Indexing Electronic Medical Records Using a Taxonomy

by John Kuranz and Barbara Gilles

EDITOR’S SUMMARY
With the move toward patient electronic medical records (EMRs), accessing information for insurance coding and research depends on standardized taxonomies to organize and index the content. Controlled vocabularies are necessary to interpret content consistently. Established quasi-taxonomies provide codes for medical conditions and treatments, but applying these codes as metadata to index the records is laborious, requiring translation from natural language in the EMR to a code’s verbal equivalent to the code. Indexing systems can streamline the categorization process for greater efficiency and accuracy by using Bayesian engines or a rule-based approach. Analyzing discrepancies between human indexing and the software system results shows where editorial intervention is needed for continual improvement, with a goal of 85% or higher accuracy. Using a categorization system with a hierarchical taxonomy enables deep, precise indexing or quick, automatic filtering to more general concepts. The accuracy of medical indexing systems varies widely, based on the degree of automation and capacity for semantic analysis.

KEYWORDS
medical records knowledge bases
subject indexing semantic analysis
machine aided indexing taxonomies

The use of electronic medical records (EMRs) has increased dramatically over the past few years. In the health industry and in medical research, EMRs have proven to be a valuable source of information for diagnostic research, coding enhancement for billing and insurance purposes and understanding of an overall patient encounter. Studies to date have revealed that the text content in EMRs may contain important additional information relevant to outcomes, concomitant diseases, procedures, interventions or test results in observational studies. However, completely manual review of EMRs is time-consuming, subjective and inefficient. Furthermore, manually applying codes and/or terms as metadata to index records is laborious. The solution to these problems is the use of machine-aided indexing software to extract diagnostic and other clinical information.

Because EMR text can contain a wide range of complex language structures, terminology and context-specific abbreviations and acronyms, the software must be able to handle those aspects of clinical language. To provide a foundation for such complex language analysis, an organization needs to select or develop an appropriate taxonomy. This provides a consistent vocabulary for making the indexing consistent. At the same time, it provides a basis for rule building (in the case of rule-based indexing systems) or system training (in the case of Bayesian indexing systems).

Building or Borrowing a Taxonomy for EMR Indexing
Consider what, essentially, you want to cover in your indexing, and make sure you cover it in your taxonomy. Diagnoses, and therefore diseases, injuries, conditions and symptoms? Surgical procedures, medications and other treatments? There are coding systems and taxonomies that cover these areas.

John Kuranz is CEO, Access Integrity. He can be reached through the company website: www.accessintegrity.com.
Barbara Gilles is a taxonomist at Access Innovations, Inc. She can be reached at barbara_gilles<at>accessinn.com.
If you are creating a taxonomy from scratch, enlist the help of subject matter experts (SMEs) who are familiar with health services and medical terminology and have them suggest and review terms. You might need some generalists to spot ambiguous and overlapping terms among the various disciplines and sub-disciplines. The SMEs should also review the hierarchical structure after it has been roughed out.

Due to the needs and expectations of the medical community, your medical taxonomy will need to be very specific. From what we’ve observed, the term hierarchy will need to go at least five or six levels deep. Depending on the services (primary care, for example) and specialties (cardiology, for example) your organization covers, the overall scope might be fairly narrow or all-encompassing. For a narrower scope, you might need only 10 top terms.

One of the standard coding systems can serve as the basis for the taxonomy. The 9th edition of the *International Statistical Classification of Diseases and Related Health Problems* (ICD-9) and the American Medical Association’s *Current Medical Terminology* (CPT), among others, are large coding systems heavily used by the medical provider community.

However, these coding systems are not taxonomies per se. And this may be an advantage for associating required codes with EMRs. However, it means that things start out somewhat backwards. For “indexing” a code, it is logical and practical to leave it as a taxonomy “term,” rather than change it to a descriptive word or phrase. However, these terms are not easy for a taxonomist, automated rule builder or human indexing rule editor to deal with without some reference to semantic content. One way of adding the semantic content back in is to put it in a synonym or non-preferred term field. Figure 1 shows an example from the ICD-9 diagnosis code set, with the descriptive phrase in the synonym field coming from the ICD-9 tabular list that explains the codes.

**Software Training and Rule Building**

In a rule-based indexing system, we can easily view the synonym and write and edit rules to capture the various ways in which the concept above (“food poisoning due to Vibrio vulnificus”) might be expressed in an EMR (Figure 2).

Unfortunately, the descriptions in the medical codebooks are not always as neat and clear as in this example. Some very common problem descriptions are “Other,” “NOS” (not otherwise specified), and “NEC” (not elsewhere classified). And then there are the multiple occurrences of descriptions (such as “varicella”) that have different meanings, depending on what section they’re in.

Good taxonomists and rule builders will check the context before building the rules that differentiate among the codes for currently having a disease, having a history of the disease, having a recent exposure to the
An indexing system needs to be trained in the specific subject or vertical concept area. In rule-based systems, this training is accomplished by (1) selecting the approved list of keywords to be used and, through matching and synonyms, building simple rules; and (2) employing phraseological, grammatical, syntactical, semantic, usage, proximity, location, capitalization and other algorithms – based on the system – for building complex rules. This approach means that, frequently, the rules are keyword-matched to synonyms or to word combinations using Boolean statements in order to capture the appropriate indexing terms in the target text.

In Bayesian engines, the system begins with a training set of indexed patient records, usually 50-60 documents. The system uses the sample to associate the keywords with text, creating scenarios for word occurrence based on the words in the training documents and how often they occur in conjunction with the approved keywords for that item. Some systems use a combination of Boolean and Bayesian engines to achieve the final indexing results.

Natural language systems can be used in conjunction with the training process. They base their application on parts of speech and the nature of language usage. Language is used differently in different applications. Think of the word plasma. It has very different meanings in medicine and in physics, although the word has the same spelling, pronunciation and etymology. Therefore, the contextual usage is what informs the application.

A natural language system trains the indexing system based on parts of speech and term usage and builds a domain for the specific area of knowledge to be covered. The following natural language techniques can be used by human rule builders and editors:

- morphological (term form – number, tense, etc.);
- lexical analysis (part of speech tagging);
- syntactic (noun phrase identification, proper name boundaries);
- numerical conceptual boundaries;
- phraseological (discourse analysis, text structure identification);
- semantic analysis (proper name concept categorization, numeric concept categorization, semantic relation extraction); and
- pragmatic (common sense reasoning for the usage of the term, such as cause and effect relationships).

For the best results, at least some of the human rule editors should have medical training or medical billing experience.

**Achieving Accuracy in EMR Indexing Systems**

The accuracy of medical indexing systems varies widely, based on the degree of automation and capacity for semantic analysis. One weakness of automated indexing systems, compared to 100% human intervention, is the frequency of what are called “false hits.” That is, the concepts selected fit the computer model but do not make sense in actual use. These terms are considered noise in the system and in application. Systems work to reduce the level of noise using the process described below.

In the world of knowledge management, the measure of the accuracy of an indexing system is based on the number of hits (exact matches with what a human indexer would have applied to the system); misses (the keywords a human would have selected that a computerized system did not); and noise (keywords selected by the computer that a human would not have selected). The statistical ratios of hits, misses and noise are the measure of how good the system is. Our experience has been that, for a system to be practical and beneficial, the threshold should be at 85% hits out of a total of 100% accurate (against human) intervention. That means that noise and misses need to be less than 15% combined.

A good system will provide an accuracy rate of 60% initially from a good foundation keyword list with simple match rules and 85% or better with training or rule building. This low starting point means that there is still a margin of error expected and that the system needs – and improves
with — human review. Some systems can maintain a statistical record of discrepancies between human indexing and the automated term suggestions or assignments; ideally, the sampling and review that this approach requires should be done at least once every few months.

While human monitoring and control of the indexing of individual records is ideal, perceived economic or workflow impacts often render frequent human participation in the indexing process unacceptable. These issues generally lead to the attempt to provide some form of fully automated indexing results, so that human indexing is not required. Fortunately, there are some techniques, other than rule development or system training, that can help increase accuracy:

- The system can be configured in such a way that only the most specific terms are used; this practice prevents the application of terms that might be considered too general for the text. (For coding, though, the rules should be written so that only the most specific code numbers are used.)

- If coding is not involved, the keywords may also be “rolled up” to ever-broader terms until only the first three levels of the hierarchy are used; this procedure casts a broader net and prevents the use of terms that might be considered too specific for the context. This second approach is preferred in some environments, where popular thinking indicates that users will not go deeper into the hierarchy.

Deeper indexing and precise application of keywords still benefit from human intervention, at least by review, in all systems. The decision then becomes how precisely and deeply the user develops the indexing for the system application and the target user groups.

**Summary**

We have investigated some methodologies used in the automatic and semi-automatic classification of text in the medical field. In practice, many of the systems use a mixture of the methods to achieve the result desired. Most systems require a taxonomy in order to start, and most systems tag text to each keyword term in the taxonomy as metadata in the keyword name or in other elements. A taxonomy that has been so enhanced enables deep, precise indexing or quick and automatic filtering to more general concepts.

There are real and reasonable differences in deciding how a literal world of data, knowledge or content should be organized. Purveyors of various systems maneuver to occupy or invent the standards high ground and to capture the attention of the marketplace, but they often bring ambiguity to the discussion of process and confusion to the debate over performance.

The processes are complex and performance claims require scrutiny against an equal standard. Part of the grand mission of rendering order out of chaos is to bring clarity and precision to the language of our deliberations. In simple terms, it’s about how to bridge the distance between questions from humans and answers from systems. When the answers are in your EMRs, taxonomies and associated indexing systems can bridge the gap.