Received: 18 December 2017 | Revised: 10 April 2018 | Accepted: 11 April 2018

DOI: 10.1002/lrh2.10057

RESEARCH REPORT

Learning Health Systems

De facto diagnosis specialties: Recognition and discovery

Aston Zhang[1] | Xun Lu[1] | Carl A. Gunter[1] | Shuochao Yao[1] | Fangbo Tao[1] | Rongda Zhu[1] | Huan Gui[1] | Daniel Fabbri[2] | David Liebovitz[3] | Bradley Malin[2]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois

[2] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee

[3] Department of Medicine, University of Chicago, Chicago, Illinois

Correspondence
Aston Zhang, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
Email: lzhang74@illinois.edu

Funding information
National Science Foundation, Grant/Award Numbers: CNS 0964392, CNS 1330491, 0964063 and 1526014; US Department of Health & Human Services, Grant/Award Number: 90TR0003-01

Abstract

A medical specialty indicates the skills needed by health care providers to conduct key procedures or make critical judgments. However, documentation about specialties may be lacking or inaccurately specified in a health care institution. Thus, we propose to leverage diagnosis histories to recognize medical specialties that exist in practice. Such specialties that are highly recognizable through diagnosis histories are de facto diagnosis specialties. We aim to recognize de facto diagnosis specialties that are listed in the Health Care Provider Taxonomy Code Set (HPTCS) and discover those that are unlisted. First, to recognize the former, we use similarity and supervised learning models. Next, to discover de facto diagnosis specialties unlisted in the HPTCS, we introduce a general discovery-evaluation framework. In this framework, we use a semi-supervised learning model and an unsupervised learning model, from which the discovered specialties are subsequently evaluated by the similarity and supervised learning models used in recognition. To illustrate the potential for these approaches, we collect 2 data sets of 1 year of diagnosis histories from a large academic medical center: One is a subset of the other except for additional information useful for network analysis. The results indicate that 12 core de facto diagnosis specialties listed in the HPTCS are highly recognizable. Additionally, the semi-supervised learning model discovers a specialty for breast cancer on the smaller data set based on network analysis, while the unsupervised learning model confirms this discovery and suggests an additional specialty for *Obesity* on the larger data set. The potential correctness of these 2 specialties is reinforced by the evaluation results that they are highly recognizable by similarity and supervised learning models in comparison with 12 core de facto diagnosis specialties listed in the HPTCS.

KEYWORDS

diagnosis specialty, electronic health record, medical informatics, machine learning

1 | INTRODUCTION

Medical specialties provide information about which health care providers (hereinafter referred to as "providers") have the skills needed to conduct 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. However, such an organization has its limitations. For instance, at a large medical center, some specialties may be lacking or inaccurately described (eg, 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 in various settings, such as staffing by providing accurate specialty information about current staff and building patient confidence by ensuring that patients are treated by specialists.[1]

In the United States, providers select from the Health Care Provider Taxonomy Code Set (HPTCS)[2] when they apply for their National Provider Identifiers (NPIs).[3] NPIs are required by the Health Insurance Portability and Accountability Act of 1996 and are used in health care–related transactions. Providers usually choose taxonomy codes according to the certifications they hold. Ideally, this mechanism would identify each provider with the taxonomy codes that most accurately describe their specialties. However, this is not always the case for several reasons.

First, the National Plan & Provider Enumeration System does not verify that the taxonomy code selections made by providers in NPI applications are accurate.[2] 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.[4] Third, not all national certifications are reflected by the taxonomy code list. Since the taxonomy codes do not correspond to certifications within the field, providers may interpret these codes inconsistently.

In view of the limitations of purely relying on the taxonomy codes, we introduce methods to leverage real-world diagnosis histories to infer and recognize actual specialties. We refer to such inferred knowledge as de facto specialties, which we define as medical specialties that exist in practice regardless of the taxonomy codes that are selected from the HPTCS.

Recognizing de facto specialties can be useful. This would enable administrative teams to verify the taxonomy codes of the providers in a health care institution. If certain providers' declared specialties failed to match their activity-based specialties, a possible redesignation of their codes or investigation might be warranted.

Moreover, there is benefit in discovering de facto specialties that are unlisted in the HPTCS. As the medical profession evolves, the HPTCS may not be comprehensive enough.[5-7] Inefficiencies and mismanagement could arise if the specialty codes are not sufficiently expressive to convey providers' specialties. For instance, if there is no official taxonomy code to express certain specialties, since no provider could declare such unlisted de facto specialties, false alarms of suspicious electronic health record (EHR) access detection might be raised. 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."[4]

However, as shown in this paper, not all specialties can be accurately recognized through diagnosis histories. Thus, the focus of this study is on "de facto diagnosis specialties" of providers that exist in practice and are highly recognizable by the diagnoses documented in the EHRs of the patients. Our goal is to recognize de facto diagnosis specialties and discover those that do not have official taxonomy codes in the HPTCS.

To demonstrate the feasibility of our methods, we study 1 year of diagnosis histories from Northwestern Memorial Hospital with 2 data sets. One is attributable and the other is full. The attributable data set is a subset of the full data set except for additional information useful for network analysis. We make the following major contributions.

- We introduce methods to leverage real-world EHRs of patients to recognize de facto diagnosis specialties. We use similarity and supervised learning models for recognition.
- We show that 12 core de facto diagnosis specialties listed in the HPTCS are highly recognizable. For instance, multilayer perceptrons achieve an F_1 score of 90.90% for the mean of these 12 specialties on the full data set.
- We propose a novel de facto diagnosis specialty discovery problem. To solve it, we introduce a general discovery-evaluation framework. Specifically, the framework begins by using a semi-supervised learning model based on heterogeneous information network analysis or an unsupervised learning model based on topic modeling for discovery. The discovered results are then evaluated by similarity and supervised learning models used in recognition. As a result, the discovery problem enriches the applications of the recognition problem by resorting to recognition models for evaluating the discovery results.
- We show that the semi-supervised model discovers a de facto diagnosis specialty for breast cancer on the attributable data set. The unsupervised learning method confirms this discovery and suggests a new de facto diagnosis specialty for *Obesity* on the larger full data set. The potential correctness of these 2 specialties is reinforced by the evaluation results that they are highly recognizable by similarity and supervised learning models in comparison with 12 core de facto diagnosis specialties listed in the HPTCS.

A preliminary version of the de facto diagnosis specialty discovery portion of this work was reported at the 6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.[8] The research reported in the current paper extends the prior work and includes comprehensive studies of both recognizing and discovering de facto diagnosis specialties. Specifically, the empirical findings of recognizing the 12 core de facto diagnosis specialties on the data set that excludes discovered specialties are new. The current paper also expands the scope of recognition and evaluation of the results with a new similarity model and a multilayer perceptron model. The results of their recognition and evaluation of discovery are only reported in

the current paper. In addition, we describe and discuss results of using procedure codes for recognition exclusively in the current paper.

2 | BACKGROUND

This section describes related work and introduces the data sets and performance measures that are used in this study.

2.1 | Related work

A driver behind inferring medical specialties is the analysis of audit logs for security and privacy purposes.[9-12] This is feasible because patient records and audit logs encode valuable interactions between users and patients.[13] 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 investigated the idea of examining accesses to patient records to determine the position of an employee.[14] This work used a Näive 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.[16]

The problem of determining which departments are responsible for treating a given diagnosis was addressed by studies on Explanation-Based Auditing System (EBAS).[17,18] They are similar to our problem of identifying a user's specialty. In these studies, the auditing system uses the access patterns of departments to determine diagnosis responsibility information in 2 ways: by analyzing (1) how frequently a department accesses patients with the diagnosis and (2) 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[1] went beyond approaches based on conditional 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 random topic access model 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 patient record access detection[19-21] is that our work outputs de facto diagnosis specialty information even for those that lack codes in the HPTCS. It has been shown that the de facto diagnosis specialty information is useful in convincing patients into trusting a provider for using their EHRs.[22,23]

2.2 | Data sets

We collect access log data from a hospital and combine it with the diagnosis lists in patient discharge records. For each encounter (visit to the hospital by a patient), there is a set of diagnoses, and for each provider, there is a record of whether the provider accesses the EHR of that patient during that encounter. We use the term users (as in EHR users) rather than providers.

TABLE 1 Attributes of the attributable and full data sets

	Attributable	Full
Accesses	35,869	4 829 376
Patients	41,603	291 562
Users (providers)	2,504	3269
Patient encounters	62,390	890 812
Taxonomy codes	161	165
Diagnoses	4,172	13 566
Procedures	740	2165

The data for this study come from the Cerner Powerchart EHR system in use at Northwestern Memorial Hospital. The data contain all user accesses (in the form of audit logs) made over a 1-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 are de-identified for this study in accordance with the Health Insurance Portability and Accountability Act Privacy Rule and conducted 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.

A small portion of the collected data has an explicit mapping between users and the diagnoses documented in the EHRs they access. In other words, such diagnoses can be attributed to the users who access them. We refer to this portion of the data as the *attributable data set*. However, the majority of the data lacks such an explicit relationship. In fact, patients may have multiple diagnoses and their EHRs may be accessed by different users without documentations about which specific diagnoses are associated with the actions of which user. Although an attributable data set is more desirable with attributable access information, it may not always be available in practice. To this end, we also expand to a more general data set that is more representative of the challenging scenarios encountered in practice. Hence, we also use all of the data after removing attributable information, which we refer to as the *full data set*. The attributes of the data sets used in this study are summarized in Table 1.

We use the Clinical Classifications Software to cluster diagnosis and procedure codes into a manageable number of clinically meaningful categories.[24] This is because ICD-9 codes are not completely indicative of patients' clinical phenotypes[25] and the sheer number of codes makes it too challenging to characterize patterns of diagnoses or procedures. The ICD-9 codes for diagnoses and the ICD-9-CM codes for procedures are aggregated into 603 and 346 Clinical Classifications Software codes, respectively. In practice, due to a lack of documentation or multiple specialties, a user can have zero or many taxonomy codes. In this study, only the primary taxonomy code, if any, is considered hereinafter. For either the attributable or the full data set, the majority of users are known to have accurate taxonomy codes as their labels.* The taxonomy codes for the remainder of the users are either inaccurate or missing and are labeled as NA. Newly discovered de

*The percentage is not reported due to the proprietary nature of the data.

facto diagnosis specialties will be assigned to NA-labeled users. In different sets of experiments, recognition models are trained on access logs whose users have accurate taxonomy codes, or together with access logs whose users are assigned with newly discovered de facto diagnosis specialties. To ensure there is a sufficient amount of data to train models, taxonomy codes with fewer than 20 associated users in either data set are filtered out.[26]

2.3 | Performance measures

We use precision, recall, and the F_1 score as the performance measures. The precision P for a specialty s is the faction of correctly classified users among those who are classified as s. The recall R for a specialty s is the fraction of users with specialty s who have been recognized over all available users with s. The precision of a recognition model 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 recognition model is defined similarly. The F_1 score is the harmonic mean of the precision (P) and recall (R): $F_1 = 2 \cdot P \cdot R/(P + R)$. In general, a higher F_1 score indicates a better performance.

3 | RECOGNIZING DE FACTO DIAGNOSIS SPECIALTIES

Here, we illustrate the concept and recognition models in greater detail.

3.1 | De facto diagnosis specialty

Intuitively, it should be easier to characterize a urologist with medical diagnoses for conditions of the kidney, ureter, and bladder, as opposed to an anesthesiologist, whose duties are more crosscutting 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 simple similarity recognition model based on diagnosis codes. To gain intuition into the general idea, let us delay the technical discussions of the recognition model in Section 3.2 and 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 of each diagnosis is accessed by each user, with some adjustment for its frequency. This technique is typified by term frequency-inverse document frequency (TF-IDF).[27] Then based on TF-IDF, we represent each diagnosis specialty by its most relevant diagnoses and represent each user by the diagnoses in the most frequently accessed EHRs. Finally, the similarity model can classify users according to the specialties with which they share the diagnoses in the EHRs that are frequently accessed. Using the full data set as described in Section 2.2, we observe that urologists tend to access EHRs with diagnoses such as "retention of urine" and "urinary tract infection," whereas anesthesiologists tend to access EHRs with diagnoses such as "hemorrhage of rectum and anus" and "nausea with vomiting." When using the diagnoses in frequently accessed EHRs by either of the 2 specialists as the features for the similarity model, the results are decent for recognizing the urologists, yielding an F_1 score of 70.35%. However, the results for recognizing anesthesiologists are much poorer, yielding an F_1 score of 11.30%. If we use a supervised learning model, such as support vector machines, we can achieve substantially better results: recognizing anesthesiologists with an F_1 score of 48.98%. However, this performance is still weaker than that of recognizing urologists, which achieves an F_1 score of 97.44%. The experimental results show that urology is more recognizable than anesthesiology by the diagnoses inherent in EHRs. Thus, urology is more likely a diagnosis specialty than anesthesiology.

Based on the guidance of several clinicians and hospital administrators, we identify 12 taxonomy codes from the HPTCS as diagnosis specialties: cardiovascular disease, dermatology, gastroenterology, infectious disease, neonatal-perinatal medicine, neurological surgery, neurology, obstetrics and gynecology, ophthalmology, orthopaedic surgery, pulmonary disease, and urology. We refer to this group as the *12 core diagnosis specialties* or simply *12 core classes*. Recall that de facto diagnosis specialties are specialties that exist and are highly recognizable through diagnosis histories. In our experiments, we aim to recognize these specialties from the data sets as described in Section 2.2. If a de facto diagnosis specialty that is unlisted in the HPTCS is discovered, then its recognition performance will be compared with those of such 12 core de facto diagnosis specialties listed in the HPTCS.

3.2 | Recognition models

Ideally, a de facto diagnosis specialty can be recognized accurately through diagnosis histories. 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 use. The users can be likened to readers of an archive of documents. The words in each document correspond to diagnoses. Users 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 recognition problem, we aim to develop a classifier that characterizes this common interest with respect to the documents that they have read, or the EHRs that they have accessed. For instance, if there are a group of readers who 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.

In the rest of this section, we describe the similarity and supervised learning models for recognition. Essentially, all of these models are classifiers. They require feature vectors as the inputs of classification, which are described as follows.

3.2.1 | Feature vector

Suppose that u is a user in the set of users U whose cardinality is $|U|$ and d is a diagnosis in the set of diagnoses D whose cardinality is $|D|$. Let $n_{u,d}$ be the number of times that the user u accesses EHRs with the documented diagnosis d. We use m_d to represent the number of users who access EHRs with the documented diagnosis d. To apply

recognition models to the data sets, each user $u \in U$ is mapped to the following TF-IDF weighted diagnosis vector

$$\mathbf{v}_u = \left[v_{u,d_1} \ldots v_{u,d_{|D|}}\right]^\top \in \mathbb{R}^{|D|}, \quad (1)$$

where each element of the vector is the relevance score of its corresponding diagnosis $d \in D$ to u and is computed according to TF-IDF,

$$v_{u,d} = \log\left(n_{u,d} + 1\right) \log\left(\frac{|U|}{m_d}\right). \quad (2)$$

Note that (2) applies logarithms to penalize higher frequencies with a pseudo count of one to mitigate bias from lower frequencies, especially zero frequencies. The feature vector $\mathbf{v}_u$ in (1) serves as the input to the recognition models, which we now describe.

3.2.2 | Similarity model

The similarity model proceeds by finding the most relevant diagnoses of each diagnosis specialty and the diagnoses in the most frequently accessed EHRs by each user. Users are classified according to the specialties with which they share the most common diagnoses.

Let s be a diagnosis specialty that takes the form of a taxonomy code. The set of diagnosis specialties is denoted by S. For the user u, we define $V_u^{(l)}$ as the set of diagnoses whose corresponding relevance scores in (2) are the largest l elements in (1). Thus, $V_u^{(l)}$ can be considered as the l most representative diagnoses for the user u that are frequently accessed by u and have high distinctions as reflected by the idea of TF-IDF.

We go on to find the l most representative diagnoses for each diagnosis specialty in a similar way. We use $U^{(s)}$ to represent the set of users whose taxonomy code is s. Then, the TF-IDF weighted diagnosis vector of a diagnosis specialty s is

$$\mathbf{v}_s = \frac{\sum_{u \in U_s} \mathbf{v}_u}{|U_s|}, \quad (3)$$

where $\mathbf{v}_u$ is the TF-IDF weighted diagnosis vector of u as defined in (1). Similarly, for the specialty s, we define $V_s^{(l)}$ as the set of diagnoses whose corresponding relevance scores are the largest l elements in $\mathbf{v}_s$ as given by (3). Likewise, $V_s^{(l)}$ can be thought of as the most representative diagnoses for the diagnosis specialty s.

The similarity between a user u and a diagnosis specialty s is characterized by the Jaccard similarity coefficient of $V_u^{(l)}$ and $V_s^{(l)}$. In this way, the similarity model will recognize a user u's diagnosis specialty s as

$$\text{argmax}_{s \in S} \frac{|V_u^{(l)} \cap V_s^{(l)}|}{|V_u^{(l)} \cup V_s^{(l)}|}, \quad (4)$$

where l is a parameter that will be tuned experimentally.

3.2.3 | Supervised learning models

We further use 5 recognition models based on supervised learning: decision trees (J48), random forests, k nearest neighbors with principal component analysis (PCA-KNN), support vector machines, and multilayer perceptrons. Specifically, to mitigate the curse of dimensionality problem by k nearest neighbors, we use principal component analysis by selecting a small number of the principal components to perform dimension reduction. For support vector machines, a Gaussian kernel is used.

4 | RECOGNITION EXPERIMENTS

This section describes the experimental setting for recognizing de facto diagnosis specialties and the results of the experiments.

4.1 | Experimental setting

To recognize the 12 core classes as listed in Section 3.1, users whose taxonomy codes are not in the 12 core classes are excluded from the full data set. For the decision trees, random forests, support vector machines, and multilayer perceptrons, we use the default parameter values in Weka.[28] We split the data instances in the full data set so that 20% is used for parameter tuning and the remaining 80% is for performance evaluation. Within the 20% portion of the data set, we use one half for model training and the other half for validation to select the parameter values that achieve the highest F_1 scores for the mean of the 12 core classes. The number of nearest neighbors K and the number of principal components are set to 9 and 50, respectively, for KNN-PCA. The parameter l of the similarity model in (4) is set to 100. We use 10 × 2 cross-validation for recognition with classifiers on the 80% portion of the data set. In each of the 10 rounds, the data instances are randomly split into 2 equal-sized sets. Then a model is trained on one set and tested on the other set and vice versa. After these 10 rounds, the average of the 20 testing results is reported.

4.2 | Feature study

Since both diagnosis and procedure codes are used separately in EHRs, we devise 2 preliminary experiments using features from *diagnoses* only and *procedures* only. As discussed in Section 3.2.1, features from *diagnoses* means that TF-IDF vectors (1) of diagnosis codes are used as input features for recognition models. The experiments for *procedures* take input features from the procedure codes in a similar manner.

We evaluate the F_1 scores of the mean of 12 core classes with multilayer perceptrons under the 10 × 2 cross-validation using a paired t test with $P < 0.05$. The results indicate that *diagnoses* (F_1 score = 90.90%) yield statistically significantly better results than *procedures* (F_1 score = 85.48%). Similar results are obtained for the other recognition models. Such findings indicate that it is easier to characterize and classify users with respect to medical diagnoses than procedures. Hence, we use the diagnosis features in the remainder of our experiments.

4.3 | Recognition results

The results of the experiments for recognizing 12 core de facto diagnosis specialties are presented in Table 2. It is important to highlight that multi-class classification (12 classes in our case) is generally more challenging than binary classification. Table 2 shows that, in general, the 12 core de facto diagnosis specialties listed in the HPTCS are highly recognizable. Despite its simplicity, the similarity model is

TABLE 2 De facto diagnosis specialty recognition performance on the full data set (in percent)

Specialty	Similarity			Decision Trees			Random Forests		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Mean of 12 Core classes	**67.63**	**66.53**	**67.07**	**71.47**	**67.98**	**69.68**	**73.54**	**73.55**	**73.55**
Cardiovascular disease	80.00	69.84	74.58	72.31	74.60	73.44	66.67	53.97	59.65
Dermatology	53.66	57.89	55.70	72.41	55.26	62.69	76.19	84.21	80.00
Gastroenterology	75.00	71.05	72.97	62.50	78.95	69.77	55.88	50.00	52.78
Infectious Disease	50.00	57.69	53.57	65.71	88.46	75.41	77.42	92.31	84.21
Neonatal-Perinatal Medicine	50.00	53.57	51.72	90.91	35.71	51.28	95.83	82.14	88.46
Neurological Surgery	48.00	60.00	53.33	100.00	45.00	62.07	81.82	45.00	58.06
Neurology	66.67	55.32	60.47	60.53	48.94	54.12	67.44	61.70	64.44
Obstetrics & Gynecology	75.24	72.48	73.83	75.22	77.98	76.58	69.05	79.82	74.04
Ophthalmology	73.81	73.81	73.81	44.74	80.95	57.63	78.72	88.10	83.15
Orthopaedic Surgery	51.35	65.52	57.58	72.73	55.17	62.75	86.21	86.21	86.21
Pulmonary Disease	70.37	79.17	74.51	89.47	70.83	79.07	80.00	83.33	81.63
Urology	76.47	65.00	70.27	73.68	70.00	71.79	80.95	85.00	82.93
Specialty	PCA-KNN			Support Vector Machines			Multilayer Perceptrons		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Mean of 12 core classes	**79.29**	**78.10**	**78.69**	**91.27**	**90.29**	**90.78**	**91.31**	**90.50**	**90.90**
Cardiovascular disease	78.87	88.89	83.58	93.44	90.48	91.94	95.31	96.83	96.06
Dermatology	64.86	63.16	64.00	72.34	89.47	80.00	77.08	97.37	86.05
Gastroenterology	86.49	84.21	85.33	85.00	89.47	87.18	90.48	100.00	95.00
Infectious Disease	57.14	61.54	59.26	92.31	92.31	92.31	83.33	96.15	89.29
Neonatal-Perinatal Medicine	100.00	67.86	80.85	96.43	96.43	96.43	92.86	92.86	92.86
Neurological Surgery	56.25	45.00	50.00	100.00	50.00	66.67	100.00	50.00	66.67
Neurology	74.00	78.72	76.29	79.66	100.00	88.68	87.76	91.49	89.58
Obstetrics & Gynecology	90.11	75.23	82.00	96.00	88.07	91.87	100.00	88.99	94.17
Ophthalmology	88.10	88.10	88.10	100.00	97.62	98.80	88.64	92.86	90.70
Orthopaedic Surgery	61.90	89.66	73.24	89.66	89.66	89.66	83.87	89.66	86.67
Pulmonary Disease	78.57	91.67	84.62	91.67	91.67	91.67	78.26	75.00	76.60
Urology	78.26	90.00	83.72	100.00	95.00	97.44	100.00	90.00	94.74

effective. It attains an F_1 score of 67.07% for the mean of the 12 core classes. Under the 10 × 2 cross-validation, according to a paired t test with $p < 0.05$ for the F_1 score of the mean of the 12 core classes, all of the supervised learning models perform statistically significantly better than the similarity model. For instance, multilayer perceptrons achieve an F_1 score of 90.90% for the mean of these 12 specialties.

5 | DISCOVERING DE FACTO DIAGNOSIS SPECIALTIES

Next, we aim to discover de facto diagnosis specialties that lack official taxonomy codes in the HPTCS. The recognition results of the de facto diagnosis specialties in Section 4.3 suggest that an unlisted de facto diagnosis specialty, if discovered, may be evaluated by those recognition models.

5.1 | Discovery-evaluation

It is important to emphasize that there is no ground truth for the de facto diagnosis specialty discovery problem. Hence, we solve it under a general discovery-evaluation framework.

5.1.1 | Discovery

We first use a semi-supervised learning model to leverage the mapping between users and their specifically accessed diagnoses of EHRs in the attributable data set. Next, we consider a more challenging scenario where such attributable access information is unavailable. In this case, we use an unsupervised learning model for discovery in the larger full data set. Since the attributable data set is a subset of the full data set except for the attributable access information, the discovery results can be reinforced if they exhibit common findings on both data sets.

5.1.2 | 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 use similarity and supervised learning models to evaluate the recognition performance of the discovered de facto diagnosis specialties by comparing them with the recognition performance of the listed de facto diagnosis specialties, such as the 12 core classes as described in Section 3.1. Ideally, their recognition performance should be similar. We evaluate such recognition performance using the same recognition models as described in Section 3.2.

5.2 | PathSelClus for discovery

Intuitively, discovering de facto diagnosis specialties through diagnosis histories may rely upon effective clustering techniques. Such techniques may divide a pool of users into groups that have high intragroup similarities but low intergroup similarities. We anticipate that new de facto diagnosis specialties may emerge from these clusters. As noted in Section 2.2, the attributable data set has an explicit mapping between users and the diagnoses documented in the EHRs they access. In fact, the structure of the attributable data set can be represented as heterogeneous information networks.[29-31] Given the heterogeneous information network setting and partially labeled ground truth, we use PathSelClus, a semi-supervised learning model based on such a network setting.[30] For context, we briefly introduce heterogeneous information networks.

5.2.1 | Heterogeneous information networks

An information network consists of objects and links. There are multiple types of objects or links in a heterogeneous information network. This type of network explicitly distinguishes between object and link types. For instance, there exist 3 types of objects in the attributable data set: users, patients, and diagnoses. Links exist between users and patients through the relations of "access EHRs of" and "whose EHRs are accessed by"; links exist between users and diagnoses through the relations of "access" and "accessed by." Note that such links between users and diagnoses are only available in the attributable data set where there exists the attributable access information on users and their accessed diagnoses in the EHRs.

In heterogeneous information networks, link-based clustering groups objects based on their relations to other objects in the networks. The relations derived from a heterogeneous information network between 2 objects are called meta-paths.[32] In our case, the target object type for clustering is the user object. To cluster users, there are 2 meta-paths that capture relations between users: User (access EHRs of) Patient (whose EHRs are accessed by) User and User (access) Diagnosis (accessed by) User.

5.2.2 | Semi-supervised learning

During clustering, a decision has to be made about the weighted combination of different meta-paths to use. Such a decision can be guided by the seeded target objects in different clusters. For guided clustering on heterogeneous information networks, we use the semi-supervised learning model PathSelClus.

In PathSelClus, the guidance of clustering takes the form of object seeds in each cluster. For example, to cluster users based on the pattern of their accessed diagnosis histories, one can provide representative users who have similar access patterns as seeds in each individual cluster. These seeds provide guidance for clustering similar target objects in the heterogeneous information networks and help adjust combination weights of meta-paths during the clustering process. It is important to note that PathSelClus can handle input clusters that are unseeded. This is the exact feature that makes it possible to use PathSelClus to discover new de facto diagnosis specialties. The discovering process is illustrated as follows.

5.2.3 | Clustering users

Recall Section 2.2 that the majority of users have accurate taxonomy codes while the rest are labeled as NA. For the majority of users who have accurate taxonomy codes, we create a cluster for each specialty. Each cluster is initialized by being assigned users of the same specialty as the seeds of the cluster. We also create additional empty clusters. Each empty cluster is expected to be populated with NA-labeled users who have similar access patterns as guided by the other seeded clusters. As an output, each NA-labeled user is assigned to the cluster with the highest assignment likelihood. The clusters that are assigned to NA-labeled users can be either seeded or unseeded ones.

We can analyze the semantics of the unseeded clusters via their assigned users. We treat a cluster as a taxonomy code and find the most relevant diagnoses for each cluster. Specifically, the semantics of an unseeded cluster may be exhibited via a list of the most frequently accessed diagnoses by the users in the cluster. Note that information on user accessed diagnoses is only available in the attributable data set. Based on the semantics, the medical expert can label each unseeded cluster, which we use to interpret the discovery results.

5.3 | Latent Dirichlet allocation for discovery

In practice, attributable data sets are not always available for using PathSelClus. Hence, we also consider an unsupervised learning method based on topic modeling.

5.3.1 | Unsupervised learning based on topic modeling

Latent Dirichlet allocation (LDA) is an unsupervised learning method based on topic modeling.[33] In the language of text analysis, a corpus is a collection of documents, where each document is composed of words. With the output of LDA, on the corpus level a topic can be represented by a ranked list of words ordered by their generative likelihoods given the topic. Here, topics can be thought of as summaries of the different themes pervasive in the corpus. A topic may be interpreted from the semantics exhibited in the words most likely to be generated by the topic. Meanwhile, with the output each document can be characterized with respect to these topics in the form of a distribution over the topics, which is also known as topic allocations.

5.3.2 | Clustering users

The intuition behind our employment of LDA is from the possible existence of diagnosis topics with coherent themes in a hospital. In other words, if diagnoses documented in EHRs are considered as words, the de facto diagnosis specialties may correspond to topics. As in text analysis, if a new de facto diagnosis specialty is discovered, it may be interpreted via the semantics of the diagnoses most likely to be generated by the specialty.

To represent a document of the corpus in the hospital setting, we extract diagnoses of all the EHRs of a patient to form the words in a document. In this way, each document may be indexed by a patient $p \in P$ where P is the set of the patients in the data set. With LDA, each patient p can be characterized by a topic distribution $\boldsymbol{\pi}_p$. As an output of LDA, each diagnosis topic can be represented by a ranked list of diagnoses ordered by their generative likelihoods given the topic.

Recall that in the de facto diagnosis specialty recognition problem, models recognize de facto diagnosis specialties of users based on the diagnoses documented in their accessed EHRs. Similarly, the output diagnosis topics by LDA correspond to de facto diagnosis specialties of users. Now, we can characterize users with respect to de facto diagnosis specialties from their accessed EHRs of patients. Denote by P_u the set of patients whose EHRs are accessed by the user u. Let $|P_u|$ be the cardinality of the set P_u. Given the topic distribution $\boldsymbol{\pi}_p$ of each patient p, the topic distribution of the user u can be computed as

$$\boldsymbol{\pi}_u = \frac{\sum_{p \in P_u} \boldsymbol{\pi}_p}{|P_u|}. \quad (5)$$

With the topic distribution in (5), each user is now characterized with respect to de facto diagnosis specialties. Let $s \in S'$ be a de facto diagnosis specialty whose topic distribution value for a user u is indexed by $\boldsymbol{\pi}_u[s]$. Here, S' is the set of diagnosis specialties whose cardinality $|S'|$ is equal to the predefined number of topics. Note that $|S'|$ is also equal to the dimension of any topic distribution vector, such as $\boldsymbol{\pi}_u$. To cluster NA-labeled users based on the same de facto diagnosis specialty, a user u is assigned with a de facto diagnosis specialty

$$s = \text{argmax}_{s \in S'} \boldsymbol{\pi}_u[s],$$

where the assigned specialty s indexes the largest element value in the vector $\boldsymbol{\pi}_u$.

5.4 | Evaluation

As discussed in Sections 5.2 and 5.3, we can manually interpret the de facto diagnosis specialties via their representative diagnoses. In PathSelClus, a de facto diagnosis specialty is represented by the diagnoses that are most frequently accessed by all the users in the same cluster. Such user accessed diagnosis information is only available in the attributable data set. For LDA on the full data set, a de facto diagnosis specialty is represented by the diagnoses most likely to be generated by the specialty as a diagnosis topic. 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 unlisted specialties exist, they are considered to be potential newly discovered de facto diagnosis specialties.

It is important to highlight that there is no ground truth for the discovery results and such expert opinions are not always available in practice. We use recognition models in Section 3.2 to evaluate the recognition performance of the discovered de facto diagnosis specialties. Ideally, their recognition performance should be similar to that of the 12 core diagnosis specialties listed in the HPTCS.

6 | DISCOVERY EXPERIMENTS

This section reports on the de facto diagnosis specialties discovered by PathSelClus and LDA. When evaluating recognition performance for the discovered specialties, we use the same experimental setting as in Section 4.1. The recognition models used for evaluation are described in Section 3.2.

6.1 | Discovery results for PathSelClus

To illustrate that the user accessed diagnosis information in the attributable data set is useful for discovery with PathSelClus, we start by using PathSelClus on the full data set where such attributable access information is unavailable. The meta-paths remain the same except that in the *User (access) Diagnosis (accessed by) User* meta-path, 2 different users are related if the same diagnosis is documented in both of their accessed EHRs. The semantics of an unseeded cluster is given by a list of diagnoses from the most frequently accessed EHRs by the assigned users to the cluster.

We observe that a patient can have multiple diagnoses related to different specialties in the same EHR, such as "retention of urine" and "benign neoplasm of skin of upper limb, including shoulder". Suppose that a urologist accesses the diagnosis "retention of urine" and a dermatologist accesses the diagnosis "benign neoplasm of skin of upper limb, including shoulder" in the EHR. Such attributable access information is available in the attributable data set. However, PathSelClus considers that both specialists access both diagnoses in the same EHR. The inaccurate access mapping makes it difficult for PathSelClus to discover de facto diagnosis specialties on the full

TABLE 3 Top diagnoses of 3 specialties as discovered by PathSelClus from the full data set[a]

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

[a]None of the clusters shows a consistent theme with respect to a de facto diagnosis specialty. PathSelClus fails to discover de facto diagnosis specialties without attributable access information.

TABLE 4 PathSelClus discovers a de facto diagnosis specialty for breast cancer on the attributable data set[a]

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

[a]This specialty is represented by 10 most frequently accessed diagnoses by the users who are assigned with the breast cancer specialty.

data set. For example, Table 3 shows the top diagnoses of 3 specialties as discovered by PathSelClus (the unseeded cluster count is 3). None of the clusters exhibits a consistent theme with respect to a specialty, even when the unseeded cluster count is set to other values.

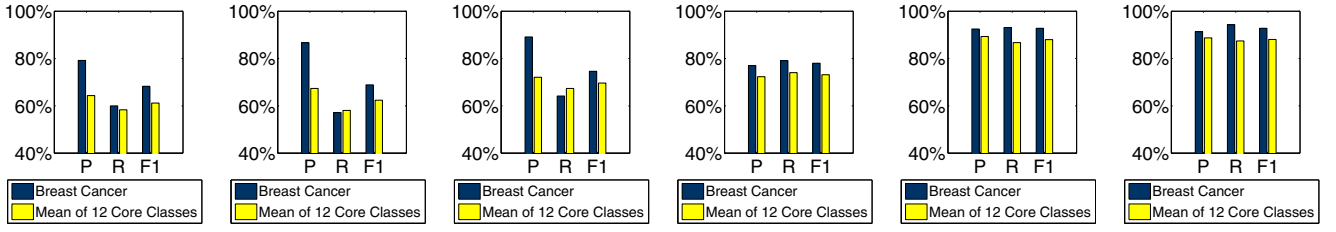
With respect to the attributable data set, PathSelClus discovers a de facto diagnosis specialty for breast cancer that does not have an official taxonomy code in the HPTCS. Table 4 lists the most frequently accessed diagnoses by the 35 users who are assigned with the breast cancer specialty when the unseeded cluster count is set to 3.

Figure 1 shows the average of the performance measures of multiclass classification on the attributable data set under 10 × 2 cross-validation. Users with the de facto breast cancer specialty discovered by PathSelClus are in one class, while users whose taxonomy codes belong to the core de facto diagnosis specialties as listed in Section 3.1 are in 12 distinct core classes. According to a paired t test with $P < 0.05$, the F_1 score of the breast cancer specialty discovered by PathSelClus is statistically significantly higher than that of the mean of 12 core classes under the 6 recognition models.

6.2 | Discovery results for LDA

We set the number of topics for LDA to 30 by minimizing the perplexity measure.[33] In the larger full data set, LDA confirms the discovery of breast cancer by PathSelClus and suggests another de facto diagnosis specialty for obesity as shown in Table 5. These specialties are represented by 10 diagnoses most likely to be generated by these 2 diagnosis topics. The breast cancer and obesity specialties are assigned to 68 and 20 users, respectively.

Figures 2 and 3 summarize the average of the performance measures of multiclass classification on the full data set under 10 × 2 cross-validation for the 2 discovered de facto diagnosis specialties. According to a paired t test with $P < 0.05$, the F_1 score of the discovered de facto breast cancer specialty by LDA is also statistically significantly higher than that of the mean of 12 core classes under all the recognition models. It reaffirms the finding by PathSelClus. The result for obesity is similar, except for PCA-KNN. Overall, both the breast cancer and obesity specialties discovered by LDA are highly recognizable on the full data set.



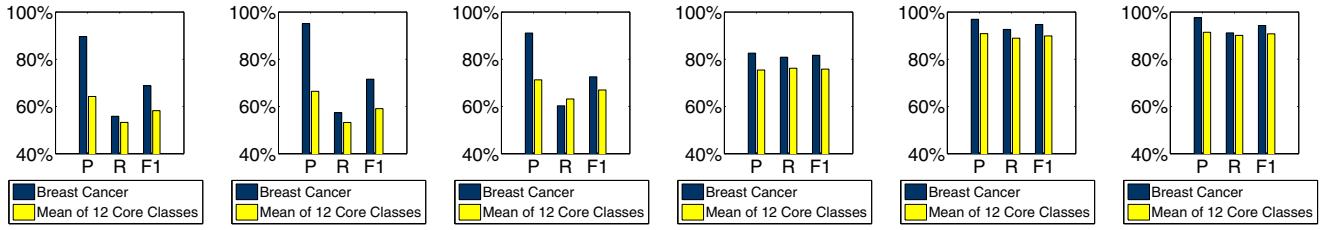
(A) Similarity (B) Decision Trees (C) Random Forests (D) PCA-KNN (E) Support Vector Mach. (F) Multilayer Perceptrons

FIGURE 1 De facto diagnosis specialty recognition performance for multiclass classification on the attributable data set. Users with the de facto breast cancer specialty discovered by PathSelClus are in one class; users whose taxonomy codes belong to the core de facto diagnosis specialties are in 12 distinct classes

TABLE 5 Latent Dirichlet allocation discovers de facto diagnosis specialties for breast cancer (left) and obesity (right) on the full data set[a]

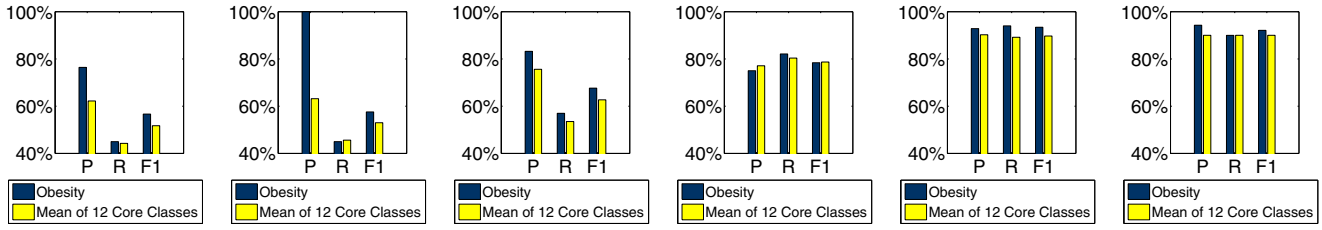
Personal history of malignant neoplasm of breast	Obesity, unspecified
Lump or mass in breast	Morbid obesity
Abnormal mammogram, unspecified	Obstructive sleep apnea
Other specified aftercare following surgery	Unspecified sleep apnea
Other sign and symptom in breast	Hypersomnia with sleep apnea, unspecified
Carcinoma in situ of breast	Paralysis agitans
Family history of malignant neoplasm of breast	Hip joint replacement by other means
Other specified disorder of breast	Edema
Benign neoplasm of breast	Other dyspnea and respiratory abnormality
Acquired absence of breast and nipple	Body Mass Index 4

[a]These specialties are represented by 10 diagnoses most likely to be generated by these 2 diagnosis topics.



(A)Similarity (B) Decision Trees (C) Random Forests (D) PCA-KNN (E) Support Vector Mach. (F) Multilayer Perceptrons

FIGURE 2 De facto diagnosis specialty recognition performance for multiclass classification on the full data set. Users with the de facto breast cancer specialty discovered by latent Dirichlet allocation are in one class; users whose taxonomy codes belong to the core de facto diagnosis specialties are in 12 distinct classes



(A) Similarity (B) Decision Trees (C) Random Forests (D) PCA-KNN (E) Support Vector Mach. (F) Multilayer Perceptrons

FIGURE 3 De facto diagnosis specialty recognition performance for multiclass classification on the full data set. Users with the de facto obesity specialty discovered by latent Dirichlet allocation are in one class; users whose taxonomy codes belong to the core de facto diagnosis specialties are in 12 distinct classes

7 | DISCUSSIONS

The recognition and discovery of de facto diagnosis specialties may assist in managing health care institutions. For instance, recognizing a new de facto specialty for a group of providers may lead to new care center's creation or enable existing operational management to better coordinate services and communication to support the providers within the group. It should be pointed out that whether the HPTCS needs to be updated or what is the most proper vocabulary for medical specialties are beyond the scope of our work. This work shows that there are ways to reuse and refine an existing vocabulary within a health care institution. Besides, this work does not suggest that we can recognize or discover every specialty. Planned future work includes exploration of other information that may be more indicative of specialties. For instance, instead of the ICD-9-CM codes used in this work, CPT codes may be worth exploring for procedures on EHRs.

8 | CONCLUSIONS

We introduced methods to leverage real-world diagnosis histories to recognize de facto diagnosis specialties. Using similarity and supervised learning models, we experimentally showed that 12 core de facto diagnosis specialties listed in the HPTCS are highly recognizable. We then proposed a de facto diagnosis specialty discovery problem under a general discovery-evaluation framework. In this framework, we used the semi-supervised learning model PathSelClus on the attributable data set and the unsupervised learning model LDA on a larger full data set for discovery. We further used the recognition models for evaluating the discovered specialties. PathSelClus discovered a de facto diagnosis specialty for breast cancer on the attributable data set. Latent Dirichlet allocation confirmed this discovery and suggested a new de facto diagnosis specialty for Obesity on the larger full data set. The potential correctness of these 2 specialties was reinforced by the evaluation results that they are highly recognizable by similarity and supervised learning models in comparison with 12 core de facto diagnosis specialties listed in the HPTCS.

ACKNOWLEDGEMENTS

Research was sponsored in part by the National Science Foundation CNS 0964392, CNS 1330491, 0964063, and 1526014 and by the US Department of Health & Human Services 90TR0003-01.

REFERENCES

1. Gupta S, Hanson C, Gunter C, Frank M, Liebovitz D, Malin B. Modeling and detecting anomalous topic access. In *IEEE International Conference on Intelligence and Security Informatics*, Pages 100–105, 2013.
2. Centers. for Medicare & Medicaid Services. Taxonomy code. http://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/MedicareProviderSupEnroll/Taxonomy.html.
3. National provider identifier. http://nppes.cms.hhs.gov/NPPES/Welcome.do.
4. The NPI taxonomy codes for psychology: APA practice organization offers guidance, advocates for change. http://www.apapracticecentral.org/reimbursement/npi/select-code.aspx.
5. Bonomo Y. Addiction medicine: a new medical specialty in a new age of medicine. *Intern Med J*. 2010;40(8):543-544.
6. Detmer DE, Munger BS, Lehmann CU. Clinical informatics board certification: history, current status, and predicted impact on the clinical informatics workforce. *Appl Clin Informat*. 2010;1(1):11-18.
7. Shulimzon TR. Interventional pulmonology: a new medical specialty. *Isr Med Assoc J*. 2014;16(6):379-384.

8. Lu X, Zhang A, Gunter CA, Fabbri D, Liebovitz D, Malin B. Discovering de facto diagnosis specialties. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, Pages 7–16, 2015.
9. Caine K, Hanania R. Patients want granular privacy control over health information in electronic medical records. *J Am Med Inform Assoc*. 2013;20(1):7-15.
10. Campos-Castillo C, Anthony DL. The double-edged sword of electronic health records: implications for patient disclosure. *J Am Med Inform Assoc*. 2015;22(e1):e130–e140.
11. Premarathne U, Han F, Liu H, Khalil I. Impact of privacy issues on user behavioural acceptance of personalized mhealth services. In *Mobile Health*, Pages 1089–1109. 2015.
12. Hedda M, Malin B, Yan C, Fabbri D. Evaluating the effectiveness of auditing rules for electronic health record systems. In *AMIA Annual Symposium Proceedings*, volume 2017. American Medical Informatics Association, 2017.
13. Soulakis ND, Carson MB, Lee YJ, Schneider DH, Skeehan CT, Scholtens DM. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. *J Am Med Inform Assoc*. 2015;22(2):299-311.
14. Zhang W, Gunter CA, Liebovitz D, Tian J, Malin B. Role prediction using electronic medical record system audits. In *AMIA Annual Symposium Proceedings*, volume 2011, pages 858–867. American Medical Informatics Association, 2011.
15. Gunter CA, Liebovitz D, Malin B. Experience-based access management: a life-cycle framework for identity and access management systems. *IEEE Secur Privacy*. 2011;9(5):48-55.
16. Zhang W, Chen Y, Gunter C, Liebovitz D, Malin B. Evolving role definitions through permission invocation patterns. In *Proceedings of the ACM Symposium on Access Control Models and Technologies*, pages 37–48. ACM, 2013.
17. Fabbri D, LeFevre K. Explaining accesses to electronic medical records using diagnosis information. *J Am Med Inform Assoc*. 2013;20(1):52-60.
18. Fabbri D, LeFevre K. Explanation-based auditing. *Proceedings of the VLDB Endowment*. 2011;5(1):1-12.
19. Martin P, Rubin AD, Bhatti R. Enforcing minimum necessary access in healthcare through integrated audit and access control. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Health Informatics, pages 946–955, 2013.
20. Menon AK, Jiang X, Kim J, Vaidya J, Ohno-Machado L. Detecting inappropriate access to electronic health records using collaborative filtering. *Mach Learn*. 2014;95(1):87-101.
21. Nimkar AV, Ghosh SK. An access control model for cloud-based emr federation. *Int J Trust Manag Comput Comm*. 2014;2(4):330-352.
22. Caine K, Tierney WM. Point and counterpoint: patient control of access to data in their electronic health records. *J Gen Intern Med*. 2015;30(1):38-41.
23. Tierney WM, Alpert SA, Byrket A, et al. Provider responses to patients controlling access to their electronic health records: a prospective cohort study in primary care. *J Gen Intern Med*. 2015;30(1):31-37.
24. 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.
25. Benesch C, Witter D, Wilder A, Duncan P, Samsa G, Matchar D. Inaccuracy of the international classification of diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*. 1997;49:660-664.
26. Hogg R, Tanis E. *Probability and Statistical Inference*. Pearson Prentice Hall; 2006.
27. Baeza-Yates R, Ribeiro-Neto B, et al. Modern Information Retrieval, *volume 463*. New York: ACM Press; 1999.
28. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 2009;11(1):10.
29. Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T. Rankclus: integrating clustering with ranking for heterogeneous information network analysis in *Proceedings of the International Conference on Extending Database Technology*, Pages 565–576, 2009.
30. Sun Y, Norick B, Han J, Yan X, Yu PS, Yu X. Pathselclus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Trans Knowl Discov Data*. 2013;7(3):11:1-11:23, Sept.
31. Zhang A, Xie X, Chang KC-C, Gunter CA, Han J, Wang X. Privacy risk in anonymized heterogeneous information networks. In Proceedings of the International Conference on Extending Database Technologies, pages 595–606, 2014.
32. Sun Y, Han J, Yan X, Yu PS, Wu T. Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment, 2011.
33. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.

How to cite this article: Zhang A, Lu X, Gunter CA, et al. De facto diagnosis specialties: Recognition and discovery. *Learn Health Sys*. 2018;2:e10057. https://doi.org/10.1002/lrh2.10057