

Facilitating Quick and Better Text Searching for ICD-10-CM Codes

Hari Krishna Nandigam

Connected Health Symposium 2015

www.iproc.org/2015/1/e4

DOI: <http://doi.org/10.2196/iproc.4637>

Abstract

Background: The Center for Medicare & Medicaid Services (CMS) published ICD-10-CM files in xml format in addition to pdf format. The coding guidelines recommend using both tabular and index files for efficient and accurate coding of the medical conditions. It would pose a challenge to clinicians to not only to correctly diagnose and provide appropriate treatment to their patients, but also to search and select a right term and right ICD code. The traditional text search involves querying using keywords and browsing for answers. In the context of text search for an ICD-10 diagnosis code, browsing through irrelevant results, or finding no results may frustrate busy clinicians. Ideally, a search for ICD-10 code should lead them to a correct answer in a quick and easy fashion. Therefore when building a search application for ICD-10 coding, considering these issues would be the key to a good design.

Methods: In our first step we pre-coordinated 'terms' nested in the 'mainTerm' in the ICD-10 Index Extensible Markup Language (xml) file and made the relations between them explicit in the <title> elements using a commercially available transformation tool. The values for the 'level' attribute in the <title> ranged from 1 to 9 representing the levels of nesting. In our next step we joined the tabular and the modified index files based on their code as the key and combined into one xml file. We then loaded this combined file into the database present at the back end. The original tabular xml file from Centers for Disease Control and Prevention (CDC) was also loaded in the database for the sake of a comparative study. We understand that both tabular and index files complement the full set of ICD-10-CM. But, this study sought to use the original files as they are and it was noted that querying against the original index file wasn't helpful. We requested clinicians to use our search tool running with one instance of combined xml and another instance of original tabular xml file at their backend. We then carried out a statistical analysis of the sensitivity and specificity of both result sets for clinical relevancy using their judgment as the gold standard. Further, in order to refine the selected code, we developed a faceted search as an add-on feature to our search tool.

Results: Our preliminary results showed that querying against the database containing our combined xml file resulted in a more comprehensive and accurate diagnoses set compared to querying against to the one that contained the original tabular file. The search engine looked through the elements namely diagnosis description, Main Term, and Inclusion Terms. In some cases querying against the original tabular file resulted in null results. We also noticed adding elements such as etiology, anatomical sites and laterality to this combined file to help in faceted search.

Conclusion: We conclude that the combined tabular and index file loaded into the database results in more accurate and comprehensive diagnoses result set on querying. Our next step is to develop a faceted search to help in navigating to highly granular ICD-10-CM codes.

Introduction

The International Classification of Diseases (ICD) is the product of various international efforts to report the causes of illness and underlying causes of death [1]. It was published by World Health Organization (WHO) and traditionally reviewed every ten years. It includes classification of diseases and other health problems recorded on many types of health and vital records including death certificate and health records [2]. Today ICD is not only used for disease classification but also as a standard for reimbursement and supporting medical necessity for a procedure or treatment or service by the physician. It would be a challenge for a physician not only correctly diagnose and provide appropriate treatment to the patient, but also to search and pick a right term and right code in the ICD. Therefore for a clinician searching the ICD-10 content for the right code is the means to end [3].

Background

Usually a disease begins with a cause (etiological factor), undergo a series of pathophysiological changes, and manifest a range of clinical symptoms and signs. Predisposing factors and other contributing factors may result in either early manifestation or aggravation of the disease. Every treating physician aim at diagnosing the disease since it helps them in identifying the root cause, provide appropriate treatment and predict the prognosis of the disease. However any disease may manifest clinical conditions ranging from one end of spectrum to another; further the causes/etiology of the disease can be due to many reasons. While naming a disease condition is useful, it may be not be enough unless the cause of the disease is identified. In United States, the ICD-10 classification system provided by Center for Disease Control includes a tabular file comprising groups of diseases based on topographic site and etiology and an alphabetical index file which is a list of acceptable or approved disease terminology arranged in a simple alphabetical arrangement of disease terms; thus differs from the ideal way of classification. The ICD has been revised periodically to incorporate changes from the medical field. The Tenth Revision (ICD-10) differs from the Ninth Revision (ICD-9) in several ways although the overall content is similar: First, ICD-10 is printed in a three-volume set compared with ICD-9's two-volume set. Second, ICD-10 has alphanumeric categories rather than numeric categories. Third, some chapters have been rearranged, some titles have changed, and conditions have been

regrouped. Fourth, ICD-10 has almost twice as many categories as ICD-9. Fifth, some fairly minor changes have been made in the coding rules for mortality [2].

The reported conditions are then translated into medical codes through use of the classification structure and the selection and modification rules contained in the applicable revision of the ICD, published by the World Health Organization (WHO). These coding rules improve the usefulness of mortality statistics by giving preference to certain categories, by consolidating conditions, and by systematically selecting a single cause of death from a reported sequence of conditions.

Methods

Optimization of ICD-10 Content for Search

In the clinical world, clinicians use search feature a lot. Search functionality is quite often used for querying patients' health records or finding clinical information in the library resources.

Clinicians use the search functionality to the maximum extent when querying a code for the diagnoses that they came up with. Traditional search involve querying and browsing for answers. Because of time constraint clinicians browsing may not be an option. Instead a search for ICD-10 code should lead them to the correct answer in quick and easy fashion. Unlike searching a piece of clinical information for knowledge, clinicians may search the ICD-10 content only for the sake of assigning a code and reimbursement. "People assume that good doctors are good searchers, but that's not true at all. [3]" Hence when building a search application for ICD-10 coding, taking these issues into consideration would be a key to good design.

Because of the strengths of xml, we used ICD-10 content available in xml format for querying and searching. The data in ICD-10 xml files are very structured. CDC represented the nomenclature of diseases in index table which is available as index files that comply with the schema called index.xsd. It represented the classification of diseases in tabular format which is available as tabular files for diseases, drugs and neoplasms that comply with the schema called tabular.xsd. Several of the diagnoses contain a chunk of text as a way to explicitly describe the details of the disease. It is particularly important to have xml query languages to select records from structured elements of the xml document as well as search for information in the text of the disease [4]. We tried to identify certain keywords in the text of the diagnoses such as anatomical sites, severity, laterality, etiology, manifest, visit type etc and develop xml elements for their corresponding diagnoses and explicitly indicate these identified keywords for each diagnosis. Since xml is extensible we were able to accommodate these features to the diagnoses. We identified that such an explicit identification of the elements help in querying and searching features to our user interface.

Since drilling down to a fine granular level of diagnoses up to a 7 digit code is recommended by CDC for getting reimbursed, it is essential for a clinician to have an interface that can lead them to the right approach and right level of granularity. When a physician enters a diagnosis of ‘Panniculitis’ then an interface that give the physician the option of selecting the anatomical site help them drill to a specific granularity level. Identifying these key clinical words that is explicitly indicated to their corresponding diagnoses would help. We also recommend the user interface to facilitate a tree based model approach where the key clinical words exists at the node points and drive the clinician to pick the accurate code.

In its guidelines to accurately code the diagnoses, CDC recommends to first locate the term in the alphabetical index and then verify the code in the tabular list. Say for example in the case of chronic renal failure, the main term is the failure and should be searched in the alphabetical index file and then drilled down to find the term – ‘renal’ and get the code. The code is then refined to further granularity by searching in the tabular List. We tried to develop an interface that facilitates this whole step wise process seamlessly.

Pre-Coordination of Terms in the Index ICD-10 Files

The Index.xml file is highly structured and when it comes to processing regular path expressions queries, it is fairly inefficient. It was also reported that in highly nested hierarchical structures, the overhead of traversing the hierarchy of xml data can be substantial. Further the situation gets aggravated if the path lengths are very long and unknown [5]. There are several studies that worked on the optimization of regular path expressions in object oriented databases [6]. As a way to optimize the query search we consolidated certain nodes in the ICD-10 Index file by doing an element-to-element join and changed the highly structured data to a semi structuring data. Particularly for searching paths that are very long or whose length is unknown, element-to-element join is highly effective [5]. As a result the nodes contained rich context and are more expressive and can be picked up by using regular expressions through the key word search. In our xqueries we used regular expression terms such as contains and matches. Figure 1 shows the XML transformation using Altova’s Map Force and Figure 2 shows a comparison of the index ICD-10 files before and after transformation.

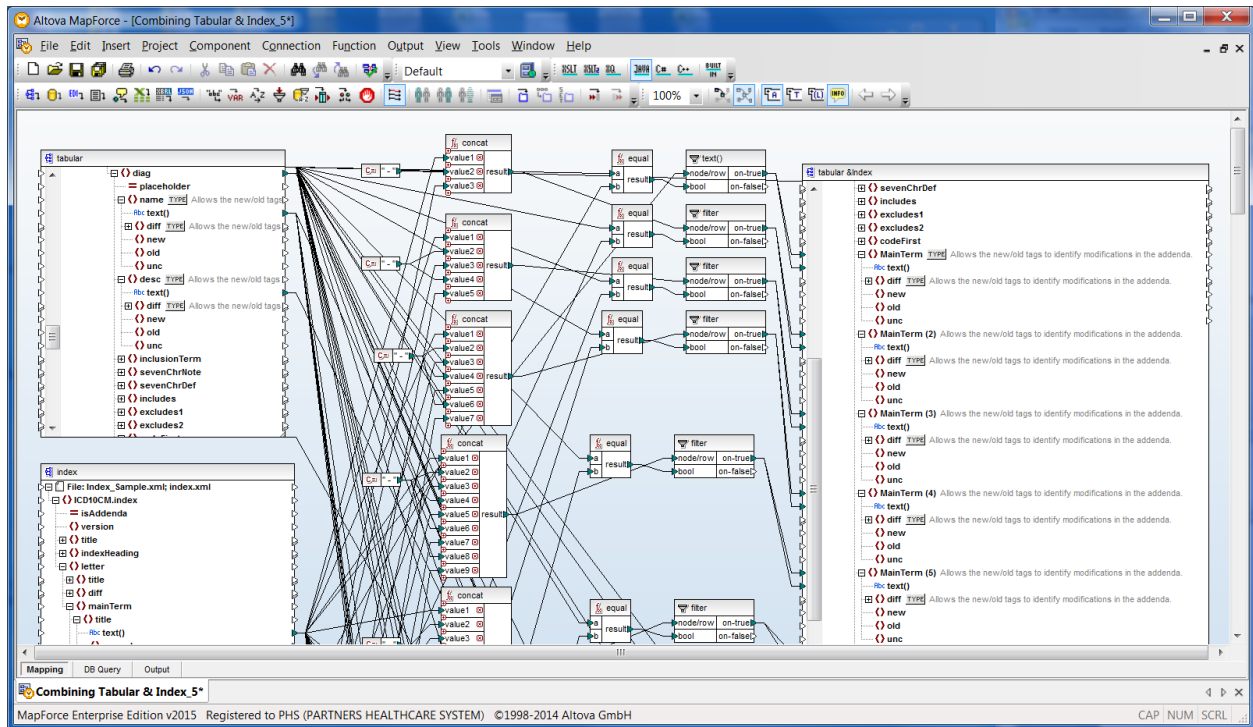


Figure 1. XML transformation using Altova's Map Force.

<pre> <mainTerm> <title>Typhoid</title> <code>A01.00</code> <term level="1"> <title>with pneumonia</title> <code>A01.03</code> </term> <term level="1"> <title>abdominal</title> <code>A01.09</code> </term> <term level="1"> <title>arthritis</title> <code>A01.04</code> </term> <term level="1"> <title>spine</title> <code>A01.05</code> </term> <term level="1"> <title>specified NEC</title> <code>A01.09</code> </term> <term level="1"> <title>ulcer<nemod>(perforating)</nemod></title> <code>A01.09</code> </term> </pre>	<pre> <mainTerm> <title>Typhoid</title> <code>A01.00</code> <term level="1"> <title>Typhoid --> with pneumonia</title> <code>A01.03</code> </term> <term level="1"> <title>Typhoid --> abdominal</title> <code>A01.09</code> </term> <term level="1"> <title>Typhoid --> arthritis</title> <code>A01.04</code> </term> <term level="1"> <title>Typhoid --> spine</title> <code>A01.05</code> </term> <term level="1"> <title>specified NEC</title> <code>A01.09</code> </term> <term level="1"> <title>ulcer<nemod>(perforating)</nemod></title> <code>A01.09</code> </term> </pre>
---	---

Figure 2. A comparison of the Index ICD-10 files before and after transformation.

Results

Federated Search by Merging/Joining the Tabular and Index ICD-10 Files

The consolidated and more expressive terms in the alphabetical Index and the tabular list are merged to form one xml file basing the ICD-10 code as their key. Since the clinician may either enter the terms in the Index file or a full specific term located in the tabular list, we merged the index file and tabular file. This would enable fast, powerful and unified queries. The resulting xml file contained new elements called Index designation a shown in the Figure 3. Please note the Index designation need not necessarily mean synonyms of the diagnosis. Say for example, the clinician may enter 'typhoid of the spine' or 'typhoid of any bone' and the result shows up as 'Typhoid Osteomyelitis', since ICD-10 tried to identify it as a single type of diagnosis for both conditions.

```

<diag>
  <name>A01.04</name>
  <desc>Typhoid arthritis</desc>
  <MainTerm>Typhoid -> arthritis</MainTerm>
  <MainTerm>Arthritis, arthritic -> due to or associated with -> typhoid fever</MainTerm>
  <MainTerm>Arthritis, arthritic -> in -> typhoid fever</MainTerm>
</diag>
<diag>
  <name>A01.05</name>
  <desc>Typhoid osteomyelitis</desc>
  <MainTerm>Osteomyelitis -> typhoid</MainTerm>
  <MainTerm>Spondylitis -> typhosa</MainTerm>
  <MainTerm>Typhoid -> osteomyelitis</MainTerm>
  <MainTerm>Typhoid -> spine</MainTerm>
</diag>
<diag>
  <name>A01.09</name>
  <desc>Typhoid fever with other complications</desc>
  <MainTerm>Post -> typhoid abscess</MainTerm>
  <MainTerm>Typhoperitonitis</MainTerm>
  <MainTerm>Abscess -> post -> typhoid</MainTerm>
  <MainTerm>Cholangiolitis -> typhoidal</MainTerm>
  <MainTerm>Cholecystitis -> typhoidal</MainTerm>
  <MainTerm>Cholelithiasis -> typhoidal</MainTerm>
  <MainTerm>Fistula -> typhoid</MainTerm>
  <MainTerm>Rheumatic -> typhoid fever</MainTerm>
  <MainTerm>Typhoid -> abdominal</MainTerm>
  <MainTerm>Typhoid -> cholecystitis</MainTerm>
  <MainTerm>Typhoid -> mesenteric lymph nodes</MainTerm>
  <MainTerm>Typhoid -> perichondritis, larynx</MainTerm>
  <MainTerm>Typhoid -> specified NEC</MainTerm>
  <MainTerm>Typhoid -> ulcer</MainTerm>
  <MainTerm>Glomerulonephritis -> in -> typhoid fever</MainTerm>
  <MainTerm>Lymphadenitis -> mesenteric -> due to Salmonella typhi</MainTerm>
  <MainTerm>Nephritis, nephritic -> due to -> typhoid fever</MainTerm>
  <MainTerm>Perichondritis -> larynx -> typhoid</MainTerm>
</diag>
...

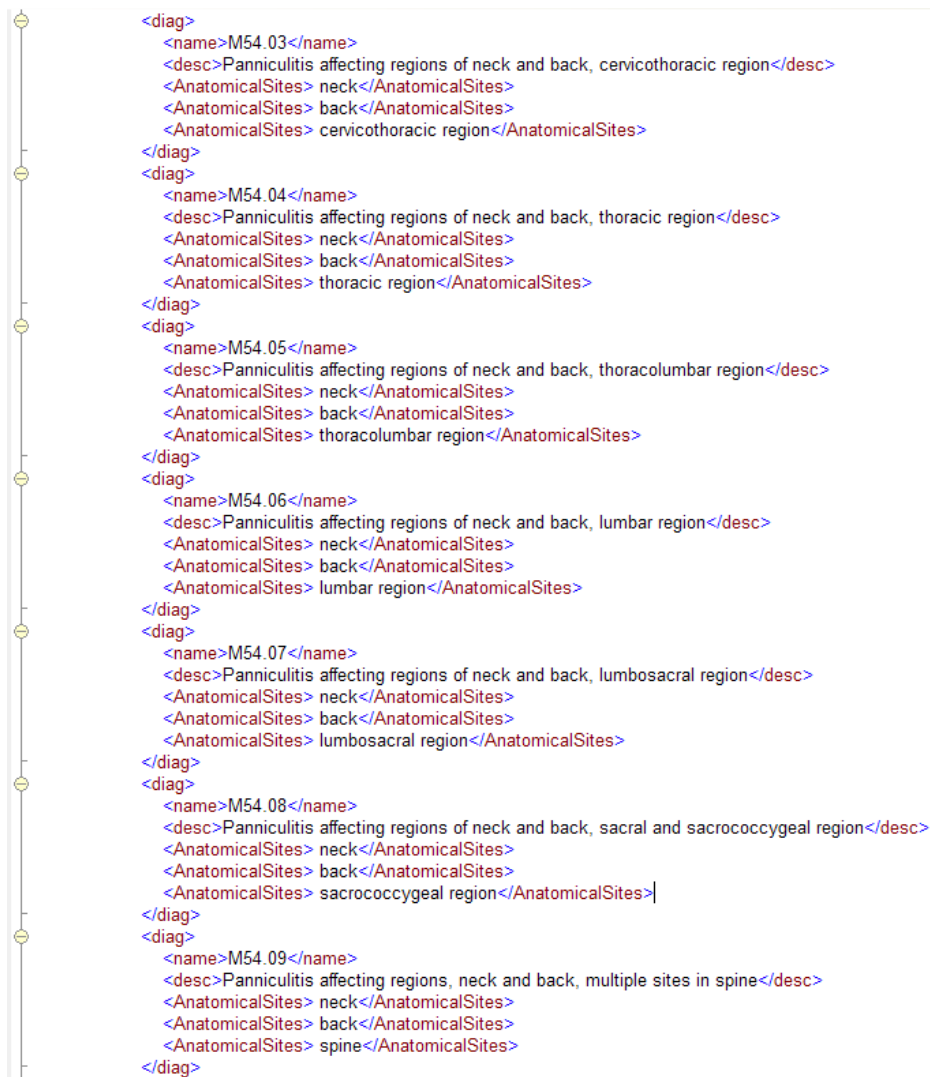
```

Figure 3. An output file created by merging/joining the tabular and Index ICD-10 files.

Faceted Search of the Tabular ICD-10 Files by Adding Elements

It is not an exaggeration to say that a search that is an iterative and interactive process has the power to suggest, define, refine, relate and educate [3]. Faceted search leverage metadata fields and values in order to provide users with visible option for clarifying and refining queries. Because most of the diagnoses expressively include etiology, anatomical site, laterality etc with the name itself, we tried to add new elements to expressively indicate them. Say for example, an element- Anatomical site has been created for ‘panniculitis affecting region of neck and back,

cervicothoracic region’ to explicitly indicate that the anatomical site here is neck and back, cervicothoracic region. Such a strategy help especially in a user interface where the clinician is trying to refine diagnoses of panniculitis. Once the clinician enters the diagnoses of ‘panniculitis’ in the search box of the application, the user interface would give the option to select the anatomical sites and give all the available anatomical sites for that diagnoses. Creating attributes to develop a faceted search is presented in Figure 4.



```

<diag>
  <name>M54.03</name>
  <desc>Panniculitis affecting regions of neck and back, cervicothoracic region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> cervicothoracic region</AnatomicalSites>
</diag>
<diag>
  <name>M54.04</name>
  <desc>Panniculitis affecting regions of neck and back, thoracic region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> thoracic region</AnatomicalSites>
</diag>
<diag>
  <name>M54.05</name>
  <desc>Panniculitis affecting regions of neck and back, thoracolumbar region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> thoracolumbar region</AnatomicalSites>
</diag>
<diag>
  <name>M54.06</name>
  <desc>Panniculitis affecting regions of neck and back, lumbar region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> lumbar region</AnatomicalSites>
</diag>
<diag>
  <name>M54.07</name>
  <desc>Panniculitis affecting regions of neck and back, lumbosacral region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> lumbosacral region</AnatomicalSites>
</diag>
<diag>
  <name>M54.08</name>
  <desc>Panniculitis affecting regions of neck and back, sacral and sacrococcygeal region</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> sacrococcygeal region</AnatomicalSites>
</diag>
<diag>
  <name>M54.09</name>
  <desc>Panniculitis affecting regions, neck and back, multiple sites in spine</desc>
  <AnatomicalSites> neck</AnatomicalSites>
  <AnatomicalSites> back</AnatomicalSites>
  <AnatomicalSites> spine</AnatomicalSites>
</diag>

```

Figure 4. Creating attributes to develop a faceted search.

Our preliminary results showed that querying against the database containing our combined xml file resulted in a more comprehensive and accurate diagnoses set compared to querying against to the one that contained the original tabular file. In some cases querying against the original tabular file resulted in null results.

Conclusion and Future Work

Federated search helped providing us relevant results. Faceted search is powerful and most often helps in navigating to highly granular 7 digit ICD-10 code. Because several of the metadata required for faceted search is not provided by CDC, it takes effort to identify the metadata based on the type of clinical practice. Identifying the right metadata is the key challenge [7]. Once metadata and its values are identified they need to be incorporated in faceted search algorithm. To improve our performance and quick search, we plan to work on the indexing the pre-coordinated elements. It is almost never that any clinician needs to know the structure of the schema of the index file. Since clinicians most often search values that have specific key words, we assume that value index and text index could be useful to facilitate better keyword search capabilities [8]. We also assume that creating an inverted file for the index.xml could facilitate speedy search [4].

One of the limitations of element-to-element join approach is the lack of flexibility. Each time when CDC provides an updated Index file, the join process needs to be computed particularly when new nodes are inserted between the existing elements.

Competing Interests

None declared.

Ethics Approval

No IRB approval was required for this study.

References

1. Moriyama IM, Ph.D., Loy RM, MBE , Robb-Smith AHT, M.D. History of the Statistical Classification of Diseases and Causes of Death 2011.
2. Grider DJ. Principles of ICD-10-CM Coding. American Medical Association.: American Medical Association.
3. Morville P, Callender J. Search Patterns: Design for Discovery: O'Reilly Media; 2010.
4. Florescu D, Kossmann D, Manolescu I. Integrating keyword search into XML query processing. Computer Networks. 2000 (33):119–35. Elsevier Science.
5. Li Q, Moon B, editors. Indexing and Querying XML Data for Regular Path Expressions. Proceedings of the 27th VLDB Conference; 2001; Roma, Italy.
6. McHugh J, Widom J, editors. Query Optimization for XML. Proceedings of the 25th VLDB Conference; 1999; Edinburgh, Scotland.

7. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology behind Search. 2 edition ed. New York: Addison-Wesley Professional; 2011 2011/02/10/. 944 p.
8. McHugh J, Widom J, Abiteboul S, Luo Q, Rajaraman A. Indexing Semistructured Data. 1998 1998. Report No.