
Tensor	No	Possible	Yes
Case	14	14	22
Control	3	6	41

For the six phenotypes labeled as possibly meaningful, four of them contained the HCC category “major symptoms and abnormalities.” The ambiguous diagnosis made it difficult for the medical expert to determine the exact medical concept. An example of the comment for these four phenotypes is “Yes, but with a diagnosis of major symptoms, abnormality, it can mean anything.” The broader class of control patients may have resulted in ambiguous diagnoses for several of the phenotypes.

Only three of the control phenotypes were labeled as not clinically meaningful. The predominant reason was the lack of a cohesive diagnosis factor. One phenotype contained the single diagnosis of “major symptoms and abnormalities”, while the other two had over ten unrelated diagnoses.

We next focus on the medical expert’s annotations of the phenotypes derived from the case patients tensor. Generally, the phenotypes labeled as clinically meaningful share the same medical concepts as those in the clinically meaningful control patients derived phenotypes. The differences in the tensor factorization results occur for the phenotypes marked as possibly meaningful or not clinically meaningful. For the fourteen phenotypes annotated as possibly meaningful, the expert’s comment was “Looks good except <insert a diagnosis> is meaningless.” Thus, 36 of

[9] Extra elements were included to present more information to the medical expert at the cost of less concise phenotype definitions.



Hyperlipidemia

Phenotype 1
(41.6% of patients)
Other Endocrine, Metabolic, and Nutritional Disorders
HMG CoA Reductase Inhibitors
Intestinal Cholesterol Absorption Inhibitors
Fibric Acid Derivatives
Antihyperlipidemics - Combinations
Nicotinic Acid Derivatives
Bile Acid Sequestrants
Oil Soluble Vitamins

Moderate Hypertension

Phenotype 2
(31.5% of patients)
Hypertension
Beta Blockers Cardio-Selective
Angiotensin II Receptor Antagonists
Loop Diuretics
Potassium
Nitrates
Alpha-Beta Blockers
Vasodilators

Uncomplicated Diabetes

Phenotype 3
(17.6% of patients)
Diabetes with No or Unspecified Complications
Sulfonylureas
Biguanides
Diagnostic Tests
Insulin Sensitizing Agents
Diabetic Supplies
Meglitinide Analogues
Antidiabetic Combinations

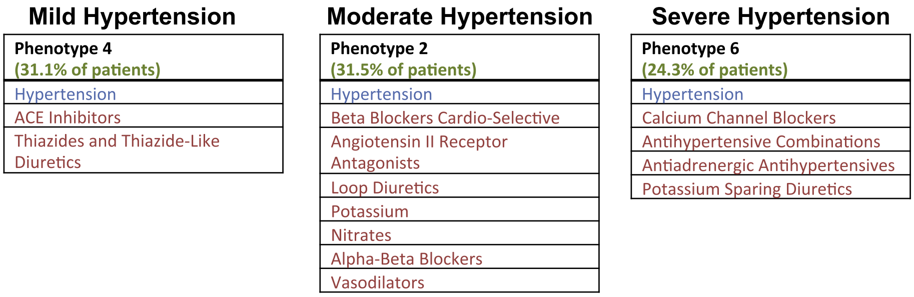
Mild Hypertension

Phenotype 4
(31.1% of patients)
Hypertension
ACE Inhibitors
Thiazides and Thiazide-Like Diuretics

Chronic Respiratory Inflammation/Infection

Phenotype 5
(36.7% of patients)
Other Ear, Nose, Throat, and Mouth Disorders
Viral and Unspecified Pneumonia, Pleurisy
Significant Ear, Nose, and Throat Disorders
Cough/Cold/Allergy Combinations
Azithromycin
Fluoroquinolones
Sympathomimetics
Penicillin Combinations
Antitussives
Glucocorticosteroids
Tetracyclines
Anti-infective Misc. - Combinations
Clarithromycin
Cephalosporins - 2nd Generation
Cephalosporins - 1st Generation
Expectorants

Fig. 13. The top five Limestone-derived phenotypes using the control patients' tensor.



Mild Hypertension

Phenotype 4
(31.1% of patients)
Hypertension
ACE Inhibitors
Thiazides and Thiazide-Like Diuretics

Moderate Hypertension

Phenotype 2
(31.5% of patients)
Hypertension
Beta Blockers Cardio-Selective
Angiotensin II Receptor Antagonists
Loop Diuretics
Potassium
Nitrates
Alpha-Beta Blockers
Vasodilators

Severe Hypertension

Phenotype 6
(24.3% of patients)
Hypertension
Calcium Channel Blockers
Antihypertensive Combinations
Antiadrenergic Antihypertensives
Potassium Sparing Diuretics

Fig. 14. Limestone-derived phenotypes from the control patients' tensor relating to hypertension.

the 50 case phenotypes, or 72% of the phenotypes, generally maps to a clinical concept.

The remaining 14 case phenotypes were labeled as not clinically meaningful. The medical expert's comments for these phenotypes were similar to the control phenotypes that lacked clinical meaning. Phenotypes either had too many unrelated diagnoses or uninformative diagnoses elements such as "history of disease", "minor symptoms, signs, findings", and "major symptoms and abnormalities."

5. Discussion

Our proposed method can identify multiple candidate phenotypes concurrently from EHR data without any user supervision. However, there remain several challenges towards its application in a high-throughput setting. First, one of the most challenging and unanswered tensor factorization questions relate to the choice of rank (the number of phenotype candidates) [40]. A small number of phenotypes may result in broad phenotype definitions while a large number of phenotypes may result in "over-specificity" in the phenotype definitions. Our heart failure case study used 50 as the number of phenotypes to minimize the burden of the annotation process while also illustrating the potential to achieve high-throughput phenotyping. One possible option is to select the number of phenotypes based on the performance of the candidate phenotypes when used for subsequent predictive modeling tasks, but discovering the best strategy will require further investigation.

The second challenge is that our clinical evaluation of Limestone relied on a single medical expert to answer a question regarding the clinical meaningfulness of the phenotypes. The clinical evaluation was designed to be a proof of concept with the knowledge that a potential bias can influence our results. To demonstrate statistical significance, it will be necessary to design a more extensive study that involves a panel of experts and asks various questions relating to the tensor derived phenotypes.

A third limitation of Limestone is that not all of the candidate phenotypes were clinically meaningful. One potential explanation is that the phenotypes labeled as possibly meaningful allude to the potential of our method for knowledge discovery, generating novel phenotypes that are currently unknown medical concepts. Moreover, the higher number of clinically meaningful phenotypes from the control population suggests that 50 phenotypes may not be present in the case population. Further exploration of the results in conjunction with a panel of experts are necessary to better understand the differences between the case and control populations.

Even though the candidate phenotypes generally mapped to a medical concept, the medical expert's annotations suggest the need for incorporating user feedback to refine phenotypes. Future work can improve the Limestone process by using tensor factorization to generate multiple candidate phenotypes from the observed data and then present the phenotypes to domain experts. The medical experts can then approve, reject, or alter the phenotypes such that all phenotypes are clinically meaningful. Although the phenotype generation process requires some human intervention, the goal would then be to minimize the interaction time necessary to produce meaningful phenotypes. Furthermore, existing phenotypes can be utilized to avoid repeated derivation of the same definitions.

Fourth, the clinical evaluation results suggest the potential ability to generate phenotypes that capture disease subtypes under an unsupervised setting. For instance, the three candidate phenotypes shown in Fig. 14 captured differing disease severities in the control population. Nonetheless, further analysis is necessary to determine whether the candidate phenotypes reflect the true patient status. Moreover, future work should analyze the efficiency of Limestone to capture disease subtypes both in this dataset and other EHR datasets.

Fifth, although our paper only focused on the tensor constructed with diagnoses and medications, the EHR tensor can be constructed using various other structured data sources. Preliminary experiments using several other data sources (though outside the scope of this specific study) such as laboratory tests, imaging results, and patient symptoms yielded similar results in terms of conciseness and predictive power. For unstructured sources, such as clinical notes, Limestone will require an additional preprocessing step (e.g. text mining or natural language processing). However, our current methodology only supports a single tensor. EHR data is comprised of multiple sources which may not naturally fit into a single tensor representation. Therefore, Limestone will need to be extended to factorize multiple tensors to fully utilize all EHR data.

Finally, our proposed method does not address portability across institutions. Candidate phenotypes generated at one site may be somewhat different from candidate phenotypes generated at another site. Thus, Limestone-derived phenotypes may not be readily transportable and executed at various other institutions. Future work should focus on the portability of Limestone-derived phenotypes while allowing variations in the phenotype definition.

6. Conclusion

This paper introduced Limestone, a nonnegative tensor factorization method to generate phenotypes without supervision. Limestone can generate numerous phenotypes simultaneously from data with minimal human intervention. The resulting tensor factors serve as phenotype candidates that automatically reveal patient clusters on specific diagnoses and medications. Moreover, our method can derive concise phenotype definitions, potentially capture disease subtypes that may not otherwise be easily defined, and produce consistent phenotype definitions for multiple factorizations of the same data.

Our results on 31,815 patient records from Geisinger Health System demonstrate the stability, conciseness, predictive power, and clinical meaningfulness of Limestone-derived phenotypes. They underscore the promise of Limestone for high-throughput phenotyping that generally results in meaningful phenotypes. Future work will focus on incorporating domain expertise in the phenotype generation process and extending the methodology to factorize multiple tensors simultaneously.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This research is supported by the Schlumberger Centennial Chair in Engineering and Army Research Office, under Grant W911NF-11-1-0258.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2014.07.001.

References

[1] Davis J, Lantz E, Page D, Struyf J, Peissig P, Vidaillet H, et al. Machine learning for personalized medicine: will this drug give me a heart attack. In: ICML workshop on machine learning for health care applications; 2008.

[2] Ramakrishnan N, Hanauer D, Keller B. Mining electronic health records. Computer 2010;43:77–81.

[3] Koh HC, Tan G. Data mining applications in healthcare. J Healthcare Inform Manage 2005;19:64–72.

[4] Wagholikar KBK, Maclaughlin KLK, Henry MRM, Greenes RAR, Hankey RAR, Liu HH, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19:833–9.

[5] Davis D, Chawla N, Christakis NA, Barabási A. Time to CARE: a collaborative engine for practical disease prediction. Data Min Knowl Discov 2010;20:388–415.

[6] Savage N. Better medicine through machine learning. Commun ACM 2012;55:17–9.

[7] Morillo DS, León Jiménez A, Moreno SA. Computer-aided diagnosis of pneumonia in patients with chronic obstructive pulmonary disease. J Am Med Inform Assoc 2013;20:e111–7.

[8] Sajda P. Machine learning for detection and diagnosis of disease. Annu Rev Biomed Eng 2006;8:537–65.

[9] Nguyen HB, Corbett SW, Steele R, Banta J, Clark RT, Hayes SR, et al. Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. Crit Care Med 2007;35:1105–12.

[10] Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Translat Med 2010;2. 48ra65-48ra65.

[11] Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306:1688–98.

[12] Lee N, Laine AF, Hu J, Wang F, Sun J, Ebadollahi S. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. In: First IEEE international conference on healthcare informatics, imaging and systems biology. IEEE; 2011. p. 250–7.

[13] West SL, Blake C, Liu Z, McKoy JN, Oertel MD, Carey TS. Reflections on the use of electronic health record data for clinical research. Health Inform J 2009;15:108–21.

[14] Westra BL, Dey S, Fang G, Steinbach M, Kumar V, Oancea C, et al. Interpretable predictive models for knowledge discovery from home-care electronic health records. J Healthcare Eng 2011;2:55–74.

[15] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev: Genet 2012;13:395–405.

[16] Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc 2012;19:684–7.

[17] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1–24.

[18] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3–23.

[19] Lin J-H, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 2008;41:1–14.

[20] Wang TD, Plaisant C, Quinn AJ, Stanchak R, Murphy S, Shneiderman B. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI); 2008. p. 457–66.

[21] Wang F, Lee N, Hu J, Sun J, Ebadollahi S. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In: Proceeding of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD); 2012.

[22] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. In: AMIA summits on translational science proceedings 2010; 2010. p. 1–5.

[23] Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48:S106–13.

[24] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2012;20:117–21.

[25] Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013;20:e226–31. http://dx.doi.org/10.1136/amiajnl-2013-001926.

[26] Wei W-Q, Tao C, Jiang G, Chute CG. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes. In: AMIA annual symposium proceedings 2010; 2010. p. 857–61.

[27] Denny JC. Mining electronic health records in the genomics era. PLoS Comput Biol 2012;8. e1002823-e1002823.

[28] Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc: JAMIA 2013.

[29] Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20:e147–54.

[30] Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and

controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc 2011;18:376–86.

[31] Conway M, Berg RL, Carrell D, Denny JC, Kullo IJ, Kho AN, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In: AMIA annual symposium proceedings 2011; 2011. p. 274–83.

[32] Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Translat Med 2011;3:79re1.

[33] Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc 2013;20:e243–52. http://dx.doi.org/10.1136/amiajnl-2013-001930.

[34] McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genom 2011;4:13.

[35] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2011;19:54–60.

[36] Carroll RJ, Eyler AE, Denny JC, Naïve electronic health record phenotype identification for rheumatoid arthritis. In: AMIA annual symposium proceedings 2011; 2011. p. 189–96.

[37] Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc 2013;20(e2):e253–9. http://dx.doi.org/10.1136/amiajnl-2013-001945.

[38] Li D, Endle CM, Murthy S, Stancl C, Suesse D, Sottara D, et al. Modeling and executing electronic health records driven phenotyping algorithms using the NQF quality data model and JBoss® drools engine. In: AMIA annual symposium proceedings 2012; 2012. p. 532–41.

[39] Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc 2013;20(e2):e311–8. http://dx.doi.org/10.1136/amiajnl-2013-001922.

[40] Chi EC, Kolda TG. On tensors, sparsity, and nonnegative factorizations. SIAM J Matrix Anal Appl 2012;33:1272–99.

[41] Singh AP, Gordon GJ. A unified view of matrix factorization models. In: ECML PKDD '08: proceedings of the European conference on machine learning and knowledge discovery in databases – Part II. Springer-Verlag; 2008.

[42] Cichocki A, Zdunek R, Phan AH, Amari S-I. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. Wiley; 2009.

[43] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.

[44] Wang Y-X, Zhang Y-J. Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng 2013;25:1336–53.

[45] Lee H, Cichocki A, Choi S. Nonnegative matrix factorization for motor imagery EEG classification. In: Proceedings of the 16th international conference on artificial neural networks (ICANN). Springer; 2006. p. 250–9.

[46] Liu W, Peng F, Feng S, You J, Chen Z, Wu J, et al. Semantic feature extraction for brain CT image clustering using nonnegative matrix factorization. In: Proceedings of the 1st international conference on medical biometrics (ICMB). Springer; 2007. p. 41–8.

[47] Liu W, Yuan K, Ye D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. J Biomed Inform 2008;41:602–6.

[48] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev 2009;51:455–500.

[49] Mørup M. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisc Rev: Data Min Knowl Discov 2011;1:24–40.

[50] Smilde A, Bro R, Geladi P. Multi-way analysis: applications in the chemical sciences. Wiley; 2004.

[51] Lu H, Plataniotis KN, Venetsanopoulos AN. A survey of multilinear subspace learning for tensor data. Pattern Recogn 2011;44:1540–51.

[52] Wang D, Kong S. Feature selection from high-order tensorial data via sparse decomposition. Pattern Recogn Lett 2012;33:1695–702.

[53] Carroll JD, Chang J-J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika 1970;35:283–319.

[54] Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. UCLA Work Papers Phonet 1970;16:1–84.

[55] Acar E, Yener B. Unsupervised multiway data analysis: a literature survey. IEEE Trans Knowl Data Eng 2009;21:6–20.

[56] Dauwels J, Garg L, Earnest A, Pang LK. Handling missing data in medical questionnaires using tensor decompositions. In: Proceedings of 8th international conference on information, communications and signal processing (ICICS); 2011. p. 1–5.

[57] Acar E, Aykut-Bingol C, Bingol H, Bro R, Yener B. Multiway analysis of epilepsy tensors. Bioinformatics 2007;23:i10–8.

[58] De Vos M, De Lathauwer L, Vanrumste B, Van Huffel S, Van Paesschen W. Canonical decomposition of ictal scalp EEG and accurate source localisation: principles and simulation study. Comput Intell Neurosci 2007:58253.

[59] Mørup M, Hansen LK, Herrmann CS, Parnas J, Arnfred SM. Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG. NeuroImage 2006;29. 10-10.

[60] Lee H, Kim Y-D, Cichocki A, Choi S. Nonnegative tensor factorization for continuous EEG classification. Int J Neural Syst 2007;17:305–17.

[61] Cichocki A, Phan AH, Caiafa C. Flexible HALS algorithms for sparse non-negative matrix/tensor factorization. In: Proceedings of 2008 IEEE international workshop on machine learning for signal processing; 2008. p. 73–8.

[62] Li Y, Ngom A. Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data. In: Proceedings of the IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2010. p. 438–43.

[63] Rodriguez G. Poisson models for count data. In: Lecture notes on generalized linear models; 2007. p. 1–14. <http://data.princeton.edu/wws509/notes/>.

[64] Heiler M, Schnörr C. Controlling sparseness in non-negative tensor factorization. In: Proceedings of the 9th European conference on computer vision. Berlin, Heidelberg: Springer; 2006. p. 56–67.

[65] Cichocki A, Zdunek R, Choi S, Plemmons R, Amari S-I. Novel multi-layer non-negative tensor factorization with sparsity constraints. In: 8th International conference on adaptive and natural computing algorithms. Springer; 2007. p. 271–80.

[66] Mørup M, Hansen LK, Arnfred SM. Algorithms for sparse nonnegative Tucker decompositions. Neural Comput 2008;20:2112–31.

[67] Liu J, Liu J, Wonka P, Ye J. Sparse non-negative tensor factorization using columnwise coordinate descent. Pattern Recogn 2012;45.

[68] Richesson RL, Rusincovitch SA, Wixted D. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013.

[69] Bach F, Jenatton R, Mairal J, Obozinski G. Optimization with sparsity-inducing penalties. Found Trends Machine Learn 2012;4.

[70] Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, et al. American heart association advocacy coordinating committee, stroke council, council on cardiovascular radiology and intervention, council on clinical cardiology, council on epidemiology and prevention, council on arteriosclerosis, thrombosis and vascular biology, council on cardiopulmonary, critical care, perioperative and resuscitation, council on cardiovascular nursing, council on the kidney in cardiovascular disease, council on cardiovascular surgery and anesthesia, and interdisciplinary council on quality of care and outcomes research, forecasting the future of cardiovascular disease in the United States: a policy statement from the American heart association. Circulation 2011;123:933–44.

[71] Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, et al. American heart association statistics committee and stroke statistics subcommittee, heart disease and stroke statistics–2012 update: a report from the American heart association. Circulation 2012;125:e2–e220.

[72] Elixhauser A, Steiner C. Readmissions to U.S. hospitals by diagnosis, 2010. Healthcare cost and utilization project (HCUP) statistical briefs, agency for healthcare research and quality; 2012.

[73] Ho KKL, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham study. J Am College Cardiol 1993;22:A6–A13.

[74] Lloyd-Jones DM, Larson MG, Leip EP, Beiser A, D'Agostino RB, Kannel WB, et al. Lifetime risk for developing congestive heart failure: the Framingham heart study. Circulation 2002;106:3068–72.

[75] Mejhert M, Kahan T, Persson H, Edner M. Predicting readmissions and cardiovascular events in heart failure patients. Int J Cardiol 2006;109:108–13.

[76] Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle heart failure model: prediction of survival in heart failure. Circulation 2006;113:1424–33.

[77] Luo D, Wang F, Sun J, Markatou M, Hu J, Ebadollahi S. SOR: scalable orthogonal regression for non-redundant feature selection and its healthcare applications. In: Proceedings of the 12th SIAM international conference on data mining (SDM); 2012. p. 576–87.

[78] Acar E, Kolda TG, Dunlavy DM. All-at-once optimization for coupled matrix and tensor factorizations. In: Proceedings of mining and learning with graphs (MLG); 2011.

[79] Acar E, Dunlavy DM, Kolda TG, Mørup M. Scalable tensor factorizations for incomplete data. Chemometr Intell Lab Syst 2011;106:41–56.

[80] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 2010;26:1205–10.