Hybrid Knowledge Approach for Determining Health Care Provider Specialty from Patient Diagnoses

Erin Lynne Plettenberg, Jeremy Vickery

Abstract**—**In an access-control situation, the role of a user determines whether a data request is appropriate. This paper combines vetted web mining and logic modeling to build a lightweight system for determining the role of a health care provider based only on their prior authorized requests. The model identifies provider roles with 100% recall from very little data. This shows the value of vetted web mining in AI systems, and suggests the impact of the ICD classification on medical practice.

Keywords**—**Ontology, logic modeling, electronic medical records, information extraction, vetted web mining.

I. Introduction

We were approached with a challenging research question: Can the specialty or professional focus of a health care provider be determined exclusively from analysis of the diagnoses of the patients he/she treats? This information was to be used to detect anomalous accesses to electronic medical records by identifying cases of providers accessing the records of patients with conditions outside the provider's specialty. The application, a health information security system, required the specialty be determined quickly, in an environment of very limited data. The determination was to be made at record access time, based only on the health care provider's recent record access history. Patient privacy protection regulations and security constraints necessitated any personally identifying information about the medical providers (e.g., human resources records) be excluded from the analysis. An analytical system was developed to determine the specialty of each provider, given the diagnoses of the patients he/she treated, in a sample dataset provided by a large local hospital. The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) [1] was selected as the vocabulary for diagnoses because it was in currently in use at the hospital. Mappings for the tenth revision (ICD-10) were included. The CMS Medicare Specialty Codes [2] were selected as the vocabulary for provider specialty types. The architecture of the system, development of each of its components, evaluation methodology and results are described below.

II. Methodology

A. System Architecture

The system has three major components: a model of the complex relationship between provider specialty and patient diagnosis, scoring metrics and a ruleset for extracting medical records data from EPIC Clarity server. The ruleset was used exclusively for purposes of evaluation. These components were combined in a HIGHFLEET XKS (eXtensible Knowledge Server) [3], which provided text extraction capabilities, querying and rule application.

B. A Model of the Provider Specialty – Patient Diagnosis Relationship

The best model of the complex relationship between provider specialty and patient diagnosis would perhaps be one constructed by large groups of staff expert clinicians, who could, based on their experience and knowledge of the field, manually construct a ruleset reflecting the guidelines providers themselves use when determining whether to treat a patient or to make a referral. For example, only surgeons perform surgery, and a clinician could identify the set of diagnoses dispersed across the ICD-9 that would always or would likely require a surgeon's care. However, given the size of the ICD-9 (over 14,000 diagnoses—increasing by 54,000 with the ICD-10), such a manual catalog would be a major undertaking.

No experienced clinicians were available to assist with this project. Expert clinical information was available, however, in the professional literature published online by medical colleges and medical billing specialists to assist providers in selecting the correct ICD-9 code for the patients in their care.

Guidance was also available from pharmaceutical and medical device manufacturers, who publish diagnosis coding materials for the providers who use their products to assist them with billing. Information relating medical disciplines to diagnosis codes was also available from insurance companies and state Medicaid agencies, who publish their restrictions for what conditions a provider of a particular type can submit for reimbursement. (For example sources see [4] (Cardiology), [5] (Gastroenterology), [6] (Surgical Pathology), and [7] (Pain Management).)

Relevant open-source materials of these types were identified, manually classified by medical specialty or specialties, harvested from the web, and fed into the text extraction model of the HIGHFLEET XKS. Each of the 118 sources was manually vetted as reputable and hand-coded for the types of providers it discussed prior to extraction.

The resulting data were used to create a model of the

Erin Lynne Plettenberg is with NonLinear Press, Parkton, MD (e-mail: lynne@nonlinearpress.com).

Jeremy Vickery is with MapFit, Washington, DC (e-mail: jeremy.vickery@gmail.com).

conditions typically treated by providers of a particular specialty. The extracted data was supplemented by a ruleset (Fig. 1) that provided keyword-based reasoning over the descriptors of the ICD-9 codes. The hierarchical structure of the ICD-9 classification was also leveraged in the ruleset, i.e., if a provider specialty was associated with a particular diagnosis, it was also associated with any sub-diagnoses.

C. Comparing Provider Records with Known Medical Specialties

A library of scoring metrics was created and added to the model to compare the diagnoses of a provider's patients with those in each modeled specialty and thereby choose a most likely specialty for each provider. The metrics included one-to-one matching, conceptual distance measures (based on the ICD-9 structure), penalties for diagnoses outside of the provider's modeled specialty, and normalization factors for overlap across the specialties and the availability of data for different provider types. The best performing metric was the "specialty count," the sum of the inverse total frequencies of each diagnosis associated with the provider, using the model as a corpus:

$$score\ (provider, j) = \sum_{d_i \in S_j} \frac{p(d_i)}{sp(d_i)} \qquad (1)$$

where $\{d_1 .. d_k\}$ is the set of unique, modeled diagnoses made by the provider, S_j is the set of diagnoses associated with specialty j by the model, $p(d_i)$ is the number of provider-patient encounters coded as d_i, and $sp(d_i)$ is the number of specialties associated with d_i by the model.

D. Evaluation

```
(<=        (providerAssociatedDiagnosis "37" "200" ?icd9)

           (and
                    (icd9 ?icd9 ?desc ?)
                    (RootCtx.substring ?item ?desc)
                    (item (listof "infant" "child" "pediatric") ?item)))
```

Fig. 1 Example of a text-based rule associating ICD-9 diagnoses with provider specialty code 37, pediatrician. "200" is a source code identifying the origin of these associations as descriptor text analysis

Evaluation data was obtained through a relationship with a large local hospital. The hospital provided an excerpt from their electronic medical records database (EPIC), and the HIGHFLEET XKS was used to extract provider-diagnosis relationships (the set of diagnoses assigned to a patient during an encounter with a provider) via a ruleset that generated queries to the relational database. Provider specialty data was not available; instead, the name of the department where the provider treated the patient was used to establish the provider's likely specialty (Fig. 2). The data was anonymized for analysis.

Of the 9,907 providers examined, 361 were associated with one or more diagnoses in the data. Of these, 335 were affiliated with a department that could be mapped to a particular specialty via the department name. The evaluation data contained very few (on average ~2.18) diagnoses per provider.

III. Results

In practice, the system evaluates an attempted patient record access event n based on the user's prior (1 .. n - 1) successful access attempts. For the purposes of that evaluation, more than the top-ranked predicted specialty may be used to determine if the attempted access is typical or anomalous. However, the cost of retrieving such information from the access logs of a large hospital records system is significant enough that the evaluation must be made from relatively few prior diagnoses, similar to the sample data. This small set could be the first few record accesses retrieved within a very short time period (as the health care provider waits for access to be granted) or be taken from a specially indexed subset of the medical record system.

Given five or more conditions treated by a provider in his/her past few appointments, the model was able to correctly predict provider specialty with 64.2% precision and 100% recall (Fig. 3). The evaluation measure used was precision-at-one, although this is perhaps more precise than the system requires. In many cases the model returned more than one correct specialty for a single provider. This is perhaps due to the hospital's large number of interdisciplinary departments, where providers from various specialties treat patients with a particular disorder or class of disorders as a team. Although the model described only 45.9% (6,685) of the 14,567 diagnoses in the ICD-9, 78.2% (1,109) of the 1,418 diagnoses in the evaluation data were modeled. This suggests that the open-source literature accurately reflects the diagnosis codes most commonly used by providers in practice.

Department Name
"EMC PLASTIC SURGERY"
"**** PEDS PLASTIC SURG"
"**** PLS MELANOMA"
"***** PLASTIC SURGERY"
"*** PROV PLAS SURG"
"*** PLASTIC SURGERY"
"** SOM PLASTIC SURGERY"
"** SOM COSMETIC"
"EMC PED PLASTIC SURGERY"
"*** SOM PLASTIC SUR2"
"**** PLASTICS BREAST"
SUMMARY PLASTIC SURG

Fig. 2 Affiliations indicating provider specialty code 24, plastic surgery. Identifying information is replaced with '**'

Precision was shown to increase with the number of diagnoses considered. However, in the vast majority of cases an accurate prediction of provider specialty could be made from very little data (2 patient diagnoses or less). One hundred ninety of the 335 providers examined were correctly classified from a single diagnosis.

IV. Discussion

These results suggest that open-source information from professional sources can effectively create an expert system in the absence of subject matter experts *or* training data. The process of identifying reputable open source material for web harvest was far less time-consuming than constructing a model by eliciting expert knowledge, and was completed in under 8 hours.

While initially we were concerned that the manually identified vetted sources would not provide sufficient information to build a robust model, these concerns proved unfounded, as such a large percent of the actual diagnoses were effectively modeled. More sources pertaining to less common conditions, treatments and specialties could easily be identified and harvested to increase the "raw" coverage of the model and capture for less frequent diagnosis codes.

Additionally, the data harvested from the web provided a corpus for tf-idf scoring in addition to a model of associations. This knowledge of the frequency or popularity of each diagnosis turned out to be crucial for measuring the overlap between the various medical specialties so that it could be accounted for when choosing a match. Such information could also have been extracted from the audit log itself, but would in that case be more subject to seasonal variation, etc. and less reliable.

The ICD, originally designed as a classification for causes of death in the 1930s [8], has become the standard for diagnosis coding in the United States. A community of interest has developed around it and other standard classifications used in medical billing. (See for example ahima.org.) This community provides guidance to health care providers and trains medical coding specialists in the classifications' use, as medical coding has become an industry in its own right. (See for example optumcoding.com, supercoder.com.)

More significantly, insurance companies and the medical regulatory community rely on the ICD to make medical coverage determinations. (See for example United HealthCare Services, Inc. "Code Summary for Providers. Preventative Services – Health Care Reform." 2015.) Much of the diagnostic coding guidance used by health care providers is designed to help them meet the standards set by medical insurance providers [9]. Assessing whether the diagnoses suggested by the open-source literature were used more commonly in practice because they describe conditions that occur more frequently in patients or because they appear more frequently in the literature is outside the scope of this study.

The results also establish some limitations of this approach. In general, diagnosis coding guidelines aim to be comprehensive, causing more relationships than necessary to be established in the model. Additionally, some types of health care providers (e.g. intensive cardiac rehabilitation specialist, oncology phlebotomist) treat the same types of patients as other types (cardiologist, oncologist), and can only be distinguished from them by considering the procedures they perform. An expansion of the model to included procedure codes as well as diagnosis codes would address this issue and likely improve the precision of the system.

V. Application

This system was developed to detect anomalous accesses to electronic medical records as part of a health information security system. The system was implemented at record-access time, to determine if patient records a provider attempted to access were within his/her typical medical focus. In early testing (with a separate data set), it correctly classified 87% of access attempt events as either typical or atypical.

Such a system capable of determining the specialty of a medical provider from limited and privacy-protected information about patient diagnosis codes also has applications to medical billing fraud detection. The model of the health care provider – patient diagnosis relationship has further applications to health care policy and quality improvement. Further uses of this approach are discussed in the section on Implications.

The model includes mappings to support SNOMED-CT and ICD-10 diagnosis coding. This broadens not only the possible applications of the system (the hospital in our evaluation used ICD-9 diagnosis codes, but other diagnosis coding models may be more common elsewhere) but also the corpus of source documents that can be drawn from the web. For this implementation, only sources referencing ICD-9 diagnosis codes were used, because the mappings were developed in parallel with the model of the provider-diagnosis relationship. However, subsequent applications need to limit themselves to ICD-9 references, but could also incorporate those referencing only ICD-10 or SNOMED-CT. This is important because it greatly increases the pool of data the model can draw from without introducing any of the ambiguity or special processing issues implicit in a natural language or keyword-based approach. Finally, any vocabulary for health care provider specialties could be used, provided a mapping to the CMS Medicare Specialty Codes was available.

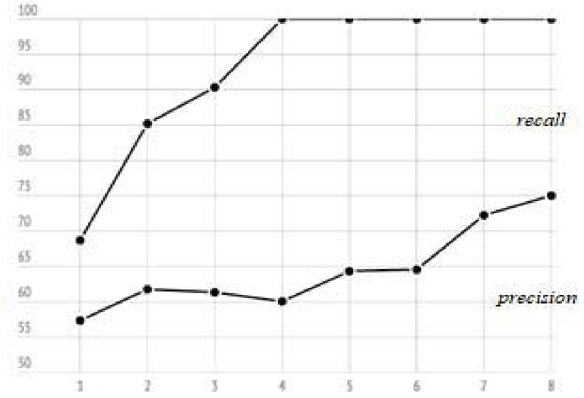

Fig. 3 Recall and precision at each data requirement threshold

VI. Implications

The maturity of open-source information in this subject area made it possible to create an accurate and surprisingly complete model of patient diagnosis by health care provider specialty with a minimum of manual effort and readily available technical tools (web harvesting, text extraction and ruleset application). Although the constraints of the application to patient medical records access control were somewhat unusual, the process and vetted web mining approach may be applicable in a variety of fields.

Vetted web mining harvests information that is both structured and authoritative. Therefore, its use did not present the possibilities for error implicit in natural language processing or mining raw or crowd-sourced data. This suggests that similarly useful “expert-models-without-the-expert” could be developed with this approach in any number of fields, such as finance, cybersecurity, or retail commerce. Prerequisites for this approach include a) an agreed-upon terminology like the ICD, and b) a trusted community of interest.

References

[1] Centers for Medicaid and Medicare Services. “International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM).” 2011.

[2] Centers for Medicaid and Medicare Services. “Medicare National Coverage Determinations (NCD) Manual.” 2015.

[3] C. Palmer, N. Chungoora, R. I. M. Young, A. G. Guendran, Z. Usman, K. Case, and J. A. Harding. "Exploiting unified modelling language (UML) as a preliminary design tool for Common Logic-based ontologies in manufacturing." International Journal of Computer Integrated Manufacturing, vol. 26, pp. 267-283, 2013.

[4] N. Maguire. “Cardiology Coding Unmasked - Part 4 Cardiology Diagnosis Coding, ICD-9 and ICD-10.” Codeapedia, no date. Available http://codapedia.com/article_592_Cardiology-Coding-Unmasked-Part-4-Cardiology-Diagnosis-Coding-ICD-9-and-ICD-10.cfm. Accessed April 24, 2015.

[5] P. A. Myer, A. Mannalithara, G. Singh, G. Singh, P. J. Pasricha, and U. Ladabaum. “Table A1. List of ICD-9-CM codes for GI disease definition.” From “Clinical and Economic Burden of Emergency Department Visits Due to Gastrointestinal Diseases in the United States.” American Journal of Gastroenterology, vol. 108, pp. 1496-1507, September 2013. Available http://www.nature.com/ajg/journal/v108/n9/fig_tab/ajg2013199t6.html. Accessed April 24, 2015.

[6] P. A. Humphrey, L. P. Dehner, and J. D. Pfeifer. “ICD-9 Codes for General Surgical Pathology.” The Washington Manual of Surgical Pathology, 2012. Available http://hemaonco.tmu.edu.tw/ ICD9.pdf. Accessed April 25, 2015.

[7] A. Howard. “Coding for Pain.” For The Record , vol. 10, p. 38, 2007. Available http://www.fortherecordmag.com/archives/ftr_05292007p38.shtml. Accessed April 25, 2015.

[8] World Health Organization. “History of the Development of the ICD.” No date. Available http://www.who.int/classifications/icd/en/HistoryOfICD.pdf. Accessed April 20, 2015.

[9] M. Pippenger, R. G. Holloway, and B. G. Vickrey. “Neurologists’ use of ICD-9CM codes for dementia.” Neurology, vol. 56, pp. 1206-1209, May 2001.