Journal of Biomedical Informatics 55 (2015) 82–93

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Building bridges across electronic health record systems through inferred phenotypic topics

You Chen [a,*], Joydeep Ghosh [b], Cosmin Adrian Bejan [a], Carl A. Gunter [c], Siddharth Gupta [c], Abel Kho [d], David Liebovitz [d], Jimeng Sun [e], Joshua Denny [a,f], Bradley Malin [a,g]

[a] Dept. of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN, USA
[b] Dept. of Electrical & Computer Engineering, University of Texas, Austin, TX, USA
[c] Dept. of Computer Science, University of Illinois at Urbana-Champagne, Champaign, IL, USA
[d] School of Medicine, Northwestern University, Chicago, IL, USA
[e] School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[f] Department of Medicine, Vanderbilt University, Nashville, TN, USA
[g] Dept. of Electrical Engineering & Computer Science, School of Engineering, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 2 December 2014
24 March 2015
Accepted 25 March 2015
Available online 1 April 2015

Keywords:
Clinical phenotype modeling
Computers and information processing
Data mining
Electronic medical records
Medical information systems
Pattern recognition

ABSTRACT

Objective: Data in electronic health records (EHRs) is being increasingly leveraged for secondary uses, ranging from biomedical association studies to comparative effectiveness. To perform studies at scale and transfer knowledge from one institution to another in a meaningful way, we need to harmonize the phenotypes in such systems. Traditionally, this has been accomplished through expert specification of phenotypes via standardized terminologies, such as billing codes. However, this approach may be biased by the experience and expectations of the experts, as well as the vocabulary used to describe such patients. The goal of this work is to develop a data-driven strategy to (1) infer phenotypic topics within patient populations and (2) assess the degree to which such topics facilitate a mapping across populations in disparate healthcare systems.

Methods: We adapt a generative topic modeling strategy, based on latent Dirichlet allocation, to infer phenotypic topics. We utilize a variance analysis to assess the projection of a patient population from one healthcare system onto the topics learned from another system. The consistency of learned phenotypic topics was evaluated using (1) the similarity of topics, (2) the stability of a patient population across topics, and (3) the transferability of a topic across sites. We evaluated our approaches using four months of inpatient data from two geographically distinct healthcare systems: (1) Northwestern Memorial Hospital (NMH) and (2) Vanderbilt University Medical Center (VUMC).

Results: The method learned 25 phenotypic topics from each healthcare system. The average cosine similarity between matched topics across the two sites was 0.39, a remarkably high value given the very high dimensionality of the feature space. The average stability of VUMC and NMH patients across the topics of two sites was 0.988 and 0.812, respectively, as measured by the Pearson correlation coefficient. Also the VUMC and NMH topics have smaller variance of characterizing patient population of two sites than standard clinical terminologies (e.g., ICD9), suggesting they may be more reliably transferred across hospital systems.

Conclusions: Phenotypic topics learned from EHR data can be more stable and transferable than billing codes for characterizing the general status of a patient population. This suggests that EHR-based research may be able to leverage such phenotypic topics as variables when pooling patient populations in predictive models.

1. Introduction

There is mounting evidence to suggest that data derived from electronic health records (EHRs) can be applied in a secondary fashion to support a wide range of activities. There are indications,

* Corresponding author at: 2525 West End Ave, Suite 1030, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37203, USA. Tel.: +1 615 343 1939; fax: +1 615 322 0502.
E-mail address: you.chen@vanderbilt.edu (Y. Chen).

http://dx.doi.org/10.1016/j.jbi.2015.03.011
1532-0464/

for instance, that EHR data can facilitate novel clinical decision support [1,2], conduct biomedical association studies [3–9], improve auditing and EHR security [10–12], and assess the cost effectiveness of treatments [13]. It is further anticipated that EHR data can be utilized to efficiently support a learning healthcare system, where information about care and operations is translated into knowledge for evidence-based clinical practice and positive change [14,15]. At the same time, there are significant challenges to reusing EHR data, including a lack of common standards to merge clinical data and translate clinical concepts between disparate healthcare systems [1,15,16]. As such, it is critical to develop scalable methods to learn clinical concepts (or phenotypes) that can be translated across disparate healthcare systems.

In recognition of such a challenge, the past several years have witnessed a movement toward strategies to engineer and implement processes that standardize EHRs and derived concepts [17–22]. These strategies are driven both by rule-based models that are specified by experts, as well as data-driven methods that attempt to learn patterns from the information within EHRs. With respect to rule-based models, researchers often rely upon billing codes (e.g., International Classification of Diseases, or ICD) or modified versions of such vocabularies (e.g., Phenome-Wide Association Study, or PheWAS, codes [23,24]) to characterize the diagnoses of patients in disparate healthcare systems (e.g., [19,20,22]). Since billing codes can be inaccurate, often other EHR data, such as medication and laboratory data, often are combined with billing data to form more accurate phenotypes [25]. However, these rule-based methods are limited by the significant amount of manual effort (e.g., physician chart reviews) required to implement them. Furthermore, these types of studies are only appropriate for known phenotypes. As a result, the process of investigating phenotypes across disparate healthcare systems is often quite slow and hampered in the discovery of new phenotypes. By contrast, data-driven methods rely upon techniques to learn phenotypic patterns from databases of EHRs (e.g., [17,18]). Yet these methods are also limited in that they learn patterns from healthcare systems independently.

This paper introduces a method to automatically learn phenotypic topics and evaluate their consistency across disparate healthcare systems. For this study, we limit our analysis to billing code data as a demonstration project to investigate the method, recognizing that if successful, the method could be applied to other discrete EHR data. Such topics can be leveraged as control variables to align patient populations across multiple systems. After validation by knowledgeable domain experts, such topics may become novel phenotypes that are worthy of further investigation.

The proposed method is composed of two primary steps. First, it infers phenotypic topics from the EHRs of each healthcare system through a generative model. Second, it measures the consistency of the learned topics for characterizing the patient populations across disparate systems. To the best of our knowledge, this is the first approach to automatically infer and test the alignment of phenotypic topics from the EHR data of multiple healthcare systems. To demonstrate feasibility, we perform an analysis on four months of inpatient billing data from two geographically distinct systems: (i) the Northwestern Memorial Hospital (NMH) and (ii) Vanderbilt University Medical Center (VUMC). The results demonstrate that learned phenotypic topics that appear to have a high degree of similarity can be found in two different healthcare systems.

The remainder of this paper is structured as follows. Section 2 introduces data-driven and expert-based phenotypic topic learning algorithms. Models of phenotypic topic learning and evaluation criteria for their consistency across multiple systems are introduced in Section 3. The design of the experimental environment is described in Section 4, while Section 5 reports on the corresponding results. A discussion of the findings, as well as limitations is provided in Section 6 and Section 7 provides a conclusion and next steps.

2. Background

Methods for modeling phenotypic topics through EHR data can roughly be categorized into (1) expert- and (2) data-driven. The former is based on the experience of clinically-knowledgeable individuals. As such, the process is often limited to known phenotypes and can be slow, particularly when validating the specification of a phenotype across disparate systems. The latter incorporates automation, which leads to significant gains in efficiency and strives to minimize manual attention. However, to date, phenotypic topics have been learned from healthcare systems independently, such that their ability to serve as common variables across healthcare systems is unknown.

2.1. Expert-based phenotypes

A significant number of healthcare organizations have implemented commercial EHR systems [26]. At times, systems are implemented, or adapted, in multiple sites according to standardized policies [27]. However, EHR systems remain highly diverse due to the fact that EHR (and terminology) utilization, as well business processes, is often site-specific [28,29]. As a result, it is difficult to perform investigations across sites [28–33]. Challenges remain in reusing such data for research, such as the mapping of the data to a common standard that can enable research across one large cohort [1,28]. As such, the research community is only beginning to use phenotypic concepts to merge patients with similar conditions (or specific diseases) from disparate systems [19,21,22,29–33].

Here, we consider several representative works for illustration. First, Tanpowpong and colleagues [31] evaluated the value of ICD-9 codes for identifying a specific phenotype in the form of celiac disease. To do so, they identified all adults with an ICD-9 code of 579.0 at three hospitals and stratified the cohort according to the presence/absence of relevant serology and endoscopy codes into four groups. Coloma and colleagues [32] moved beyond billing codes and demonstrated the potential for the integration of information from clinical narratives. Using the phenotype of acute myocardial infarction in EHR data from three European countries, it was shown that an approach using the combination of billing codes and free text yielded a better positive predictive value than an approach using codes alone.

Beyond identifying specific diseases, EHR-based phenotyping algorithms have been utilized to measure the similarity of patients from different sites. Schildcrout and colleagues [20] quantified the variability in comorbid ICD9 codes for six phenotypes across five sites, including type 2 diabetes and peripheral arterial disease. They found that patients with the same phenotype at disparate institutions appeared to exhibit more similar comorbidity profiles than those representing different phenotypes; however, there was still variability within the same phenotype at different institutions.

While a phenotyping algorithm can be specified using various terminologies, the application of an algorithm on patient cohorts in disparate settings can often yield differing results. In an attempt to address this challenge, it was indicated that standardized information modeling and meaningful use standards could be leveraged for the presentation of a phenotyping algorithm across institutions [29,30]. It was shown that a consensus model can be more effective than a single site's specification for phenotype discovery across sites.

While these studies illustrate the potential for data derived from EHRs and the need for harmonization of phenotype

definitions, they have several limitations. First, all of these studies require significant manual effort. This indicates that the speed of learning phenotypic topics across sites can be slow, lacking scalability to a large number of phenotypes. Second, such expert-based methods are restricted to known phenotypes, which limit their utility in discovery-based research.

2.2. Data-driven phenotypes

By contrast, data-driven methods aim to automate the mining of phenotypic topics from EHR data. There has been a flurry of activity in various automated learning methods for high-throughput phenotyping over the past several years.

First, it was recently shown that inductive logic programming (ILP) can be applied to EHR data to learn ICD-9 code based phenotypes [34]. However, in preparation for ILP, which was applied to identify phenotype features, the investigators needed to review and assign labels to a set of patient records that were representative of a larger corpus.

While the previous work relies upon supervised learning, more recent methods have focused on the unsupervised setting. Lasko et al. [17], for instance, introduced an unsupervised algorithm, based on deep learning methods, to discover phenotypic features from EHR data. This method relies upon Gaussian process regression, followed by a feature discovery step based on deep learning, to learn phenotypic features from sequences of serum uric acid measurements. It was shown that the learned features could accurately distinguish between the uric-acid signatures of gout and acute leukemia. Other approaches have applied matrix (or, more generally, tensor) factorization methods to derive phenotypic topics in temporal settings [35]. With respect to the latter, variations of unsupervised nonnegative tensor factorization methods have been introduced to decompose combinations of diagnoses, medications, and procedures [18,36]. This approach was applied, for instance, on a cohort of approximately 30,000 heart failure patients and illustrated that the top 40 phenotypic topics could outperform the original 640 features (which consisted of 169 diagnosis categories and 471 medication types) in learning patient clusters.

Beyond its application for mining phenotypic topics from EHR data, data-driven methods on EHR data have also been leveraged to mine communities of care providers [10,11], semantic concepts of patients [37] and clinical pathway patterns through the activity logs of healthcare systems [38–40]. For example, Huang and colleagues [39] used an altered latent Dirichlet allocation (LDA) model to infer patterns of clinical pathways from EHR activity logs. Specifically, they applied an altered LDA model on two cohorts: (1) patients treated for unstable angina and (2) patients treated in an oncological setting. The model inferred five clinical pathways for each of the two settings. Though a pilot study, it was demonstrated that learned pathway patterns can enable decision support and greater efficiency in coordinated clinical treatments. Bouarfa and Dankelman [38] derived a workflow consensus from clinical activity logs to detect outlying workflows without prior knowledge from experts. They adopted a tree-guided multiple sequence alignment approach to model the consensus of workflows. This strategy was validated over the workflow processes associated with laparoscopic cholecystectomy, where the results indicated the derived consensus conforms to the main steps of the surgical procedure as described in best practice guidelines.

The above data-driven research indicates the automated learning of concepts in the clinical domain can be efficient and scalable. However, the existing methods are limited in that they only learn phenotypic topics from the EHR of a single institution.

3. Methods

The general framework for the proposed method is depicted in Fig. 1. The framework is composed of two parts: (1) a topic learning model, which extracts phenotypic topics for each site (as depicted in the top part of the figure) and (2) three topic consistency measurements, which evaluate the consistency of phenotypic topics across disparate sites (as depicted in the bottom part of the figure).

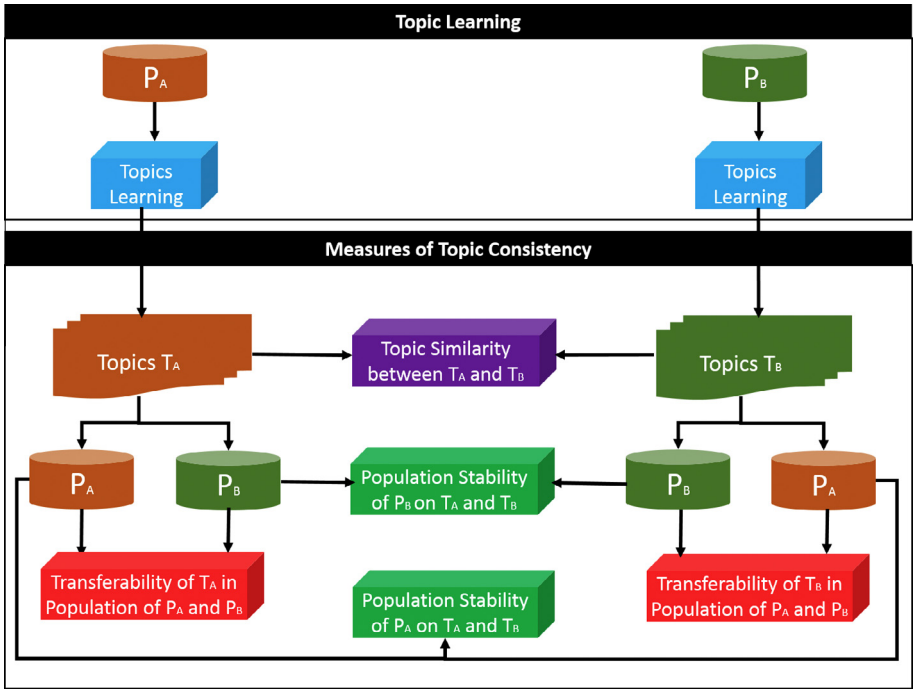


Fig. 1. A high-level overview of the architecture for extracting phenotypic topics and evaluating their consistency across healthcare systems.

Table 1
Common notation and the corresponding definitions.

Notation	Description
X	A healthcare system
$P_X = \{p_{X,1}, \ldots, p_{X,m}\}$	A set of patients from X
$C_X = \{c_{X,1}, \ldots, c_{X,n}\}$	A set of clinical terms defining patients in P_X
$T_X = \{t_{X,1}, \ldots, t_{X,k}\}$	A set of k phenotypic topics retrieved from P_X defined by n clinical terms in C_X
$G_{Y,Z} = \{g_1, \ldots, g_k\}$	A set of patient groups in P_Z clustered using k topics in T_Y
$\psi_{Y,Z}(size\, k \times m)$	A probability matrix of k topics in T_Y to characterize m patient in P_Z
$R_{Y,Z}(size\, 1 \times k)$	A vector of rates of patients in P_Z characterized by topics in T_Y

We now provide a high-level overview of the models and then proceed with a deeper dive into each component. For reference, a legend of the notation used throughout this paper is provided in Table 1.

For illustration, we assume there are two healthcare systems, A and B. We let P_A represent the set of patients from site A, where each patient is defined over a set of clinical terms in C_A. A clinical term corresponds to a phenomenon associated with the patient in the clinical domain. For instance, a clinical term could be a diagnostic billing code, a medication, a diagnosis extracted from natural language processing, or the finding of a laboratory test. The set of phenotypic topics T_A are learned in this space, and are characterized as a probability matrix of topics over clinical terms. Specifically, a topic corresponds to a pattern of co-occurring clinical terms, defined by their probability distribution given (or "conditioned on") that topic. A topic may or may not have an obviously clinical basis, but nevertheless can be useful for characterizing patients. We use $\psi_{A,A}$ and $\psi_{B,A}$ to represent matrices of probabilities that specify the likelihood that the patients in P_A are characterized by the topics in T_A and T_B, respectively. The terms P_B, C_B, T_B, $\psi_{A,B}$ and $\psi_{B,B}$ are defined similarly.

As mentioned earlier, there are numerous ways to learn from EHR data. In this work, we rely upon a general topic modeling strategy because it has a natural probabilistic interpretation. Once the phenotypic topics have been learned from each site, we evaluate their consistency from three perspectives: (1) *similarity* of topics of disparate sites, (2) *stability* of a population in the topics of disparate sites, and (3) *transferability* of a topic between disparate populations.

3.1. Topic learning model

We assume a patient is characterized by various clinical terms, such as diagnostic billing codes, and invoke an LDA model [41] to infer phenotypic topics. LDA is a probabilistic graphical model that was first developed to discover topics in natural language documents. It is a generative model that explains observations with hidden, or latent, patterns. Conceptually, patients can be thought of as documents, where the clinical terms constitute the vocabulary and the specific terms assigned to a patient's record are the "words". As such, given P_A characterized by clinical terms, we apply an LDA model to infer latent phenotypic topics T_A, each of which is composed of a probability distribution over the set of clinical terms.

The set of topics T_X is inferred from a matrix M_X (of size $m \times n$), where m is the number of patients in P_X and n is the number of distinct clinical terms in C_X. Here, $M_X(i,j)$ corresponds to the number of times clinical term $c_{X,j}$ in C_X was assigned to a patient $p_{X,i}$ in P_X.

LDA is applied to learn k latent topics $T_X = \{t_{X,1}, t_{X,2}, \ldots, t_{X,k}\}$. It is often the case that perplexity [42], an information theoretic measure, is applied to assess the fitness of an LDA model and set k. However, low perplexity is insufficient to indicate if the learned LDA model is a good fit [41,42]. In our situation, we aim to determine the k value that determines the best separation between the phenotypic topics. To do so, we calculate the average similarity of the topics:

$$\gamma(T_X) = \frac{2}{k \times (k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} cos(t_{X,i}, t_{X,j}) \quad (1)$$

where $\cos(t_{X,i}, t_{X,j})$ is the cosine similarity [43] of topics $t_{X,i}$ and $t_{X,j}$.[1]

3.2. Measures of topic consistency

We evaluated the consistency of the inferred topics using three quantitative measures: (1) similarity of topics, (2) stability of patient cohorts across topics, and (3) transferability of topics across sites.

3.2.1. Topic similarity

The first topic consistency measure directly assesses the similarity of the inferred topics from disparate sites. Note, however, that T_A and T_B have a different number of rows (i.e., diagnoses). So to compare learned phenotypic topics, we substitute T_A and T_B with a vector U_{AB} (size $n_U \times 1$) that represents the union of diagnoses, such that $U_{AB} = C_A C_B$. Thus, topics T_A and T_B are rewritten as T'_A (size $n_U \times k_A$) and T'_B (size $n_U \times k_B$). Based on this representation, the similarity of two phenotypic topics is calculated using the cosine similarity of the vectors:

$$\beta(t_{A,i}, t_{B,j}) = \frac{u_i \cdot u_j}{|u_i||u_j|} (u_i \in T'_A, 1 \leqslant i \leqslant k_A; \quad u_j \in T'_B, 1 \leqslant j \leqslant k_B) \quad (2)$$

The larger the β, the stronger the similarity of the phenotypic topics.

Our aim is to find the largest average cosine similarity, where each topic in T_A matches a topic in T_B and vice versa. We use the Hungarian algorithm [45] to perform such matches.[2] To do so, let Ω be a matrix (sized $k_A \times k_B$) that conveys the costs of matching topics between sites A and B, where cell Ω (i,j) indicates the cost of matching topic $t_{A,i}$ and topic $t_{B,j}$. We assume that if the cosine similarity of a pair of topics is 1, then the cost of this matching is 0, such that the cost of a topic matching as:

$$\Omega(t_{A,i}, t_{B,j}) = |\beta - 1| \quad (3)$$

The topic similarity is thus defined as the minimum sum of costs for the maximum matching of topics between t_A and t_B. The higher the topic similarity between two sites, the smaller the cost.

3.2.2. Population stability

The second consistency measure assesses the stability of a patient population across the topics derived from disparate sites. When the stability of a patient population is high, it is likely that the topics from one site will characterize the patients from another site.

$\psi_{Y,Z}$ is defined as a matrix of probabilities of patients in P_Z characterized by topics in T_Y. $\psi_{Y,Z}$ is retrieved by an inferred LDA model, which is based on an existing LDA model of site Y to characterize patients of site Z. According to definition of $\psi_{Y,Z}, \psi_{A,A}$ (size $k_A \times m_A$) and $\psi_{B,A}$ (size $k_B \times m_A$) represent the probabilities

[1] Kullback–Leibler divergence (KLD) is often applied to measuring the divergence between two probability distributions [44] because of its sound basis in information theory. However, there are several problems. First, KLD is asymmetric with respect to the distributions. Second, the topics should be well separated and hopefully sparse, but, unless the estimated probability distributions are smoothed (e.g., via Laplace smoothing), this can lead to KLD becoming unbounded. The cosine, by contrast, is not subject to these limitations.

[2] This algorithm is efficient when the cost matrix is small.

that the topics in T_A and T_B, respectively, characterize the patients in P_A. Specifically, a cell $\psi_{A,A}(i,j)$ corresponds to the probability that topic $t_{A,i}$ in T_A characterizes patient $p_{A,j}$ in P_A. When a patient in P_A is characterized by a phenotypic topic $t_{A,i}$ (or $t_{B,i}$) with a probability greater than a predefined threshold, we assign the patient to the topic. And thus, $\psi_{A,A}$ and $\psi_{B,A}$ can be invoked to group patients in P_A. In doing so, each phenotypic topic has a corresponding group of patients.[3]

Let $T_{Align,A}$ and $T_{Align,B}$ be a reordering of topics in T_A and T_B, respectively, such that $T_{Align,A}(i)$ most closely matches $T_{Align,B}(i)$ as per the Hungarian algorithm. For example, imagine $T_A = \{t_{A,1}, t_{A,2}, t_{A,3}\}$ and $T_B = \{t_{B,1}, t_{B,2}, t_{B,3}\}$, and the Hungarian algorithm matches $t_{A,1}$ with $t_{B,2}$,$t_{A,2}$ with $t_{B,3}$, and $t_{A,3}$ with $t_{B,1}$. Then, $T_{Align,A} = \{t_{A,1}, t_{A,2}, t_{A,3}\}$ and $T_{Align,B} = t_{B,1}, t_{B,2}, t_{B,3}$.

Now, let $G_{A,A} = \{g_1, \ldots, g_{k_A}\}$ and $G_{B,A} = \{g_1, \ldots, g_{k_B}\}$ be the sets of groups for the patients in P_A associated with the topics in $T_{Align,A}$ and $T_{Align,B}$, respectively. Moreover, let $G'_{A,A} = [|g_1|, \ldots, |g_{k_A}|]$ and $G'_{B,A} = [|g_1|, \ldots, |g_{k_B}|]$ represent the vectors with the number of patients per group. Population stability focuses on the relationship of the set of matched proportions (i.e., where each point is the rate at which patients in population A are characterized by the matched topics of sites A and B). So, we apply the Pearson correlation coefficient [46] to $G'_{A,A}$ and $G'_{B,A}$:

$$\rho(G'_{A,A}, G'_{B,A}) = \frac{COV(G'_{A,A}, G'_{B,A})}{\sigma_{G'_{A,A}} \sigma_{G'_{B,A}}} \tag{4}$$

where $cov(\cdot,\cdot)$ is the covariance and $\sigma_{G'_{A,A}}$ and $\sigma_{G'_{B,A}}$ are the standard deviations of $G'_{A,A}$ and $G'_{B,A}$, respectively. The correlation of $G'_{A,B}$ and $G'_{B,B}$ is defined similarly. The stability of a population on topics of two sites is measured via the Pearson correlation coefficient, as indicated by Eq. (4).

3.2.3. Topic transferability

The third consistency measure assesses how phenotypic topics transfer from one site to another. We aim to learn topics that characterize patients at a similar rate across the sites. This is because similar rates suggest that the sites manage similar populations.

To assess the transferability of topics in T_A, we define the following regression model:

$$\log R_{A,A} = \log I + \alpha \log R_{A,B} \tag{5}$$

where $R_{A,A}$ ($R_{A,B}$) is a vector of the rates at which patients from P_A (P_B) are characterized by the learned phenotypic topics in T_A, α is the slope of the regression and I is the intercept.[4]

Transferability of topics within a site is defined as the mean and standard deviation of distances for all of its phenotypic topics to the regressed line. To illustrate, consider a topic $t_{A,i}$ in T_A. The distance of this topic to the line is:

$$\begin{aligned} \mathrm{dis}(r_{i,B}, r_{i,A}) = {} & \frac{\max(r_{i,A}, 10^I \times r_{i,B}^{\alpha})}{\min(r_{i,A}, 10^I \times r_{i,B}^{\alpha})} \log(|r_{i,A} - 10^I \times r_{i,B}^{\alpha}| + 1) \\ & \times (r_{i,A} \in R_{A,A}, r_{i,B} \in R_{A,B};\ 1 \leqslant i,j \leqslant k_A) \end{aligned} \tag{6}$$

where $r_{i,A}$ and $r_{i,B}$ are the rates at which a learned topic $t_{A,i}$ is expressed by patients at site A and B respectively. $(r_{i,B}, 10^I \times r_{i,B}^{\alpha})$ is the corresponding point on the regressed line for $(r_{i,B}, r_{i,A})$. The term $\frac{\max(r_{i,A}, 10^I \times r_{i,B}^{\alpha})}{\min(r_{i,A}, 10^I \times r_{i,B}^{\alpha})}$ is a scaling factor that magnifies the effect of outliers on the transferability of phenotypic topics, the justification for which is in Appendix A1. A logarithmic transformation is applied for normalization and ensures that the distance of a point that falls on the regressed line is equal to zero.

4. Experimental design

4.1. Datasets

We evaluate the reliability of phenotypic topics on de-identified data from the EHRs of two distinct healthcare systems. The first dataset corresponds to four months of inpatient records from the StarPanel EHR system of the Vanderbilt University Medical Center (VUMC) [47]. The second dataset corresponds to four months of inpatient records from the EHR system of Northwestern Memorial Hospital (NMH) [37]. There are 14,606 and 17,947 inpatients at NMH and VUMC, respectively. Additional summary information about the datasets are provided in Table 2.

While we recognize that the clinical status of a patient is complex, this work focuses on a proof of concept and relies upon the billing codes to learn phenotypic topics. Such codes do not provide a complete picture of the status of a patient, but they are common in biomedical research and can provide insight into the capability of such a strategy. Nonetheless, multiple billing codes can be used to describe the same clinical disease [48,49], such that various EHR-driven phenotyping investigations (e.g., [19,29,30]) have instead adopted the Phenome-Wide Association Study (PheWAS) vocabulary [23]. PheWAS codes correspond to groups of ICD-9 codes more closely match clinical or genetic understandings of diseases and reduce variability in identifying diseases. Based on this expectation, and to be in accordance with prior work in phenotyping, we translate a patient's ICD-9 codes to PheWAS codes. All of the learned phenotypic topics reported in this paper are based on the PheWAS codes.

4.2. Setting the number of phenotypic topics

We train LDA model by using Gibbs sampling which is a typical technique for parameter estimation and then check the negative log-likelihood at each iteration to judge when a model has converged upon a solution. To parameterize the number of phenotypic topics for the LDA model, we minimize (1) the perplexity score and (2) the average similarity of the topics within a site. Based on these measures, we set the number of topics to 25 for each site. Further details of this process can be found in Appendix A2.

4.3. Consistency of NMH and VUMC topics

For topic similarity, we calculate the cost of matching NMH and VUMC topics using the Hungarian algorithm on the cost matrix defined in Eq. (3). For each VUMC phenotypic topic, we match a NMH topic and vice versa. If each phenotypic topic in one site has a matching topic in another site with a low cost, it implies that the topics are common across the sites.

For stability, we calculate the Pearson correlation coefficient of a patient population characterized by NMH and VUMC topics respectively. The higher coefficient, the more stable for a population characterized on NMH and VUMC topics. We use the Pearson correlation coefficient (Eq. (4)) to calculate the stability.

For transferability, we learn a regression model for the NMH and VUMC phenotypic topics, respectively (Eq. (5)). We then compute the distance of a topic to the regressed line (Eq. (6)). We use

[3] The value for such a threshold is dependent on the application. A value of 0.5, which we use in this work, signifies that the majority of the patient's status is captured by a single concept.

[4] This model builds on the observation in [20] that the rate of occurrence for billing codes in disparate sites is distributed around a centered line in the log scale.

Table 2
Summary information for four months of inpatient data derived from the EHRs.

	Northwestern	Vanderbilt
Patients	14,606	17,947
Unique ICD-9 codes	4543	5176
Unique PheWAS codes	1447	1413
Unique ⟨ICD9 code, patient⟩ assignments	114,133	84,331
Unique ⟨PheWAS code, patient⟩ assignments	90,732	74,192

the variance of the regression line to characterize the transferability of the corresponding model. To demonstrate the transferability of learned topics, we also conduct an analysis that compares the transferability of the learned phenotypic topics with ICD-9 and PheWAS codes. The rate at which ICD-9 and PheWAS codes transpire in the patients have a wide range (as discussed below), such that we use a binning approach to reduce the standard error of the linear regression [50]. Specifically, we use 50 bins, where each bin maps the set of points in a rectangular area of the distribution to a mean value which is supplied to the linear regression.

5. Results

To orient the reader, this section begins with a depiction of several learned phenotypic topics. We then report on the similarity, stability, and transferability of the derived topics.

5.1. Learned phenotypic topics

To better understand our experimental results, this section exemplifies a selected set of phenotypic topics inferred by the LDA model from the NMH and VUMC datasets. In our framework, each topic is expressed as a probability distribution over approximately 1500 PheWAS codes. To illustrate each topic in a succinct manner, we show the top five most probable PheWAS codes that best describe the corresponding topic. Details on topics not listed in this section are provided in Appendices A3 (for VUMC) and A4 (for NMH).

Figs. 2–4 depict several notable groups of topics. Fig. 2 shows a pair of topics (N_{13} and V_7) that exhibits high similarity. Fig. 3 shows three NMH topics (N_2, N_4 and N_{17}), that are similar to the same VUMC topic (V_4), and generally correspond to a collection of conditions associated with pregnancy and birth. Fig. 4 shows four topics (V_1, V_{15}, V_{18}, and N_{24}), each of which lacks a corresponding topic at the other site.

PheWAS code	Description	Probability
V_7		
411.4	Coronary atherosclerosis	0.31
411.2	Myocardial infarction	0.22
496	Chronic airway obstruction	0.10
415	Pulmonary heart disease	0.08
418	Nonspecific chest pain	0.06
N_{13}		
411.4	Coronary atherosclerosis	0.56
411.2	Myocardial infarction	0.15
272.1	Hyperlipidemia	0.11
443.9	Peripheral arterial disease	0.02
418	Nonspecific chest pain	0.02

Fig. 2. The top five PheWAS codes in the pair of phenotypic topics with the highest similarity (a score of 0.86).

5.2. Consistency of phenotypic topics

5.2.1. Similarity of topics

The similarity of each phenotypic topic pair from NMH and VUMC is depicted in the heatmap in Fig. 5a. It can be seen that, for the majority of the topics, the similarity is high for the best match. To show the pairs with strong relations more clearly, Fig. 5b displays a bipartite network of the similarity scores with values larger than 0.2. Here, it can be seen that almost every NMH phenotypic topic has at least one corresponding VUMC phenotypic topic. The only NMH topic that fails to have a partner is topic N_{24}, which is primarily associated with thrombosis. Similarly, almost every VUMC topic has a corresponding NMH topic. The exceptions are V_1, V_{15}, and V_{18}, which are most associated with perinatal conditions, internal injuries to organs, and burns.

The results and corresponding cost of the alignment of the topics is reported in Appendix A5. It was found that the total cost for a maximum matching[5] of topics between NMH and VUMC is 15.26. The average cost for each pair of phenotypic topics is 0.61, which indicates that the average cosine similarity for a pair of aligned phenotypic topics is 0.39 (Eq. (3)). The cost of alignment for the learned phenotypic topics is statistically significantly smaller than that of alignments for phenotypic topics in a random setting (details of the hypothesis test are in Appendix A5).

To illustrate the result of the alignment, Fig. 6 compares matched phenotypic topics N_{13} and V_7. This relationship appears natural because both topics are primarily associated with "coronary atherosclerosis" and "myocardial infarction" (each exhibits a high probability within the topics). At the same time, it should be noted that these topics include additional terms, such as "chronic airway obstruction", "pulmonary heart disease", "hyperlipidemia" and "peripheral arterial disease". Yet these terms exhibit lower probabilities, suggesting the topics consist of a core and ancillary set concepts, the latter of which are nuanced and may be driven by population-specific issues.

5.2.2. Stability of a patient population over topics

The second consistency measure assesses the stability of a patient population (e.g., VUMC patient population) on phenotypic topics learned from the NMH and VUMC datasets. The goal of this portion of the investigation is to measure the relations between a patient population characterized by its own phenotypic topics and that characterized by the corresponding topics of the other site. To do so, we aligned the VUMC and NMH topics and get the corresponding clusters of patients from a site (e.g., VUMC). The alignment is shown in Table A1 of Appendix A5 and the resulting size of the clusters is shown in Fig. 7.

The Pearson correlation coefficient of the VUMC and NMH populations is 0.957 and 0.649, respectively. This indicates there is generally high stability in the learned phenotypic topics across the sites. While the correlation for the NMH patients is clearly smaller than that which is observed for the VUMC patients, this is mainly because NMH has a higher volume of patients with certain conditions: 278.1 – *Obesity*; 649 – *Mother Complicating Pregnancy;* 665 – *Obstetrical/Birth Trauma*; and 645 – *Late Pregnancy and Failed Induction*, which are captured by three NMH topics (N_2, N_4 and N_{17}), but only one VUMC topic (V_4). The composition of these topics is summarized in Fig. 3.

Note that, as depicted in Fig. 7a, phenotypic topic V_4 is expressed by over 30% of the NMH patients. Based on this observation, we performed a sub-analysis on the patient population that

[5] A maximum matching transpires when every topic in NMH has a corresponding topic in VUMC and vice versa.

PheWAS code	Description	Probability
V_4		
649	Other conditions of the mother complicating pregnancy	0.13
278.1	Obesity	0.11
665	Obstetrical/birth trauma	0.09
655.1	Abnormality in fetal heart rate or rhythm	0.06
645	Late pregnancy and failed induction	0.05
N_2		
665	Obstetrical/birth trauma	0.32
649	Other conditions of the mother complicating pregnancy	0.27
1010	Other tests	0.20
663	Umbilical cord complications during labor and delivery	0.11
659	Indications for care or intervention related to labor and delivery NEC	0.04
N_4		
665	Obstetrical/birth trauma	0.25
663	Umbilical cord complications during labor and delivery	0.12
659	Indications for care or intervention related to labor and delivery NEC	0.10
661	Fetal distress and abnormal forces of labor	0.10
655.1	Abnormality in fetal heart rate or rhythm	0.09
N_{17}		
652	Malposition and malpresentation of fetus or obstruction	0.15
649	Other conditions of the mother complicating pregnancy	0.07
659	Indications for care or intervention related to labor and delivery NEC	0.07
651	Multiple gestation	0.06
658	Problems associated with amniotic cavity and membranes	0.05

Fig. 3. Three phenotypic topics from Northwestern that are well matched with topic 4 from Vanderbilt.

was not explained by $\{N_2, N_4, N_{17}\}$ and their corresponding aligned topics $\{V_{14}, V_4, V_1\}$ as depicted in Table A1 of Appendix A5. The correlation marginally increases for the VUMC patients (0.988), and substantially increases for the NMH patients (0.812). This suggests that a patient population on the learned phenotypic topics may be more stable when the sites are focused on a broad variety of patients (i.e., beyond several specific conditions).

To illustrate the stability of a patient population more specifically, let us consider a brief case study of N_{13} and V_7. Fig. 8 illustrates the intersection of NMH (a) and VUMC (b) patients assigned to these topics. It can be seen that both topics are expressed by most of the patients with a probability larger than 0.5.[6]

We calculate the rate of patients in common for these two phenotypic topics using the Jaccard measure:

$$r(V_7, N_{13}) = \frac{E_{V7,VUMC} \cap E_{N13,VUMC}}{E_{V7,VUMC} \cup E_{N13,VUMC}}, \quad (7)$$

where $E_{V7,VUMC}$ and $E_{N13,VUMC}$ are the sets of VUMC patients assigned to topics V_7 and N_{13}, respectively. The degree of commonality for the NMH and VUMC patients is 0.35 and 0.51, which indicates a relatively high rate of patients in common.

5.2.3. Transferability of topics

To determine if phenotypic topics are more transferable than expert-derived vocabularies for characterizing patient populations, we compared their variance of transferability to ICD-9 and PheWAS codes. For illustration, the distribution of the rate at which codes from the expert-derived vocabularies are expressed by patients is depicted in Fig. 9.

Notably, certain codes associated with common chronic diseases, such as ICD-9 401.9 and PheWAS 401.1 (a translation of ICD-9 401, 401.1 and 401.9), which are both associated with hypertension, are stable across the VUMC and NMH patient populations. However, there are certain instances where the codes exhibit a large variance in the population. Clear examples of this case are ICD-9 codes V05.3 – *need for prophylactic vaccination and inoculation against viral hepatitis* and V30.0 – *Single liveborn, born in hospital, delivered without mention of cesarean*, as well as PheWAS codes 656 – *Other perinatal conditions* and 637 – *Short gestation; low birth weight; and fetal growth retardation*.

The regression models for assessing transferability are summarized in Appendix A6. In summary, the average distance (and its corresponding standard deviation) of the ICD-9 and PheWAS codes to their regressed lines, are depicted in Fig. 11. It can be seen that the ICD-9 codes (0.0109 ± 0.2215) exhibit a larger variance than the PheWAS codes (0.0108 ± 0.1299). This is due, in part, to the fact that most of the codes which are rare at one site (i.e., the upper left and bottom right of the plots in Fig. 9) have a wider variance to the regressed line. By contrast, the codes that are more common (i.e., the upper right of the plots in Fig. 9), such as essential hypertension exhibit low variance and, thus, are more stable for expressing the patient population than those locate in the left-bottom corner. The PheWAS codes exhibit a smaller variance than the ICD9 codes, which suggests the codes are consistently utilized to represent patients with a particular clinical notion across disparate sites.

For the learned phenotypic topics, we compute the regression models (which we refer to as *N-Topic* and *V-Topic*) and calculate the distance of topics to the regressed line. Fig. 10 depicts the rate at which the phenotypic topics occur in the NMH and VUMC

[6] Recall, a patient is considered assigned to a phenotypic concept when the probability is greater than 0.5.

PheWAS code	Description	Probability
V_1		
656	Other perinatal conditions	0.19
637	Short gestation; low birth weight; and fetal growth retardation	0.11
647	Infectious & parasitic conditions complicating pregnancy	0.08
656.2	Other respiratory conditions of fetus and newborn	0.06
452	Venous embolism & thrombosis	0.05
V_{15}		
1008	Internal injury to organs	0.27
958	Early complications of trauma or procedure	0.08
807	Fracture of ribs	0.08
560.1	Paralytic ileus	0.06
348	Other conditions of brain	0.05
V_{18}		
1000	Burns	0.30
854	Mechanical complications of cardiac/vascular device, implant, and graft	0.06
174.11	Breast Cancer	0.04
480.3	Pneumonia due to fungus	0.04
289.4	Lymphadenitis	0.02
N_{24}		
452	Venous embolism & thrombosis	0.19
452.2	Deep vein thrombosis	0.10
286.5	Hemorrhagic disorder due to intrinsic circulating anticoagulants	0.09
415.1	Pulmonary embolism and infarction	0.06
819	Skull fracture and other intracranial injury	0.04

Fig. 4. Three phenotypic topics from Vanderbilt and one topic from Northwestern lack a corresponding topic of other site with a similarity greater than 0.2.

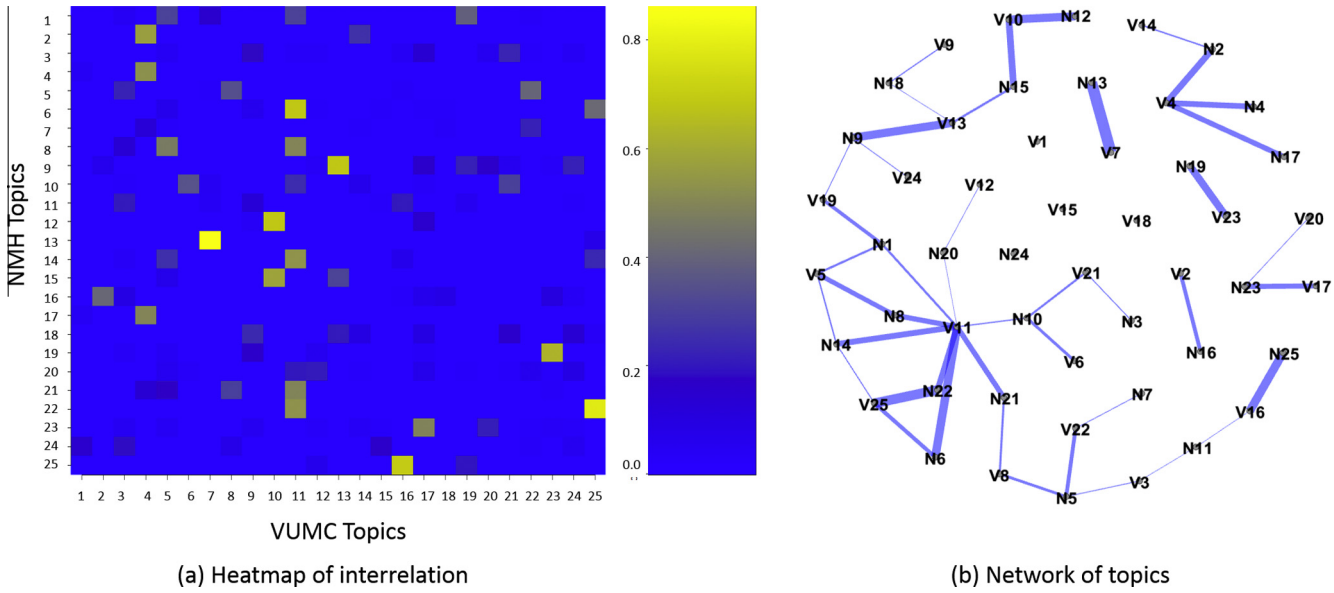


Fig. 5. Similarity of NMH and VUMC topics in (a) heatmap form and (b) network form (for scores $\geqslant 0.2$). Lines drawn in (b) are connections only for pairs of topics at different sites. The wider thickness of the line indicates tighter relations of a pair of topics.

populations. It can be seen that the distribution of phenotypic topics exhibits smaller variance than the ICD-9 and PheWAS codes. This is more formally confirmed in Fig. 11, which shows that the *N-Topic* (0.0055 ± 0.07) and *V-Topic* (0.00202 ± 0.06) models have smaller standard deviations than those of ICD-9 (0.0109 ± 0.2215) and PheWAS codes (0.0108 ± 0.13).

However, there are three outliers in the NMH topics (Fig. 10a) and one in the VUMC topics (Fig. 10b). This is because a large proportion of the NMH patients are related with Obstetrical/birth trauma conditions. As alluded to earlier, these conditions are expressed by topics N_2, N_4, N_{17}, and V_4, which form a community. The proportion of patients characterized by these four topics is high, which will

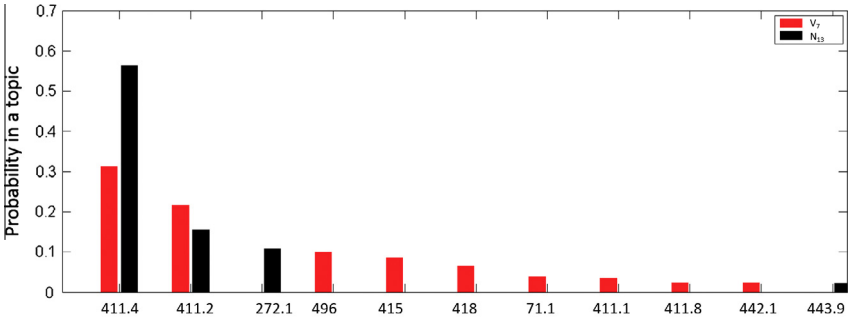


Fig. 6. Comparison of the top PheWAS codes associated with topics N_{13} and V_7.

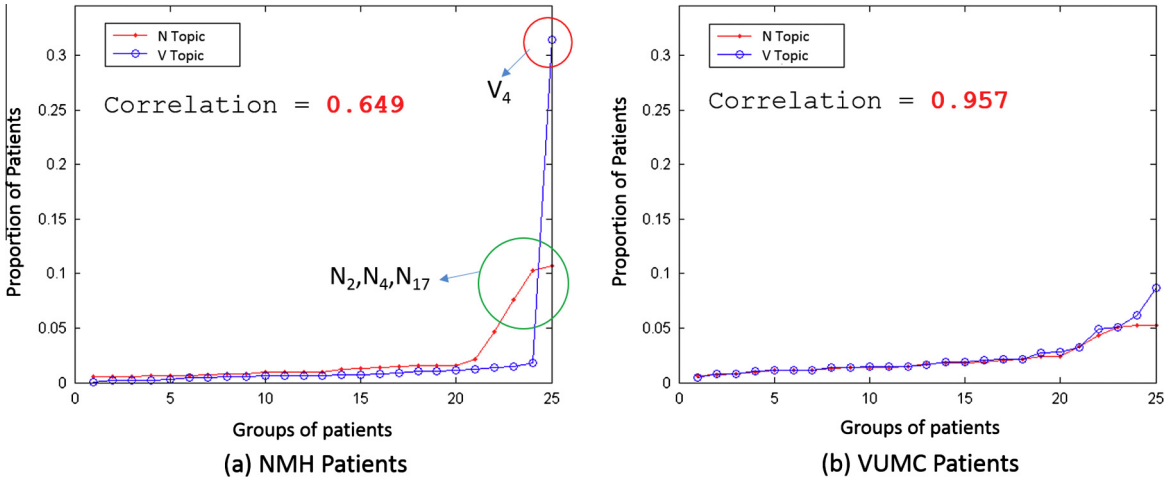


Fig. 7. Pearson correlation between the rate at which (a) NMH and (b) VUMC patients are characterized by phenotypic topics derived from the two sites.

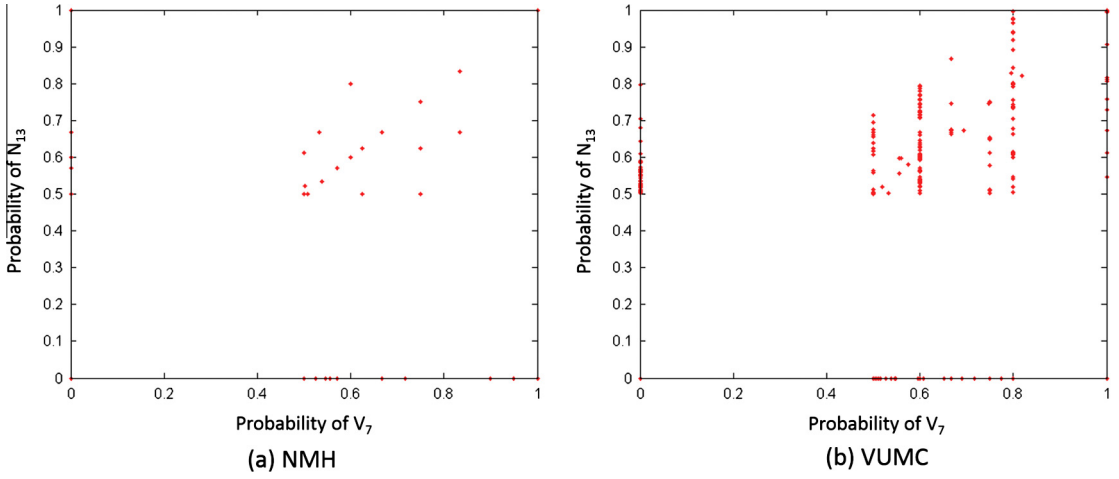
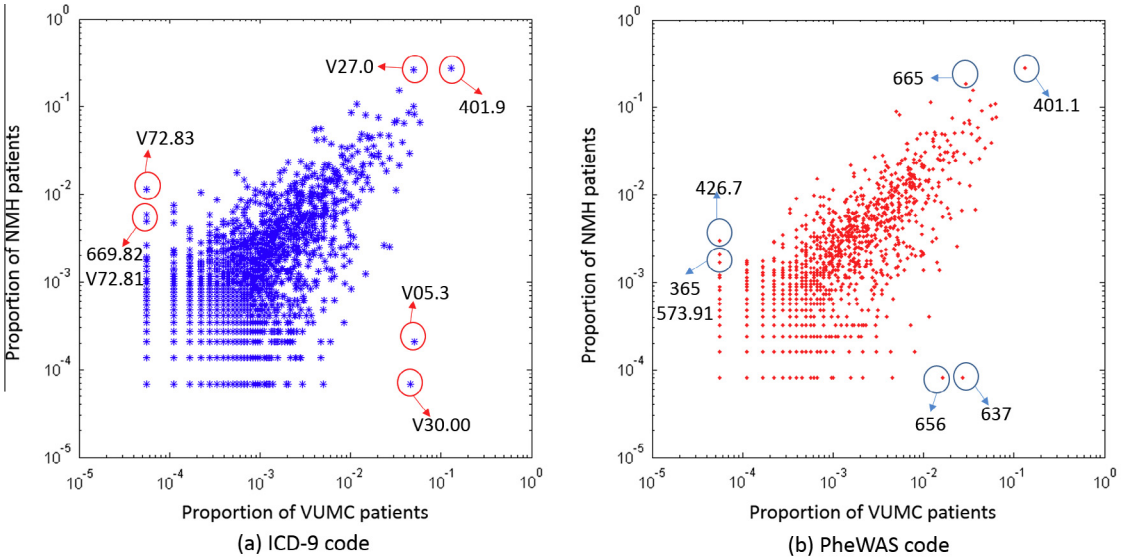


Fig. 8. Extent to which (a) NMH and (b) VUMC patients are expressed by phenotypic topics N_{13} and V_7. Each point corresponds to the probability a specific patient is characterized by a topic.

dominate the variances of other topics. We thus removed the outliers and remodeled the patients as *N-Topic-Reduced* (0.00087 ± 0.0021) and *V-Topic-Reduced* (0.0011 ± 0.001), the results for which are also shown in Fig. 11. It can be seen these models exhibit the smallest variance, suggesting they are the most transferable for characterizing the patients across the sites.



401.9 Unspecified essential hypertension
V27.0 Outcome of delivery, single liveborn
V05.3 Need for prophylactic vaccination and inoculation against viral hepatitis
V30.00 Single liveborn, born in hospital, delivered without mention of cesarean section
V72.83 Other specified pre-operative examination
669.82 Other complications of labor and delivery, delivered, with mention of postpartum complication
V72.81 Pre-operative cardiovascular examination

401.1 Essential hypertension
665 Obstetrical/birth trauma
637 Short gestation; low birth weight; and fetal growth retardation
656 Other perinatal conditions
426.7 Abnormal electrocardiogram
365 Glaucoma
573.91 Elevated levels of transaminase or lactic acid dehydrogenase

Fig. 9. Rate at which (a) ICD-9 and (b) PheWAS codes are expressed in the VUMC and NMH inpatient populations.

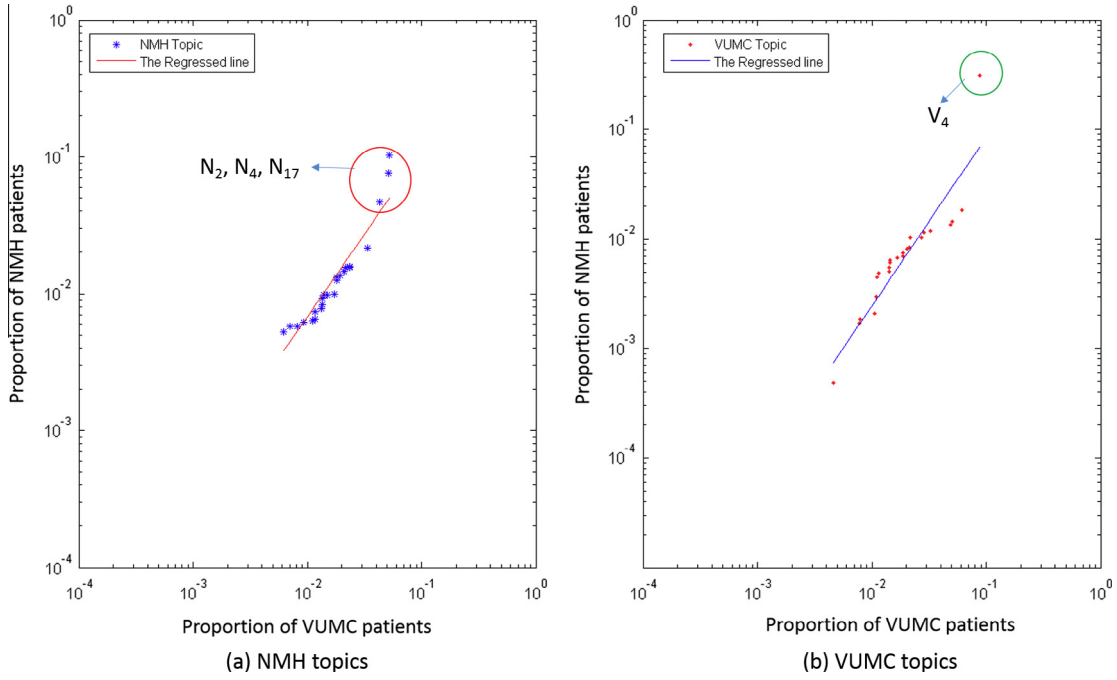


Fig. 10. Rate at which phenotypic topics learned from (a) NMH and (b) VUMC occur in the NMH and VUMC patient populations.

6. Discussion

In general, the experimental results suggest that phenotypic topics, learned through a generative topic modeling strategy (i.e., LDA) in the inpatient populations of two distinct healthcare systems, exhibit high consistency. This finding has several notable implications. First, the learned phenotypic topics could be invoked as covariates when investigating expert-defined phenotypes across healthcare systems. For example, in a diabetes-related investigation, the phenotypic topics V_{11} and V_{22}, which capture aspects of coronary heart disease, may serve as control variables that represent the complexity of such confounding clinical problems. Second, the learned topics may enable novel quality-based studies across systems in their own right. For instance, it would be possible

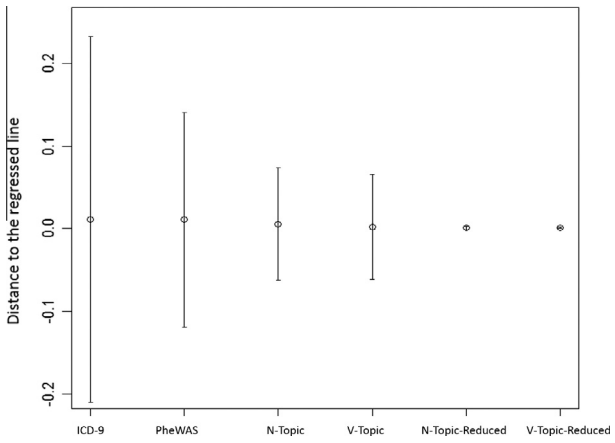


Fig. 11. Average distance (+/- 1 standard deviation) to the regressed line of vocabulary-based and learned phenotypic topic model.

to investigate how the quality of outcomes for phenotypes associated with a complex pregnancy (e.g., V_4 integrates delivery, obesity, and fetal heart rate).

At the same time, there are several limitations to this investigation. First, our notion of transferability is based on the premise that a topic should occur at the same rate at disparate healthcare organizations. However, if a topic occurs at varying rates, it does not imply that the topic is useless. Rather, it could imply that the organizations have different types of populations. The topics themselves may still be notable and worthy of further investigation, but we stress that they limit the extent to which population-based results at each institution are directly relatable.

Second, we acknowledge that this is a pilot study, which only focuses on the phenotypic topics that can be discovered through the ICD-9 (and PheWAS) codes assigned to patients while they are admitted to the hospital. While diagnosis codes do not provide a complete view of a patient, they are common in biomedical research. However, it should be noted that the methodological component of this work is not dependent upon diagnoses codes, or any particular clinical vocabulary, such that it can readily be extended to create more complex and robust phenotypes. As this work is extended, it will be necessary to enhance the approach and account for the semantics of the patient and healthcare setting (e.g., inpatient vs. outpatient), where the distribution of such terms may be utilized at other rates.

Third, while the model proposed in this paper is scalable to other types of structured data (e.g., medications, clinical concepts extracted from natural language notes, laboratory test findings), it is not trivial to combine other data types across different EHR systems. More work is needed to determine how such additional factors can be combined with diagnoses to learn more comprehensive and nuanced phenotypic topics.

Fourth, as this approach is rolled out to a larger number of healthcare systems, it will be critical to devise and apply reliability measures that account for more than two sites. We anticipate that this may be accomplished through the extension of the basic bivariate correlation to a multiple correlation model.

Fifth, our work focuses on the development of a methodology to learn phenotypic topics to align disparate patient populations. However, we did not validate the clinical meaning of such topics nor the semantics of the similarity between identified groups of patients. If such topics are to be used in association studies, their meaning must be interpreted by clinically knowledgeable experts.

Finally, the case study was performed with two healthcare systems only, which themselves may cover different types of patients. As such, it is not clear if the phenotypic topics, or the transferability of the topics discovered in this study, are directly applicable to other healthcare systems. Moreover, the case study focused on all inpatients in the system simultaneously. At VUMC, this population includes patients from multiple hospitals, including the primary hospital, children's hospital, as well as psychiatric and rehabilitation hospitals. In doing so, we incorporated neonatal, pediatric, and adult populations, which may confound the learning process. Furthermore, NMH does not have a focus on birth or children, such that the VUMC and NMH populations are not quite the same. We suspect that the learning process has the ability to discover phenotypic topics that are specific to certain demographics (age and gender), but note that this warrants further investigation.

7. Conclusions

Data derived from electronic health record (EHR) systems has the potential to enable large studies that incorporate disjoint healthcare providers, as well as support learning healthcare systems. However, it is challenging to automate learning across such systems due to a lack of standards in the use of clinical vocabularies. In this paper, we investigated the extent to which an automated learning approach, based on latent Dirichlet allocation, could be leveraged to infer phenotypic topics that are consistently defined across healthcare systems.

Specifically, we evaluated the approach with four months of inpatient data from two large geographically distinct hospital systems. The results illustrate that latent topics can reduce dimensionality and increase the stability and transferability of phenotypic topics studied across such sites. In particular, the findings suggest such an approach can enable the characterization of complex phenotypic topics that could be invoked as covariates in multi-site studies or analyzed in comparative consistency assessments for healthcare systems. Nonetheless, we stress that there are several opportunities for enhancing the proposed strategy. In particular, the current study focused solely on diagnosis codes, but more comprehensive and nuanced phenotypic topics should be discovered via an expansion of the vocabulary to contain additional phenomena, such procedures, medications, and laboratory tests.

Conflict of interest

There is no conflict of interest.

Acknowledgments

We would like to thank Daniel Schneider and Prasanth Nannapaneni for gathering and supplying the data from Northwestern Memorial Hospital and Steve Nyemba from Vanderbilt University. Thank Dr. Daniel Fabbri and Dr. Tom Lasko for helpful discussions in the early days of this research. This work is supported by the National Institutes of Health, under Grant R01LM010207 and R01LM010685, the National Science Foundation, under Grants CCF-0424422, CNS-0964063, and SCH1418504 and the Office of the National Coordinator for Health IT, under Grant HHS-90TR0003/01.

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.2015.03.011.

References

[1] H.U. Prokosch, T. Ganslandt, Perspectives for medical informatics – reusing the electronic medical record for clinical research, Methods Inf. Med. 48 (2009) 38–44.
[2] D.C. Ramick, Data warehousing in disease management programs, J. Healthc. Inf. Manage. 15 (2001) 99–105.
[3] M. Lang, N. Kirpekar, T. Bürkle, S. Laumann, H.U. Prokosch, Results from data mining in a radiology department: the relevance of data quality, Medinfo 12 (2007) 576–580.
[4] J.A. Lyman, K. Scully, J.H. Harrison Jr., The development of health care data warehouses to support data mining, Clin. Lab. Med. 28 (2008) 55–71.
[5] G. Tusch, M. Müller, K. Rohwer-Mensching, K. Heiringhoff, J. Klempnauer, Data warehouse and data mining in a surgical clinic, Stud. Health Technol. Inform. 77 (2000) 784–789.
[6] M. Silver, T. Sakata, H.C. Su, C. Herman, S.B. Dolins, M.J. O'Shea, Case study: how to apply data mining techniques in a healthcare data warehouse, J. Healthc. Inf. Manage. 15 (2001) 155–164.
[7] M. Lang, T. Bürkle, S. Laumann, H.U. Prokosch, Process mining for clinical workflows: challenges and current limitations, Stud. Health Technol. Inform. 136 (2008) 229–334.
[8] R.D. Melamed, H. Khiabanian, R. Rabadan, Data-driven discovery of seasonally linked diseases from an Electronic Health Records system, BMC Bioinformatics 15 (Suppl. 6) (2014) S3.
[9] D.C. Crawford et al., eMERGEing progress in genomics – the first seven years, Front. Genet. (2014), http://dx.doi.org/10.3389/fgene.2014.00184.
[10] Y. Chen, S. Nyemba, B. Malin, Detecting anomalous insiders in collaborative information systems, IEEE Trans. Dependable Secure Comput. 9 (2012) 332–344.
[11] Y. Chen, S. Nyemba, B. Malin, Auditing medical record accesses via healthcare interaction networks, Proc. AMIA Symp. (2012) 93–102.
[12] Y. Chen, N. Lorenzi, S. Nyemba, J.S. Schildcrout, B. Malin, We work with them? Health workers interpretation of organizational relations mined from electronic health records, Int. J. Med. Inform. 83 (2014) 495–506.
[13] F. Muranaga, I. Kumamoto, Y. Uto, Development of site data warehouse for cost analysis of DPC based on medical costs, Methods Inf. Med. 46 (2007) 679–685.
[14] L.M. Etheredge, Rapid learning: a breakthrough agenda, Health Aff. 33 (2014) 1155–1162.
[15] M. Weiner, P. Embi, Toward reuse of clinical data for research and quality improvement: the end of the beginning, Ann. Int. Med. 151 (2009) 359–360.
[16] N. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (2013) 144–151.
[17] T.A. Lasko, J.C. Denny, M.A. Levy, Computational phenotype discovery using unsupervised feature learning over noisy sparse, and irregular clinical data, PLoS ONE 8 (2013) e66341.
[18] J.C. Ho, J. Ghosh, S. Steinhubl, W. Stewart, J.C. Denny, B. Malin, J. Sun, Limestone: high-throughput candidate phenotype generation via tensor factorization, J. Biomed. Inform. 52 (2014) 199–211.
[19] C.L. Overby, J. Pathak, O. Gottesman, A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury, J. Am. Med. Inform. Assoc. 20 (2013) e243–e252.
[20] J.S. Schildcrout, M. Basford, J. Pulley, D.R. Masys, D.M. Roden, D. Wang, et al., An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records, J. Biomed. Inform. 43 (2010) 914–923.
[21] J. Pathak, K.R. Bailey, C.E. Beebe, S. Bethard, D.C. Carrell, P.J. Chen, et al., Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium, J. Am. Med. Inform. Assoc. 20 (2013) e341–e348.
[22] D.A. Springate, E. Kontopantelis, D.M. Ashcroft, I. Olier, R. Parisi, E. Chamapiwa, D. Reeves, ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records, PLoS ONE 9 (2014) e99825.
[23] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, et al., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, Bioinformatics 26 (2010) 1205–1210.
[24] A.N. Kho et al., Electronic medical records for genetic research: results of the eMERGE consortium, Sci. Transl. Med. 3 (79) (2011) 79re1.
[25] J.C. Denny et al., Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data, Nat. Biotechnol. 31 (12) (2013) 1102–1110.
[26] A.K. Jha, C.M. DesRoches, E.G. Campbell, K. Donelan, S.R. Rao, T.G. Ferris, et al., The use of electronic health records in U.S. hospitals, N Engl. J. Med. 360 (2009) 1628–1638.
[27] A. Takian, D. Petrakaki, T. Cornford, A. Sheikh, N. Barber, Building a house on shifting sand: methodological considerations when evaluating the implementation and adoption of national electronic health record systems, BMC Health Services Res. 12 (1) (2012) 105.
[28] J.Y. Sun, Y.G. Fang, Cross-domain data sharing in distributed electronic health record systems, IEEE Trans. Parallel Distrib. Syst. 21 (2010) 754–764.
[29] J. Pathak, A.N. Kho, J.C. Denny, Electronic health records-driven phenotyping: challenges, recent advances, and perspectives, J. Am. Med. Inform. Assoc. 20 (2013) e206–e211.
[30] J.F. Ludvigsson, J. Pathak, S. Murphy, Use of computerized algorithm to identify individuals in need of testing for celiac disease, J. Am. Med. Inform. Assoc. 20 (2013) e306–e310.
[31] P. Tanpowpong, S. Broder-Fingert, J.C. Obuch, D.O. Rahni, A.J. Katz, D.A. Leffler, et al., Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease, Ann. Epidemiol. 23 (2013) 136–142.
[32] P.M. Coloma, V.E. Valkhoff, G. Mazzaglia, M.S. Nielsson, L. Pedersen, M. Molokhia, et al., Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries, BMJ Open. 3 (2013) e002862.
[33] K.M. Newton, P.L. Peissig, A.N. Kho, S.J. Bielinski, R.L. Berg, V. Choudhary, et al., Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, J. Am. Med. Inform. Assoc. 20 (2013) e147–154.
[34] P. Peissig, V. Santos, Costa, M.D. Caldwell, C. Rottscheit, R.L. Berg, E.A. Mendonca, D. Page, Relational machine learning for electronic health record-driven phenotyping, J. Biomed Inform. 52 (2014) 260–270.
[35] J. Zhou, F. Wang, J. Hu, J. Ye, From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 135–144.
[36] J. Ho, J. Ghosh, J. Sun, Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 115–124.
[37] S. Gupta, C. Hanson, C. Gunter, M. Frank, D. Liebovitz, B. Malin, Modeling and detecting anomalous topic access, in: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, 2013, pp. 100–105.
[38] L. Bouarfa, J. Dankelman, Workflow mining and outlier detection from clinical activity logs, J. Biomed. Inform. 45 (2012) 1185–1190.
[39] Z.X. Huang, W. Dong, L. Ji, C.X. Gan, X.D. Lu, H.L. Duan, Discovery of clinical pathway patterns from event logs using probabilistic topic models, J. Biomed. Inform. 47 (2014) 39–57.
[40] H. Zhang, S. Mehotra, C. Gunter, D. Liebovitz, B. Malin, Mining deviations from patient care pathways via electronic medical record system audits, ACM Trans. Manage. Inform. Syst. 4 (2014) 17.
[41] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[42] D. Newman, A. Asuncion, P. Smyth, M. Welling, Distributed inference for latent Dirichlet allocation, in: Proceedings of Neural Information Processing Systems (NIPS), 2007, pp. 1–9.
[43] J. Chuang, S. Gupta, C.D. Manning, J. Heer, Topic model diagnostics: assessing domain relevance via topical alignment, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 612–620.
[44] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[45] H.W. Kuhn, The Hungarian Method for the assignment problem, Naval Res. Logist. Quart. 2 (1955) 83–97.
[46] L.I. Lin, A concordance correlation coefficient to evaluate reproducibility, Biometrics 45 (1989) 255–268.
[47] D. Giuse, Supporting communication in an integrated patient record system, AMIA Annu Symp Proc (2003) 1065.
[48] M. Chan, P.L. Lim, A. Chow, M.K. Win, T.M. Barkham, Surveillance for Clostridium difficile infection: ICD-9 coding has poor sensitivity compared to laboratory diagnosis in site patients, PLoS ONE 6 (2011) e15603.
[49] E.B. Deych, A.D. Waterman, Y. Yan, D.S. Nilasena, M.J. Radford, B.F. Gage, Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors, Med. Care. 43 (2005) 480–485.
[50] E. Perkins, J.R. Williams, Generalized spatial binning of bodies of different sizes, Discrete Element Methods (2002) 52–55.