Discovering De Facto Diagnosis Specialties

Xun Lu[*]
Dept. of Computer Science
University of Illinois at
Urbana-Champaign
xunlu2@illinois.edu

Aston Zhang[*]
Dept. of Computer Science
University of Illinois at
Urbana-Champaign
lzhang74@illinois.edu

Carl A. Gunter
Dept. of Computer Science
University of Illinois at
Urbana-Champaign
cgunter@illinois.edu

Daniel Fabbri
Dept. of Biomedical
Informatics
Vanderbilt University
daniel.fabbri@vanderbilt.edu

David Liebovitz
Dept. of Medicine
Northwestern University
davidl@northwestern.edu

Bradley Malin
Dept. of Biomedical
Informatics
Vanderbilt University
b.malin@vanderbilt.edu

ABSTRACT

In health care institutions, medical specialty information may be lacking or inaccurate, in part because there is no official code to express such specialties. Diagnosis histories offer information on which medical specialties may exist in practice, regardless of whether they have official codes. We refer to such specialties that are predicted with high certainty by diagnosis histories *de facto* diagnosis specialties. The objective of our research is to discover *de facto* diagnosis specialties under a general discovery–evaluation framework. Specifically, we employ a semi-supervised learning model (based on heterogeneous information network analysis) and an unsupervised learning method (based on topic modeling) for discovery. We further employ four supervised learning models for evaluation. We use one year of diagnosis histories from a major medical center, which consists of two data sets. One is fine-grained and has diagnoses assigned to 41,603 patients that are accessed by 2,504 medical service providers. The other is general and has diagnoses assigned to 291,562 patients that are accessed by 3,269 medical service providers. The semi-supervised learning model discovers a specialty for *Breast Cancer* on the fine-grained data set; while the unsupervised learning method confirms this discovery and suggests another specialty for *Obesity* on the larger general data set. The evaluation results reinforce that these two specialties can be recognized accurately by supervised learning models in comparison with 12 common diagnosis specialties defined by the Health Care Provider Taxonomy Code Set.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Medical information system; H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—Data mining

General Terms

Algorithm, Experimentation

Keywords

Data Mining, Electronic Health Records, Medical Informatics

1. INTRODUCTION

Medical specialties provide information about which medical service providers (hereinafter referred to as "providers") have the skills needed to carry out key procedures or make critical judgments. They are useful for training and staffing, as well as providing confidence to patients that their providers have the expertise required to address their problems.

Health care institutions have many ways to express and take advantage of staff specialties, including organizing them into departments or wards. However, such an organization has its limitations. For instance, at a large and diverse medical center, some specialties may be lacking or inaccurately described (*e.g.*, they are not always entered for new hire documents), employees can change roles, and encoded departments do not always align with specialties. As a result, there could be a gap between the diagnosis histories of certain providers and their specialties. There is thus an opportunity to design and apply data-driven techniques that assist in the management of health care operations, such as staffing (by providing accurate specialty information about current staff), quality control (by verifying that providers practice consistently with their declared specialties), and building patient confidence (by ensuring that patients are treated by specialists) [16].

Health care providers select from the Health Care Provider Taxonomy Code Set (HPTCS) [14] when they apply for their National Provider Identifiers (NPIs) [1]. NPIs are required by the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and are used in health care-related transactions. Health care providers usually choose their taxonomy codes according to the certifications they hold. Ideally, this mechanism would identify each health care provider with the taxonomy codes that most accurately describe their specialties. However, this is not always the case for several reasons.

[*] These two authors contributed equally.

BCB'15, September 09 - 12, 2015, Atlanta, GA, USA
ACM 978-1-4503-3853-0/15/09...$15.00
DOI: http://dx.doi.org/10.1145/2808719.2808720

First, the National Plan & Provider Enumeration System does not verify that the taxonomy code selections made by providers in NPI applications are accurate [14]. Second, certain taxonomy codes do not correspond to any nationwide certifications that are approved by a professional board. For example, the specialty for *Men and Masculinity* is a well-recognized area of interest, study, and activity in the field of psychology; however, there is no certification or credential available to identify psychologists who might work in this area [2]. Third, some national certifications are not reflected by the taxonomy code list. Since the taxonomy codes do not correspond to certifications within the field, providers may interpret these codes in inconsistent ways.

In view of the aforementioned limitations of purely relying on NPI taxonomy codes, we propose to leverage real-world diagnosis histories to infer and recognize actual specialties. We refer to such inferred knowledge as *de facto* specialties. *De facto* specialties are medical specialties that exist in practice regardless of the specialty codes (NPI taxonomy codes). To illustrate, imagine that there is a method for recognizing providers' *de facto* specialties based on their actual activities related to diagnosis histories. This enables us to verify the NPI taxonomy codes of the providers in a health care institution. If certain providers' declared specialties failed to match their activity-based specialties, such as electronic health record (EHR) access, an investigation and possible re-designation of their codes might be warranted.

As the medical profession evolves, the HPTCS needs to be updated to be more comprehensive [5, 10, 25]. Problems and inefficiencies could arise if the specialty codes are not sufficiently expressive to convey providers' specialties. For instance, if there is no official code to express such specialties and no providers declared them, false alarms of suspicious EHR access detection might be raised because such unlisted *de facto* specialties could not be assigned to any providers. Other concerns have been voiced by the American Psychological Association: " *... several national certifications that do exist are not reflected on the specialty code list. Since the specialty codes do not correspond to certifications within the field, psychologists will interpret these codes in different ways. Use of the specialty codes by psychologists therefore will not be uniform and will not provide meaningful information about a psychologist's practice.*" [2]

The focus of our research is on ***de facto* diagnosis specialties** of providers that exist in practice and *are highly predictable by the diagnoses in the EHRs of the patients they treat.*. Our goal is to discover *de facto* diagnosis specialties that do not have corresponding codes in the Health Care Provider Taxonomy Code Set. In this study, we use a subset of such codes for both discovery and evaluation. To provide intuition into the problem, let us consider a perfect scenario where *every* NPI code correctly reflects specialties in a data set. If machine learning models are trained on this data set and exhibit decent performance, we believe that such models would reliably discover *de facto* diagnosis specialties in a new data set; the new data set may be provided by another health care institution that needs more reliable *de facto* diagnosis specialty discovery. However, in practice this perfect scenario will not be realized. In this work, we consider a more challenging scenario where we assume that *majority* of the NPI codes correctly reflect specialties in our collected data set.

This study makes three contributions. First, we propose a novel *de facto* diagnosis specialty discovery problem. To solve this problem, we introduce a discovery–evaluation framework. Specifically, *de facto* diagnosis specialties are proposed and their recognition accuracy is subsequently evaluated in comparison with existing diagnosis specialties listed in the HPTCS. Although we rely on expert opinions to interpret our discovery results, we consider evaluation important because expert opinions may not always be available in practice.

Second, under the discovery–evaluation framework, we employ a semi-supervised learning model (based on heterogeneous information network analysis) on a fine-grained data set and an unsupervised learning method (based on topic modeling) on a larger general data set for discovery. We further employ four supervised learning models for evaluation. Details of the two data sets are described in Section 3.

Third, we perform an empirical investigation using one year of diagnosis histories from a major medical center, which consists of two data sets. One is fine-grained and has diagnoses assigned to 41,603 patients that are accessed by 2,504 providers. The other is general and has diagnoses assigned to 291,562 patients that are accessed by 3,269 providers. The semi-supervised model discovers a *de facto* diagnosis specialty for *Breast Cancer* on the fine-grained data set; the unsupervised learning method confirms this discovery and suggests a new *de facto* diagnosis specialty for *Obesity* on the larger general data set. The evaluation results reinforce that these two specialties can be recognized accurately by supervised learning models in comparison with 12 common *de facto* diagnosis specialties defined by HPTCS.

2. DE FACTO DIAGNOSIS SPECIALTY

In Section 1, we define *de facto* diagnosis specialties as medical specialties that exist in practice and are highly predictable by the diagnoses inherent in EHRs. Here we illustrate this concept in more detail.

Intuitively, it should be easier to characterize a urologist in terms of medical diagnoses for conditions, for example, of the kidney, ureter, and bladder, as opposed to an anesthesiologist, whose duties are more cross-cutting with respect to diagnoses, concerning essentially all conditions related to surgeries. To orient the reader using a concrete example, let us test this hypothesis with a naïve classifier based on diagnosis codes. To gain intuition into the general idea, consider the following steps. First, we begin with a data set that indicates which EHRs have been accessed by urologists and anesthesiologists, and view each patient as a document whose words are diagnoses in their EHRs. Next, we create a weighting for how many diagnoses of each kind are accessed by each provider, with some adjustment for how common the diagnosis is. This technique is typified by term frequency–inverse document frequency (TF-IDF, with details in Section 4.4.5). We believe such a naïve classifier is the type of model that an administrator might define and apply to infer a specialty from a diagnosis history. The technique proceeds by finding the most relevant diagnoses of each di-

Table 1: **A summary of the attributes for NMH audit logs for the fine-grained and general data sets.**

	Fine-Grained	**General**
Accesses	35,869	4,829,376
Patients	41,603	291,562
Providers	2,504	3,269
Patient encounters	62,390	890,812
Taxonomy codes	161	165

Table 2: **A summary of the attributes for patient records in NMH audit logs for the fine-grained and general data sets.**

	Fine-Grained	**General**
Provider job titles	167	171
Locations	242	251
Services	101	104
Diagnoses	4,172	13,566
Procedures	740	2,165

agnosis specialty (taxonomy code) and the most frequently accessed diagnoses of each provider. Finally, providers are classified according to the specialties with which they share the most commonly accessed diagnoses.

Using the general data set for the empirical study below (details in Section 3), we observe that urologists tend to access diagnoses such as "retention of urine" and "urinary tract infection", whereas anesthesiologists tend to access diagnoses such as "other aftercare" and "other screening". When we use the 20 conditions most accessed by either of the two specialties as the features for the naïve classifier, the results are decent for urology, yielding an F_1 score of 70.35% in predicting the urologists[1]. However, the results for anesthesiologists are poorer, yielding F_1 score of 11.30%. If we use a machine learning technique, such as SVM (described in Section 4.4.4), we can achieve substantially better results: finding anesthesiologists with an F_1 score of 48.98%. However, this performance is still weaker than the classifier learned for urologists, which achieves an F_1 score of 97.44%.

De facto specialties that are highly predictable by diagnosis histories are *de facto* diagnosis specialties. Note that there is no ground truth to determine the validity of a discovered *de facto* diagnosis specialty. Ideally, a discovered *de facto* diagnosis specialty can be recognized by classifiers as accurately as the existing listed diagnosis specialties. To illustrate how this is possible, consider an analogy with respect to the classification of documents, an area that has inspired many of the techniques we apply. The providers U can be likened to readers of documents, where A represents an archive of documents in which the words in each document correspond to diagnoses. A function $T(u)$ indicates the collection of documents that a provider u has read. Providers with specialties are groups of readers who (presumably) have a common *de facto* diagnosis specialty and interest in the same group. To solve the *de facto* diagnosis specialty discovery problem we aim to develop a classifier that characterizes this common interest in terms of the documents that they have read, if possible. For instance, if there are a group of readers that are ophthalmologists and they are inordinately interested in documents on disorders of the eyes, then we can use this proclivity to serve as a discriminatory feature.

3. DATA

Following the aforementioned analogy to the document classification, we use access log data from a hospital and combine it with the diagnosis lists in patient discharge records. That is, for each encounter (visit to the hospital by a patient) we have a set of diagnoses, and for each provider we have a record of whether the provider accessed the chart of that patient during the time of that encounter. If a provider u accessed the patient during that encounter, we include the diagnosis set for that encounter in $T(u)$. We will refer to users (as in chart users) rather than providers for our technical discussion.

We collect data for this study via the Cerner Powerchart EHR system in use at Northwestern Memorial Hospital (NMH). The data contain all user accesses (in the form of audit logs) made over a one-year period, as well as insurance billing code lists, in the form of International Classification of Diseases–ninth revision (ICD-9), for patient encounters during this period. All data were de-identified for this study in accordance with the HIPAA Privacy Rule and carried out under Institutional Review Board approval. Since specialties are mainly focused on physicians, we filter out users with other positions (*e.g.*, nurses and dieticians) from the data set.

A small portion of the collected data has an explicit mapping between users and diagnoses of the EHRs they accessed. However, majority of the data lacks such an explicit relationship. This is because patients may have multiple diagnoses and their EHRs may be accessed by different users without documentation on which specific diagnoses were associated with the actions of which user. We refer to the former portion as the *fine-grained data set*. As fine-grained data may not always be available, we expand to a more general data set for our study that may be more representative of the challenging scenarios encountered in practice. Hence, we use the entire data after removing all such fine-grained mapping information to form the other data set, which we call it the *general data set*. The attributes of the data sets used in this study are summarized in Table 1—2.

We use Clinical Classifications Software (CCS) to cluster diagnosis and procedure codes into a manageable number of clinically meaningful categories [11]. This is because ICD-9 codes are not completely indicative of patients' clinical phenotypes [3] and the sheer number of codes (on the order of 10,000) makes it challenging to characterize patterns of diagnoses or procedures. The ICD-9 codes for diagnoses are mapped down to 603 CCS codes and the ICD-9-CM codes for procedures are mapped down to 346 CCS codes. A key characteristic of the data set relevant to this study is that it also contains NPI taxonomy codes for 60% of the providers. About 150 classes of NPI taxonomy codes are listed in the data sets, but most have fewer than 10 user instances. Figure 1 shows the frequency distribution of 100

[1] A higher F_1 score indicates a better performance (more details are provided in Section 5.1).

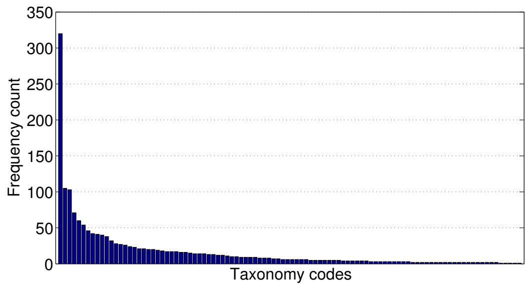

Figure 1: **The frequency distribution for the 100 most frequent taxonomy codes in the general data set.**

most frequent taxonomy codes in the data set.

To ensure there is a sufficient amount of data to train machine learning models, we filter out NPI taxonomy codes with fewer than 20 user instances [18]. Based on the guidance of several clinicians and hospital administrators, we further identify 12 NPI taxonomy codes as diagnosis specialties: *Obstetrics & Gynecology, Cardiovascular Disease, Neurology, Ophthalmology, Gastroenterology, Dermatology, Orthopaedic Surgery, Neonatal-Perinatal Medicine, Infectious Disease, Pulmonary Disease, Neurological Surgery*, and *Urology*. We refer to this group as the *core NPI taxonomy codes*. As discussed in Section 1, we assume that a majority of these codes correctly reflect specialties in the data.

4. METHODS

4.1 Discovery–Evaluation

We highlight that there is no ground truth for the *de facto* diagnosis specialty discovery problem. Hence, we address this challenge under a general discovery–evaluation framework.

4.1.1 Discovery

We invoke machine learning to discover potential *de facto* diagnosis specialties in the data set that lack corresponding codes in the HPTCS. In this study, we first employ a semi-supervised learning model (in the form of PathSelClus [29]) to leverage the mapping between users and their specifically accessed diagnoses of EHRs in the fine-grained data set. Then we consider a more challenging scenario where such fine-grained mapping is not available. In this case, we employ an unsupervised learning model (in the form of Latent Dirichlet Allocation [4]) for discovery in the larger general data set. Since the fine-grained data set is a subset of the general data set, except for the fine-grained mapping information, the discovery results can be reinforced when they exhibit common findings.

4.1.2 Evaluation

To interpret the discovery results, we rely on expert opinions. However, we acknowledge that in practice such opinions may not be available. Hence, we also make use of supervised learning models to evaluate the recognition accuracy of the discovered specialty by comparing our approach with the existing listed diagnosis specialties, such as the core NPI taxonomy codes described in Section 3. Ideally, their recognition accuracy should be similar. In this study, we evaluate such recognition accuracy using four classifiers, namely, decision trees, random forests, PCA-KNN, and SVM.

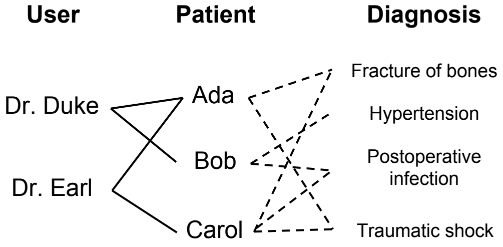

Figure 2: **A toy example for visualizing the data set in the view of a heterogeneous information network. There are multiple types of nodes, such as users, patients and diagnoses; and multiple types of links between different types of nodes.**

4.2 PathSelClus for Discovery

In general, discovering *de facto* diagnosis specialties from the diagnosis histories of providers may rely upon effective clustering techniques that can divide a pool of providers into groups that have high inter-group distances (distinctiveness), but low intra-group distances (coherence). We anticipate that, new diagnosis specialties may emerge from these clusters. The structure of our data sets can be represented as a typical *heterogeneous information network* [28, 29, 31]. Therefore, we use PathSelClus [29], a state-of-the-art semi-supervised learning model based on heterogeneous information networks for user-guided clustering. For context, we begin with a brief introduction to heterogeneous information networks.

4.2.1 Heterogeneous Information Networks

A heterogeneous information network consists of multiple types of objects and/or multiple types of links. A heterogeneous information network explicitly distinguishes between object types and relationship types in the network, which is quite different from traditional networks. For example, if a relation exists from type A to type B, denoted as ARB, then the inverse relation R^{-1} holds naturally for $BR^{-1}A$. R and its inverse R^{-1} are usually not equal, unless the two types are the same and R is symmetric.

Figure 2 depicts our data in the form of a heterogeneous information network and the corresponding schema. It contains 3 types of objects, namely user (U), patient (P) and diagnosis (D). Links exist between users and patients by the relation of “access” and “accessed by”; links exist between patient and diagnosis by the relation of “diagnosed with” and “assigned to”.

Link-based clustering in heterogeneous information networks groups objects based on their connections to other objects in the networks. The possible relations derived from a heterogeneous information network between two types of objects in a meta-level is called a *meta-path* [27]. In our case, the *target object type* to cluster is U (users). There are two meta-paths: $U \xrightarrow{access} P \xrightarrow{accessed\ by} U$ and $U \xrightarrow{access} P \xrightarrow{diagnosed\ with} D \xrightarrow{assigned\ to} P \xrightarrow{accessed\ by} U$.

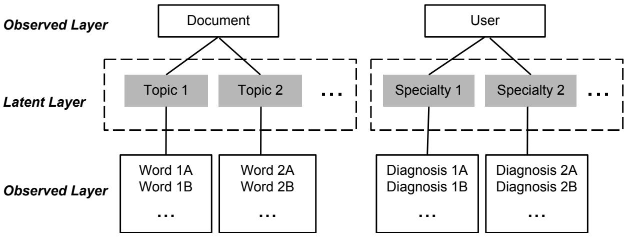

Figure 3: **An analogy of the *User–Specialty–Diagnosis* hierarchy in a *de facto* diagnosis specialty discovery problem to the *Document–Topic–Word* hierarchy in a topic modeling problem.**

4.2.2 *User-Guided Clustering*

During clustering, a decision has to be made about the weighted combination of different meta-paths to use. This is where user guidance comes into play. We use the semi-supervised learning model PathSelClus for user-guided clustering. In PathSelClus, user guidance is provided in the form of object seeds in each cluster. For example, to cluster users based on the pattern of the diagnoses of EHRs they access, one can provide several representative users as seeds for each pattern. These seeds provide guidance for clustering the target objects in the heterogeneous information networks and help select the most relevant meta-paths for the clustering task.

PathSelClus is designed to handle unseeded initial clusters because in practice, there may not be sufficient information to seed all the clusters. This is the exact feature that makes it possible to use PathSelClus to discover new diagnosis specialties. Now, let the number of listed diagnosis specialties be N and the number of *de facto* diagnosis specialties we want to discover be δ. We create $N + \delta$ empty clusters at the initiation of PathSelClus and seed N of them with corresponding specialists. The inputs of PathSelClus include all the users regardless of whether they have a taxonomy code.

As an output, each user is assigned to the cluster with the highest assignment likelihood. The δ unseeded clusters should also be filled with users. We can analyze the semantics of the unseeded clusters via the users they contain. We treat a cluster as a taxonomy code and calculate the most relevant diagnoses for each cluster. Then the medical expert labels the clusters, which we use to interpret the discovery results.

4.2.3 *Fine-Grained Data Set for PathSelClus*

We emphasize that PathSelClus is a network-based learning model and relies on the mapping between users and their specifically accessed diagnoses in the fine-grained data set. In Section 5.2, we empirically compare and analyze PathSelClus in more detail on both the fine-grained data set and the general data set.

4.3 Latent Dirichlet Allocation (LDA) for Discovery

In practice, fine-grained data sets may not be available for PathSelClus. Hence, we also employ an unsupervised learning method [4], which is based on topic modeling.

4.3.1 *General Data Set for LDA*

In Latent Dirichlet Allocation (LDA) [4], topics act as summaries of the different themes pervasive in the corpus and documents are characterized with respect to these topics. The intuition behind our employment of LDA is diagnosis topics with coherent themes in a hospital. By treating each provider as a document in which the provider's associated diagnoses are the words and applying LDA to model all these documents, we can obtain an allocation of diagnosis topics for each provider. This analogy is illustrated in Figure 3. We can further cluster the providers using their topic allocations by the topic simplex[2] that they are closest to.

LDA does not leverage network information and does not require a fine-grained mapping between users and their accessed diagnoses. Instead, LDA models specialties with respect to different diagnosis themes. In this study, all the data sets to which LDA is applied refer to the larger general data set.

4.3.2 *Representation of Users*

Diagnoses in our data set are provided with respect to patients, but not users. Therefore, we associate users with diagnoses via the patients they access. We consider two approaches for accomplishing this task.

User-document approach: For any user u_i, find the set of patients P_i whose EHR is accessed by u_i. Then, for each patient $p_j \in P_i$, let D_j be the set of diagnoses associated with p_j. We add diagnoses to D_j that occurred during the encounter of u_i and p_j to a set of diagnoses that represent u_i. The diagnosis topics and their allocations for users are discovered directly by applying LDA.

Patient-document approach: In this alternative approach, we start by applying LDA on the patient dimension to obtain a topic distribution in diagnoses for patients rather than users. Let T_{p_j} denote the topic distribution in diagnoses of patient p_j. Let T_{u_i} be the topic distribution of user u_i and let P_i be the set of patients whose EHRs are accessed by u_i. Then, the topic distribution for user u_i is

$$T_{u_i} = \frac{1}{|P_i|} \times \sum_{p_j \in P_i} T_{p_j}.$$

Both approaches were tested on the general data set. Table 3 shows one sample topic summary for both approaches. It is notable that the topic obtained from the user-document approach exhibits no clear theme, whereas the topic obtained from the patient-document approach has a consistent theme related to Urology. This is due to the fact that, in the user-document approach, each document contains the union of the diagnoses of all the accessed patients, whereas in the patient-document approach only the diagnoses of a single patient are in the document. The hodgepodge of many patients' diagnoses is likely to contain diverse and inconsistent themes, thus rendering the topics generated by the user-document approach not easily interpretable. Since discovering *de facto* diagnosis specialties requires experts to interpret such topics, we use the patient-document approach.

[2] This can be visualized by plotting the providers by their topic distributions.

Table 3: **A comparison of two sample *de facto* diagnosis specialties obtained by two different LDA approaches on the general data set. They are represented by 10 most probable diagnoses according to LDA. The user-document approach obtains more semantically random diagnoses, whereas the patient-document approach obtains a specialty with diagnoses consistent with a Urology theme.**

(a) User-document approach

Other hypertensive complications
Hypotension
Cancer of ovary
Coma, stupor, and brain damage
Hyposmolality
Ascites
Hematuria
Acute myocardial infarction
Backache, unspecified
Other connective tissue disease

(b) Patient-document approach

Calculus of kidney
Elevated prostate specific antigen
Hematuria
Impotence of organic origin
Incomplete bladder emptying
Bladder neck obstruction
Urinary frequency
Hydronephrosis
Unspecified retention of urine
Other testicular hypofunction

4.3.3 *Choice of Topic Number*

An important parameter for LDA is the number of topics k. There is no consensus on how to determine the best value of k. The sign of a good topic number is that the resulting topic summaries are semantically meaningful. The general rule for picking k is the perplexity measure [4]. This is an estimate of the expected number of equally likely words. Minimizing perplexity corresponds to maximizing the captured topic variance. Based on the perplexity measure, k is set to 30 in this study.

4.3.4 *Clustering Users*

After applying LDA, each user is assigned to an allocation in the specialty topic simplex. A higher frequency in a specialty indicates that the user is more likely to access patients with diagnoses popular in that specialty. Therefore, if we cluster users by *de facto* diagnosis specialties, it is reasonable to cluster users by the closest specialties. This is because this specialty has the highest proportion in the specialty topic simplex:

$$C_{u_i} = \operatorname*{argmax}_{t \in T} P(u_i, t),$$

where C_{u_i} denotes the specialty cluster assignment for the user u_i and T denotes the set of specialty topics, and $P(u_i, t)$ denotes the proportion of the topic t for the user u_i.

4.4 Classifiers for Evaluation

In PathSelClus, a *de facto* diagnosis specialty is represented by the most accessed diagnoses by all users in the same cluster that have such a specialty. In LDA, a *de facto* diagnosis specialty is represented by the most probable diagnoses as an output of the model. To interpret the discovered *de facto* diagnosis specialties, we rely on physicians (authors) with medical expertise. The experts reviewed the diagnosis summaries of the specialty and labeled each with one or a few medical themes that are pervasive in the specialty. After labeling, we compare the labeled specialties with the HPTCS to see if there are specialties that have pervasive themes but are not listed in the code set. If such specialties exist, they are considered to be potential newly discovered *de facto* diagnosis specialties. Since there is no ground truth for the discovery results, we use supervised learning models to evaluate the recognition accuracy of the discovered *de facto* diagnosis specialty. We briefly describe the four classifiers used in this study.

4.4.1 *Decision Trees*

A decision tree (J48) is constructed in a top-down recursive divide-and-conquer manner. To start, all the training examples are at the root. Examples are partitioned recursively based on selected attributes. Test attributes are selected on the basis of a heuristic or statistical measure. A decision tree is a popular nonlinear classifier because it is convertible to classification rules that can be reviewed and interpreted by experts.

4.4.2 *Random Forests*

To aggregate decision trees, we can use random forests. To do so, for $b \in \{1, \ldots, B\}$, we draw samples from the training data and grow a big tree T_b with some restrictions: at each split, randomly select m features from the p features and pick the best split among them. The recommended value (used in this study) for m is $\sqrt{p}$. Then the forests are represented as a collection of trees $\{T_b\}_{b=1}^{B}$. To classify a testing instance, we conduct majority voting among $T_1(x), ..., T_B(x)$.

4.4.3 *KNN-PCA*

K-Nearest Neighbors (KNN) is an instance-based learning method. It stores training examples and delays the processing until a new instance must be classified. All instances correspond to points in the n-dimensional space. The nearest neighbors are defined in terms of Euclidean distance. KNN returns the most common label among the K training examples nearest to the new testing instance. KNN is sensitive to the "curse of dimension" such that the distance between neighbors could be dominated by irrelevant attributes when the dimension of space goes higher. To mitigate this problem, we use principal component analysis (PCA) by selecting a small number of the principal components to perform dimension reduction.

4.4.4 *SVM*

A support vector machine (SVM) is a classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data set into a higher

Table 4: **Three inconsistent *de facto* diagnosis specialties are obtained by PathSelClus when the number of unseeded clusters δ is set to 3 on the general data set. They are represented by the top 10 most accessed diagnoses by all the users that are in each cluster respectively. None shows a consistent theme with respect to a specialty.**

Other bacterial infections	Chronic kidney disease	Abdominal pain
Other non-traumatic joint disorders	Essential hypertension	Other and unspecified lower respiratory disease
Convulsions	Other cardiac dysrhythmias	Nonspecific chest pain
Other upper respiratory disease	Abdominal pain	Urinary tract infection; site not specified
Phlebitis and thrombophlebitis	Phlebitis and thrombophlebitis	Diabetes mellitus without complication
Malaise and fatigue	Other fluid and electrolyte disorders	Essential hypertension
Other skin disorders	Anemia; unspecified	Other nervous system symptoms and disorders
Fever of unknown origin	Pleurisy; pleural effusion	Pneumonia; organism unspecified
Cardiomyopathy	Acute renal failure	Phlebitis and thrombophlebitis
Substance-related disorders	Hyperpotassemia	Other and unspecified circulatory disease

dimension. We used a Gaussian kernel in this study. The SVM searches for the optimal linear separating hyperplain in this new space by using support vectors that lie closest to the decision boundary. In particular, SVM is effective on a high-dimensional data set because the complexity of the trained classifier is characterized by the number of support vectors rather than the dimension of the data set.

4.4.5 *Classification*

To apply these classifiers to our data, we map each user u_i in the set of users U to a TF-IDF weighted diagnosis vectors $v_i' = \{\mathsf{tfidf}_{\mathsf{d}_1}, ..., \mathsf{tfidf}_{\mathsf{d}_\mathsf{k}}\}$ according to:

$$\mathsf{tfidf}_{d_j} = \log\left(\frac{v_i(d_j)}{a_i} + 1\right) \times \log\left(\frac{|U| + 1}{r_{d_j}}\right),$$

where d_j is the diagnosis with the globally unique index j, and each user u_i has a vector $v_i = \{c_1, ..., c_k\}$ where c_j denotes the number of times that the user has accessed patients with d_j. Let a_i be the total count of all diagnoses in v_i, and let r_{d_j} be the number of users that have accessed patients with d_j. This vector, along with each user u_i's primary taxonomy code, serves as the input to these classifiers, with a length of 603. For KNN-PCA, we perform dimension reduction via PCA to the vectors before applying KNN. We do not use procedure codes because they are less expressive than diagnosis codes [20].

5. EXPERIMENT

5.1 Setup and Evaluation Measures

We use Weka [17] for decision trees (J48), random forests, and SVM with the default parameter values. In PCA-KNN, the number of nearest neighbors K is set to 9 with 50 principal components, based on a cross-validation tuning process [9].

In the evaluation stage, we use precision, recall, and F_1 score to assess performance. For a specialty s, the true positive count $TP(s)$ is the number of users with the specialty s that are correctly classified. The false positive count $FP(s)$ is the number of users with a specialty other than s that are classified as s. The false negative $FN(s)$ count is the number of users with the specialty s that are wrongly classified. The precision P for a specialty s is computed as $\frac{TP(s)}{TP(s)+FP(s)}$ and the recall R is $\frac{TP(s)}{TP(s)+FN(s)}$. The precision of a classifier is the weighted average of precision for each specialty; the weight for a specialty s is the ratio of the number of users with s to the total number of users. The recall of a classifier is defined similarly. The F_1 score is the harmonic mean of the precision (P) and recall (R): $F_1 = \frac{2PR}{P+R}$. We use 5×2 cross-validation for evaluation with classifiers. In each of the 5 rounds, observations are split into two equal-sized sets A and B. Then a classifier is trained on A and tested on B and *vice versa*. After 5 rounds, the average of the 10 results is reported.

Table 5: **The *de facto* diagnosis specialty Breast Cancer is discovered by PathSelClus. It is represented by the top 10 most accessed diagnoses by all the users that are associated with the Breast Cancer specialty.**

Lump or mass in breast
Diffuse cystic mastopathy
Galactorrhea not associated with childbirth
Benign neoplasm of breast
Unspecified breast disorder
Abnormal mammogram, unspecified
Malignant neoplasm of upper-inner quadrant of female breast
Benign neoplasm of lymph nodes
Personal history of malignant neoplasm of breast
Other sign and symptom in breast

5.2 Results for PathSelClus

In Section 4.2.3, we mentioned PathSelClus relies on the mapping between users and their specifically accessed diagnoses of EHRs in the fine-grained data set. Table 4 shows inconsistent *de facto* diagnosis specialties by PathSelClus when the number of unseeded clusters δ is set to 3. None exhibits a consistent theme with respect to a specialty and it remains the same when δ is set to other values.

One reason why PathSelClus leads to inconsistent themes is that the general data set does not contain the aforementioned fine-grained mapping information. As a consequence, all of the diagnoses that belong to patients can be mapped to users that access such patients. We observe that a patient can have multiple encounters, such as delivering a baby and returning several months later due to a infectious disease. Therefore, in the general data set, clustering users based on all the diagnoses of their accessed patients may not be accurate (as shown in Table 4).

Table 6: **Average accuracy of multi-class classification on the fine-grained data set under 5×2 cross-validation (in percent). Users with the *de facto* Breast Cancer specialty discovered by PathSelClus are in one class; users with core NPI taxonomy codes are in 12 distinct core classes. The boldfaced result with the superscript † denotes that, the F_1 score of the discovered *de facto* Breast Cancer specialty is significantly higher than that of mean of 12 core classes (paired t-test with $p < 0.05$).**

Specialty	**Decision Trees**			**Random Forests**			**PCA-KNN**			**SVM**		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Breast Cancer	86.67	57.14	**68.87**†	89.13	64.16	**74.61**†	77.00	79.09	**78.03**†	92.50	93.11	**92.80**†
Mean of 12 Core Classes	67.37	58.07	62.38	72.08	67.36	69.64	72.30	74.02	73.15	89.30	86.72	87.99
Urology	70.59	60.00	64.86	68.42	65.00	66.67	81.82	90.00	85.71	100.00	95.00	97.44
Neurology	71.05	57.45	63.53	71.05	57.45	63.53	65.57	85.11	74.07	81.48	93.62	87.13
Pulmonary Disease	100.00	54.17	70.27	93.33	58.33	71.79	71.43	83.33	76.92	95.83	95.83	95.83
Orthopaedic Surgery	93.33	48.28	63.64	93.33	48.28	63.64	69.70	79.31	74.19	100.00	89.66	94.55
Neonatal-Perinatal Medicine	87.50	25.00	38.89	89.43	89.29	85.36	92.59	89.29	90.91	96.15	89.29	92.59
Gastroenterology	67.86	50.00	57.58	69.23	47.37	56.25	69.23	94.74	80.00	95.00	100.00	97.44
Obstetrics & Gynecology	42.23	97.25	58.89	49.03	94.50	64.56	87.18	93.58	90.27	98.99	89.91	94.23
Neurological Surgery	100.00	35.00	51.85	100.00	35.00	51.85	33.33	5.00	8.70	100.00	35.00	51.85
Ophthalmology	73.91	40.48	52.31	90.04	71.43	79.66	80.56	69.05	74.36	54.67	97.62	70.09
Cardiovascular Disease	63.93	61.90	62.90	62.12	65.08	63.57	71.95	93.65	81.38	96.83	96.83	96.83
Infectious Disease	79.17	73.08	76.00	79.17	73.08	76.00	63.64	53.85	58.33	89.29	96.15	92.59
Dermatology	78.95	39.47	52.63	78.95	39.47	52.63	71.43	52.63	60.61	100.00	68.42	81.25

P: Precision; R: Recall; F_1: F_1 Score

On the fine-grained data set, PathSelClus discovers a specialty for *Breast Cancer* that does not have a corresponding code in HPTCS, as shown in Table 5 ($\delta = 3$). Setting δ between 1 to 4 generates this discovery although a larger value of δ makes the discovery less clear. In the fine-grained data set, 35 users are found to be associated with the Breast Cancer specialty.

Table 6 summarizes the average accuracy of multi-class classification on the fine-grained data set under 5×2 cross-validation. Users with the *de facto* Breast Cancer specialty discovered by PathSelClus are in one class; users with core NPI taxonomy codes as discussed in Section 3 are in the 12 distinct core classes. The F_1 score of the discovered *de facto* Breast Cancer specialty is significantly higher than that of mean of 12 core classes under all the four classifiers (paired t-test with p $<$ 0.05).

5.3 Results for LDA

With a larger general data set, LDA confirms the discovery of Breast Cancer by PathSelClus and suggests another *de facto* diagnosis specialty for *Obesity* as shown in Table 7. The Breast Cancer and Obesity specialties are found to be associated with 68 and 20 users, respectively.

Tables 8 and 9 summarize the average accuracy of multi-class classification on the general data set under 5×2 cross-validation for the two discovered specialties. The F_1 score of the discovered *de facto* Breast Cancer specialty by LDA is also significantly higher than that of mean of 12 core classes under all the four classifiers, confirming the finding from PathSelClus (paired t-test with $p < 0.05$). The result for Obesity is similar except that PCA-KNN is not statistically significantly better than the other classifiers.

6. RELATED WORK

The discovery of *de facto* diagnosis specialties is critical to managing health care institutions and allocating resources to clinicians. This work shows that such discovery is possible and that existing vocabularies may be insufficient or incomplete. To date, there has been little investigation into automated learning for the *de facto* diagnosis specialty discovery; however, we wish to note that the approaches introduced in this work are related to those that have been developed for health care role prediction and access control management. Here, we take a moment to review relevant work in such areas.

A driver behind inferring medical specialties is the analysis of audit logs for security and privacy purposes [6, 8, 24]. This is feasible because EHRs and their audit logs encode valuable interactions between users and patients [26]. Users have roles in the health care institutions. If these roles are not respected by the online activities of the users, there may be an evidence of a security or privacy violation. An early study on this theme examined the idea of examining accesses to patient records to determine the position of an employee [33]. This work used a Naïve Bayes classifier and had generally poor performance on many positions, often because such positions could not easily be characterized in terms of the chosen attributes. Moreover, Experience Based Access Management envisioned such studies as part of a general effort to understand roles by exploiting information about institutional activities through the study of audit logs [15]. Another study in this direction sought to infer new roles from ways in which employees acted in their positions by iteratively revising existing positions based on experiences [32].

The problem of determining which departments are responsible for treating a given diagnosis was addressed by studies on Explanation-Based Auditing System (EBAS) [13, 12]. They are similar to our problem of identifying an employee's specialty. In these studies the auditing system utilizes the access patterns of departments to determine diagnosis responsibility information in two ways: by analyzing (i) how frequent a department accesses patients with the diagnosis, and (ii) how focused the department is at treating the given diagnosis. For instance, EBAS could use this approach to determine that the Oncology Department is responsible for chemotherapy patients, while the Central Staffing Nursing Department is not. The random topic access model (RTAM) [16] went beyond approaches based on conditional

Table 7: ***De facto* diagnosis specialties Breast Cancer and Obesity are discovered by LDA. They are represented by 10 most probable diagnoses respectively as an output of LDA.**

(a) Breast Cancer

Personal history of malignant neoplasm of breast
Lump or mass in breast
Abnormal mammogram, unspecified
Other specified aftercare following surgery
Other sign and symptom in breast
Carcinoma in situ of breast
Family history of malignant neoplasm of breast
Other specified disorder of breast
Benign neoplasm of breast
Acquired absence of breast and nipple

(b) Obesity

Obesity, unspecified
Morbid obesity
Obstructive sleep apnea
Unspecified sleep apnea
Hypersomnia with sleep apnea, unspecified
Paralysis agitans
Hip joint replacement by other means
Edema
Other dyspnea and respiratory abnormality
Body Mass Index 4

Table 8: **Average accuracy of multi-class classification on the general data set under 5×2 cross-validation (in percent). Users with the *de facto* Breast Cancer specialty discovered by LDA are in one class; users with core NPI taxonomy codes are in the 12 distinct core classes. The boldfaced result with the superscript † denotes that, the F_1 score of the discovered *de facto* Breast Cancer specialty is significantly higher than that of mean of 12 core classes (paired t-test with $p < 0.05$).**

Specialty	**Decision Trees**			**Random Forests**			**PCA-KNN**			**SVM**		
	P	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$
Breast Cancer	95.12	57.35	**71.56**†	91.11	60.29	**72.57**†	82.58	80.88	**81.69**†	96.92	92.65	**94.74**†
Mean of 12 Core Classes	66.42	53.21	59.08	71.28	63.18	66.99	75.45	76.21	75.83	90.84	88.93	89.88
Urology	75.00	45.00	56.25	75.00	45.00	56.25	78.26	90.00	83.72	100.00	95.00	97.44
Neurology	65.52	40.43	50.00	64.52	42.55	51.28	72.73	85.11	78.43	80.36	95.74	87.38
Pulmonary Disease	87.50	58.33	70.00	87.50	58.33	70.00	70.37	79.17	74.51	95.65	91.67	93.62
Orthopaedic Surgery	76.92	34.48	47.62	89.43	79.31	84.07	68.57	82.76	75.00	100.00	93.10	96.43
Neonatal-Perinatal Medicine	100.00	14.29	25.00	100.00	82.29	90.28	92.59	89.29	90.91	96.15	89.29	92.59
Gastroenterology	65.38	44.74	53.12	65.38	44.74	53.12	75.00	94.74	83.72	97.44	100.00	98.70
Obstetrics & Gynecology	55.80	70.64	62.35	57.03	66.97	61.60	90.83	90.83	90.83	99.02	92.66	95.73
Neurological Surgery	100.00	35.00	51.85	88.89	40.00	55.17	50.00	10.00	16.67	100.00	20.00	33.33
Ophthalmology	23.12	95.24	37.21	69.36	95.24	80.27	86.11	73.81	79.49	100.00	88.10	93.67
Cardiovascular Disease	64.29	57.14	60.50	64.15	53.97	58.62	76.62	93.65	84.29	96.77	95.24	96.00
Infectious Disease	76.19	61.54	68.09	73.91	65.38	69.39	52.00	50.00	50.98	96.00	92.31	94.12
Dermatology	75.00	31.58	44.44	82.11	68.42	74.64	79.17	50.00	61.29	55.22	97.37	70.48

P: Precision; R: Recall; F_1: F_1 Score

Table 9: **Average accuracy of multi-class classification on the general data set under 5×2 cross-validation (in percent). Users with the *de facto* Obesity specialty discovered by LDA are in one class; users with core NPI taxonomy codes are in 12 distinct core classes. The boldfaced result with the superscript † denotes that, the F_1 score of the discovered *de facto* Obesity specialty is significantly higher than that of mean of 12 core classes (paired t-test with $p < 0.05$).**

Specialty	**Decision Trees**			**Random Forests**			**PCA-KNN**			**SVM**		
	P	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$
Obesity	100.00	40.41	**57.56**†	83.22	56.98	**67.64**†	75.01	82.12	78.40	92.85	94.01	**93.43**†
Mean of 12 Core Classes	63.18	45.62	52.98	75.62	53.51	62.68	77.12	80.36	78.70	90.23	89.19	89.70
Urology	100.00	50.00	66.67	100.00	55.00	70.97	86.36	95.00	90.48	100.00	95.00	97.44
Neurology	85.71	38.30	52.94	70.71	57.45	63.39	63.49	85.11	72.73	82.14	97.87	89.32
Pulmonary Disease	100.00	45.83	62.86	100.00	45.83	62.86	71.43	83.33	76.92	100.00	87.50	93.33
Orthopaedic Surgery	100.00	3.45	6.67	82.15	37.93	51.90	62.16	79.31	69.70	96.43	93.10	94.74
Neonatal-Perinatal Medicine	100.00	39.29	56.41	100.00	39.29	56.41	89.29	89.29	89.29	100.00	89.29	94.34
Gastroenterology	82.35	36.84	50.91	82.35	36.84	50.91	76.09	92.11	83.33	94.87	97.37	96.10
Obstetrics & Gynecology	30.59	99.08	46.75	38.12	93.58	54.17	94.50	94.50	94.50	100.00	94.50	97.17
Neurological Surgery	100.00	40.00	57.14	100.00	50.00	66.67	33.33	5.00	8.70	100.00	40.00	57.14
Ophthalmology	100.00	4.76	9.09	100.00	4.76	9.09	88.24	71.43	78.95	86.96	95.24	90.91
Cardiovascular Disease	76.92	63.49	69.57	76.92	63.49	69.57	75.64	93.65	83.69	95.31	96.83	96.06
Infectious Disease	78.95	57.69	66.67	78.95	57.69	66.67	66.67	53.85	59.57	88.89	92.31	90.57
Dermatology	100.00	2.63	5.13	87.21	39.47	54.34	66.67	52.63	58.82	67.27	97.37	79.57

P: Precision; R: Recall; F_1: F_1 Score

probabilities to work with topic models that characterize the common activities of employees in certain positions in the hospital. The evaluation of our work can be seen as merging ideas from EBAS and RTAM to explore when a *de facto* diagnosis specialty can be described with a classifier.

An advantage of our work comparing with the other recent work on inappropriate EHR access detection [21, 22, 23] is that our work outputs *de facto* diagnosis specialty information even for those that lack codes from the HPTCS. It has been known that the *de facto* diagnosis specialty information is useful in convincing patients into trusting a provider for using their EHRs [7, 30, 19].

7. CONCLUSIONS

Medical specialties are important but may be lacking or inaccurate in part because there is no official code to express them. We first proposed a novel and challenging *de facto* diagnosis specialty discovery problem under a general discovery–evaluation framework. Under this framework, we then employed a semi-supervised learning model on a fine-grained data set and an unsupervised learning model on a larger general data set for discovery; we further employed four supervised learning models for evaluation. Finally, we experimented on one year of diagnosis histories from a major medical center. The semi-supervised learning model discovered a *de facto* diagnosis specialty for Breast Cancer on the fine-grained data set; the unsupervised learning model confirmed this discovery and suggested a new *de facto* diagnosis specialty for Obesity on the larger general data set. The evaluation results reinforced that these two specialties can be recognized accurately by classifiers in comparison with 12 common diagnosis specialties defined by the Health Care Provider Taxonomy Code Set.

Acknowledgment This work was supported in part by NSF CNS 09-64392 (EBAM), HHS 90TR0003-01 (SHARPS), NSF CNS 13-30491 (THaW), NSF IIS-1418504, NSF CNS-0964063, NSF CCF-0424422, and NIH R01LM010207. The views expressed are those of the authors only.

8. REFERENCES

[1] National provider identifier. http://nppes.cms.hhs.gov/NPPES/Welcome.do.

[2] The npi taxonomy codes for psychology: Apa practice organization offers guidance, advocates for change. http://www.apapracticecentral.org/reimbursement/npi/select-code.aspx.

[3] C. Benesch, D. Witter, A. Wilder, P. Duncan, G. Samsa, and D. Matchar. Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*, 49, 1997.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] Y. Bonomo. Addiction medicine: a new medical specialty in a new age of medicine. *Internal Medicine Journal*, 40(8):543–544, 2010.

[6] K. Caine and R. Hanania. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20(1):7–15, 2013.

[7] K. Caine and W. M. Tierney. Point and counterpoint: Patient control of access to data in their electronic health records. *Journal of general internal medicine*, 30(1):38–41, 2015.

[8] C. Campos-Castillo and D. L. Anthony. The double-edged sword of electronic health records: implications for patient disclosure. *Journal of the American Medical Informatics Association*, 22(e1):e130–e140, 2015.

[9] P. Cunningham and S. J. Delany. k-nearest neighbour classifiers. *Multi Classifier System*, pages 1–17, 2007.

[10] D. E. Detmer, B. S. Munger, and C. U. Lehmann. Clinical informatics board certification: history, current status, and predicted impact on the clinical informatics workforce. *Applied Clinical Informatics*, 1(1):11, 2010.

[11] A. Elixhauser and E. McCarthy. *Clinical classifications for health policy research, version 2: hospital inpatient statistics.* Number 96. US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, 1996.

[12] D. Fabbri and K. LeFevre. Explanation-based auditing. *Proceedings of the VLDB Endowment*, 5(1):1–12, 2011.

[13] D. Fabbri and K. LeFevre. Explaining accesses to electronic medical records using diagnosis information. *Journal of the American Medical Informatics Association*, 20(1):52–60, 2013.

[14] Centers. for Medicare & Medicaid Services. Taxonomy code. http://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/MedicareProviderSupEnroll/Taxonomy.html.

[15] C. A. Gunter, D. Liebovitz, and B. Malin. Experience-based access management: A life-cycle framework for identity and access management systems. *IEEE security & privacy*, 9(5):48, 2011.

[16] S. Gupta, C. Hanson, C. Gunter, M. Frank, D. Liebovitz, B. Malin, et al. Modeling and detecting anomalous topic access. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 100–105. IEEE, 2013.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11, 2009.

[18] R. Hogg and E. Tanis. *Probability and Statistical Inference.* Pearson Prentice Hall, 2006.

[19] K.-M. Kuo, C.-C. Ma, J. W. Alexander, et al. How do patients respond to violation of their information privacy? *Health Information Management Journal*, 43(2):23, 2014.

[20] X. Lu. Diagnosis based specialist identification in the hospital. *Thesis, University of Illinois at Urbana-Champaign*, 2014.

[21] P. Martin, A. D. Rubin, and R. Bhatti. Enforcing minimum necessary access in healthcare through integrated audit and access control. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 946. ACM, 2013.

[22] A. K. Menon, X. Jiang, J. Kim, J. Vaidya, and L. Ohno-Machado. Detecting inappropriate access to electronic health records using collaborative filtering. *Machine learning*, 95(1):87–101, 2014.

[23] A. V. Nimkar and S. K. Ghosh. An access control model for cloud-based emr federation. *International Journal of Trust Management in Computing and Communications*, 2(4):330–352, 2014.

[24] U. Premarathne, F. Han, H. Liu, and I. Khalil. Impact of privacy issues on user behavioural acceptance of personalized mhealth services. In *Mobile Health*, pages 1089–1109. Springer, 2015.

[25] T. R. Shulimzon. Interventional pulmonology: a new medical specialty. *The Israel Medical Association journal*, 16(6):379–384, 2014.

[26] N. D. Soulakis, M. B. Carson, Y. J. Lee, D. H. Schneider, C. T. Skeehan, and D. M. Scholtens. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. *Journal of the American Medical Informatics Association*, 22(2):299–311, 2015.

[27] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 2011.

[28] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the International Conference on Extending Database Technology*, pages 565–576. ACM, 2009.

[29] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Trans. Knowl. Discov. Data*, 7(3):11:1–11:23, Sept. 2013.

[30] W. M. Tierney, S. A. Alpert, A. Byrket, K. Caine, J. C. Leventhal, E. M. Meslin, and P. H. Schwartz. Provider responses to patients controlling access to their electronic health records: a prospective cohort study in primary care. *Journal of general internal medicine*, 30(1):31–37, 2015.

[31] A. Zhang, X. Xie, K. C.-C. Chang, C. A. Gunter, J. Han, and X. Wang. Privacy risk in anonymized heterogeneous information networks. In *Proceedings of the International Conference on Extending Database Technologies*, pages 595–606, 2014.

[32] W. Zhang, Y. Chen, C. Gunter, D. Liebovitz, and B. Malin. Evolving role definitions through permission invocation patterns. In *Proceedings of the 18th ACM symposium on Access control models and technologies*, pages 37–48. ACM, 2013.

[33] W. Zhang, C. A. Gunter, D. Liebovitz, J. Tian, and B. Malin. Role prediction using electronic medical record system audits. In *AMIA Annual Symposium Proceedings*, volume 2011, page 858. American Medical Informatics Association, 2011.