A Comparison of Two Methods for Retrieving ICD-9-CM data: The Effect of Using an Ontology-based Method for Handling Terminology Changes

Alexander C. Yu, MD, MPhil, James J. Cimino, MD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

Terminology changes may affect reusability of data, hence the need for methods for managing changes. Along these lines, we have developed a formal representation of the concept-term relationship, around which we have also developed a methodology for management of terminology changes. We have implemented our methodology in a terminology maintenance tool. To evaluate our methodology, we compared two methods for retrieving ICD-9-CM data, based on their recall when retrieving data for ICD-9-CM terms whose codes had changed but which had retained their original meaning. Our results show that recall is either the same or better with a retrieval method that takes into account the effects of terminology changes. Statistically significant differences were detected ($p<0.05$) with the McNemar test for two terms whose codes had changed. Furthermore, when all the cases are combined in an overall category, our method 2 also performs statistically significantly better than default method ($p < 0.05$).

Introduction

Most existing controlled terminologies can be characterized as collections of terms that are arranged in a simple list or organized in a hierarchy. These terminologies are useful for standardizing terms and encoding data and are used in many existing information systems. Therefore, large amounts of data have been recorded using these terminologies. Moreover, these terminologies evolve over time in order to suit the needs of users. As has been described before, there are a number of types of changes that occur in terminologies that can have an effect on the reusability of data that are encoded with these terminologies [1]. For example, when a *major name change* occurs, the term associated with a code changes so much that the meaning is essentially different, even as the code remains the same. Other types of changes include *simple addition, refinement, deletion due to obsolescence, deletion due to redundancy, minor name change, precoordination, disambiguation, code change,* and *code reuse*. In this article we focus on *code change*, which occurs when the code for a term changes but the term's meaning remains exactly the same. For example, in the 2005 version of ICD-9-CM, the code for "Meconium aspiration syndrome" is 770.1, but in the 2006 version, its code has been changed to 770.10. Recall can be decreased when searching for ICD-9-CM-encoded data because when the code for a term changes, a retrieval method that does not properly manage the code change may miss cases that are encoded with the new code for the term.

Formal Representation of the Concept-Term Relationship

We have developed the *ConceptTermRelation* method, which is based on a formal representation that captures information about the relationship between concepts and terms. This representation is meant to be used in conjunction with a domain ontology constructed according to formal ontological principles. While the domain ontology serves as a representation of the things in a domain, the *ConceptTermRelation* methodology is used to represent associations between terms and concepts.

The concept-term relationship itself is represented as a reified *ConceptTermRelation* concept, which has as its attributes:

(1) the *hasCode* attribute that is filled by the terminology's code for the concept.
(2) the *hasTerm* attribute that is filled by the terminology's term for the concept.
(3) the *hasStartDatetime* attribute that is filled by the date when the particular code and term begin to be used or associated with each other in the terminology.
(4) the *hasEndDatetime* attribute that is filled by the date when the particular code and term case cease to be used or associated with each other in the terminology.

Specific relationships between concepts and terms are represented by instantiating the *ConceptTermRelation* concept as particular *ConceptTermRelation* instances. Figure 1 shows how we defined the *ConceptTermRelation* concept in OWL (Web Ontology Language)[2], and Figure 2 illustrates the

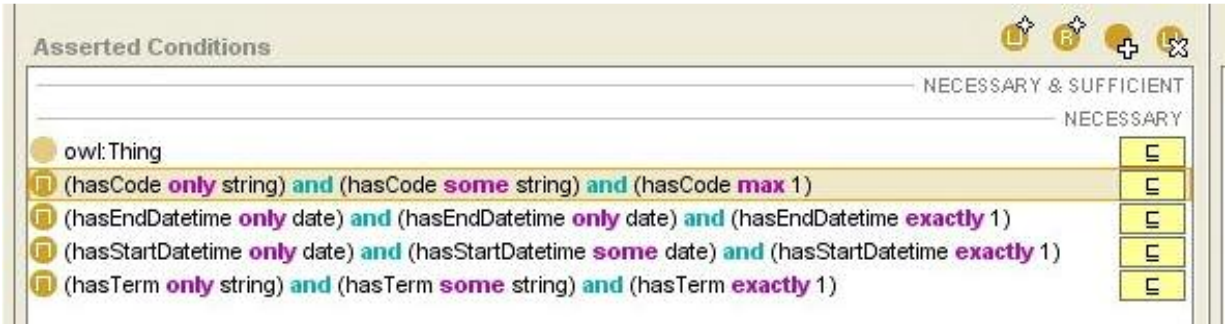


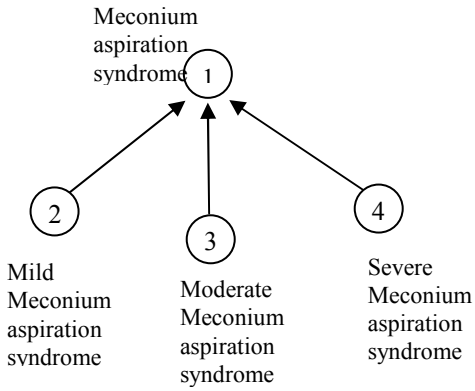
Figure 1. Representation of the *ConceptTermRelation* concept in OWL using the OWL editing environment in Protégé.

use of the *ConceptTermRelation* methodology to represent and handle inferences about code changes.

We have implemented our methodology in a terminology maintenance tool as an extension to the Protégé ontology editor. We used the tool to interface a domain ontology derived from SNOMED CT (2005) and parts of ICD-9-CM (1997 version). We also used the tool to handle successive changes to ICD-9-CM (1998-2006). Then, as part of our evaluation of the methodology, we compared two information retrieval methods based on their performance on retrieving ICD-9-CM-encoded data after the occurrence of a code change. Method 1 ("Default") was based on a static view of the ICD-9-CM terminology (i.e. without regard for changes in concepts, terms, and codes); this is the approach typically used in retrospective studies [3-5]. Method 2 ("Informed") was based on an approach for managing ICD-9-CM that took into account the effects of terminological changes using an ontological view. In the case of a code change, Method 2 utilizes the information about the code change, so that searches against an ICD-9-CM-encoded database will retrieve cases with the new code as well as the old code. Since Method 1 does not utilize this information, only cases encoded with the old code will be retrieved.

Methods

ICD-9-CM-encoded clinical data were obtained from the Columbia University Medical Center (CUMC) clinical data repository (January 1, 1997 – May 1, 2006). All data were formatted and deidentified using Perl scripts and stored on a secure database running



ConceptTermRel A
hasConceptTermRelation(-): 1
hasCode: 770.1
hasTerm: Meconium aspiration syndrome
hasStartDate: 10/01/1997
hasEndDate: 9/30/2006

ConceptTermRel B
hasConceptTermRelation(-): 1
hasCode: 770.10
hasTerm: Meconium aspiration syndrome
hasStartDate: 10/1/2006
hasEndDate:

Figure 2. An example that illustrates the use of *ConceptTermRelation*. On the left is a taxonomic hierarchy. Solid arrows stand for subsumption (is-a) relations. To the right of the graph are the *ConceptTermRelation* instances (ConceptTermRel A and B) for some of the concepts. The *hasConceptTermRelation* relations (shown as inverses) link concepts to *ConceptTermRelation* instances. There are two *ConceptTermRelation* instances shown for concept 1. Although the terms for concepts 2-5 are not included in ICD-9-CM, the ontology includes these concepts, and their instances are coded with "770.1" (before 10/1/2006) and "770.10" (from 10/1/2006 onwards) because they inherit the link to *ConceptTermRel's A* and *B* from Concept 1.

on MySQL. We analyzed all the changes that occurred in ICD-9-CM and categorized them according to the terminology change types originally described by Cimino [1]. Terminology changes that occurred across the 1997-2006 versions of the ICD-9-CM were listed and sorted into bins of change types.

We obtained a random sample of ten code changes from the population of terminology changes in the bin for code changes. For each code change, we generated a corpus of cases using a combination of keyword- and pattern matching-based searches to screen for cases that had the discharge diagnosis from the CUMC clinical data repository. A case was defined as a unique hospital inpatient course that occurred over a period of time beginning at the admission date and ending at the discharge date. Therefore, the same patient could be associated with more than one case. Discharge dates had to be within the time period encompassing the consecutive 24 months (i.e., before October) preceding the code change, and the consecutive 15 months afterward. The date range was based on results of test runs of the case-retrieval program we used to screen for cases.

Cases were allocated equally among five judges. For each of the cases retrieved by the screening process, the judges reviewed the diagnosis section (also called the "impression" or "assessment" section) of the discharge summary associated with each case and made a judgment that served as the reference standard for that case. Judges were asked to determine whether any one of the ten diagnoses was made by the primary physician in each case: After reading the diagnosis section of the discharge summary, the judges determined whether the diagnosis was documented by the primary physician in the case. Possible responses included "Yes", "Maybe", "Cannot tell", and "No". A *positive case* was defined as a case where the diagnosis was given by the primary physician. A *negative case* was defined as a case where the diagnosis was not given by the primary physician. The raters' responses were dichotomized into the two categories in three different ways. This was done because it was found that there was some variation in how judges interpreted the "maybe" and "cannot tell" categories. In the first dichotomization, "Yes" responses were binned into the *positive case* category, and "Maybe", "Cannot tell", and "No" responses were binned into the *negative cases* category. In the second dichotomization, "Yes" and "Maybe" responses were binned into the *positive case* category, and "Cannot tell" and "No" responses were binned into the *negative case* category. Finally, in the third dichotomization, "Yes", "Maybe", and "Cannot tell" responses were binned into the *positive case* category, and only "No" responses were binned into the *negative case* category.

For each code change, we carried out parallel SQL queries using Methods 1 and 2 against the corpus of cases generated for recall measurement. Recall was computed as the proportion of all cases in the corpus that were classified as a positive case by the human experts and also retrieved by the method, based on the actual codes in the patient records. The recall performance of Methods 1 and 2 on each code change were compared using the McNemar Test, which is appropriate when the data consist of paired observations of nominal data [6].
Finally, in order to estimate the reliability of the measurement process, the judges were asked to give their responses to each case in a separate set of 96 cases. There were five (5) judges and three (3) categories ("Yes","Cannot tell OR Maybe yes", and "No"). Inter-rater reliability was measured using an intraclass correlation coefficient (two-way mixed, single measures, absolute agreement – corresponding to Shrout and Fleiss's ICC(3,1) [7]) using the SPSS statistical software program [8].

Results

Table 1 shows the ten selected code changes. A total of 675 cases were retrieved by the screening process and were presented to the judges. Two of the judges had finished 3 years of specialty training in Internal Medicine, 1 judge had finished 1 year of specialty training in Internal Medicine, and 1 judge had finished medical school.

Table 2 shows the results of measuring the recall performance of Methods 1 and 2 on the code changes. The results show that Method 2 performed significantly better ($p<0.05$) than Method 1 for 2 of the ICD-9-CM terms whose codes had changed (Code changes 3 and 9), regardless of how judges' responses were dichotomized in the reference standard. For a third code change (Code change 8), Method 2 performed better than Method 1 using dichotomization 3. Finally, when all the cases were combined in an "overall" category, Method 2 also performed statistically significantly better ($p < 0.05$) than Method 1. The calculated interclass correlation coefficient was sufficiently large at 0.599 (95% CI 0.503, 0.689). The results of the inter-rater reliability study show that reliability was sufficient for the judges' responses to be used as a reference standard.

Discussion

Our method builds upon previous work on using a formal analytical approach to detecting and managing

Code Change	Year	Term	Previous Code	New Code
1	2006	Meconium aspiration syndrome	770.1	770.10
2	2001	Ulcer of lower limbs	707.1	707.10
3	2005	Decubitus ulcer	707.0	707.00
4	2005	Dysplasia of cervix	622.1	622.10
5	2005	Endometrial hyperplasia	621.3	621.30
6	2005	Prolapse of vaginal wall without mention of uterine prolapse	618.0	618.00
7	2006	Urinary obstruction	599.6	599.60
8	2005	Hyperparathyroidism	252.0	252.00
9	2001	Hyperplasia of prostate	600	600.0
10	1998	Staphylococcal septicemia	038.1	038.10

Table 1. ICD-9-CM Code changes that were used in the recall study

terminology changes. Furthermore, we adopt a formal representation of terminology changes that is compatible with the widely-adopted Web Ontology Language (OWL) representation for ontologies. An ontology-based approach is that it allows us to handle changes in the terminology in a way that propagates these changes down a class hierarchy correctly (inheritance). Furthermore, the method also handles other types of terminology changes that are not solved by simply querying for the class. For example, in major name changes, the meaning of the term/code changes but the code remains the same. Our method allows for the representation and handling of these kinds of changes.

One notable result is that the recall performance of either method is low, and this finding may be explained by the fact that the automated retrieval of cases (based on incomplete coding of cases by human coders) is measured against the responses of human expert judges who were asked about specific diagnoses. It is not hard to understand why Method 2's recall performance can be significantly better for terms whose codes have changed. Even though there is bound to be some lag between the official start date of a code change in ICD-9-CM and full compliance with that change, over time, human coders will become better at using the new (and correct) code for the diagnosis. A method that did not take into account code changes would miss more cases with the correct diagnosis, since the method would not know that cases with the new code should be retrieved. On the other hand, a method that took into account the new code would retrieve cases with the old code (prior to the date when the change is enforced) as well as cases with the new code (subsequent to the date when the change is enforced).

Precision was not measured in this study, because cases were retrieved with either method based on coding by the hospital coders with the relevant codes, even if the judges did not do so based on the limited abstracts they reviewed. However, Method 2 has the potential to improve precision, because it takes code reuse into account. Code reuse is a type of terminology change that occurs in ICD-9-CM when the name associated with a code is changed in such a way as to change its meaning (the converse of a code change). Although none of the codes were affected by code reuse, this type of change does occur in ICD-9-CM. If one of the codes had been reused during the study period, Method 1 would incorrectly retrieve cases coded with that code subsequent to its change in meaning, thereby increasing the number of false positive cases and reducing precision.

While ICD-9-CM is by no means representative of all terminologies, the pervasiveness of its use in health care, as well as the fact that many of the difficulties of handling "real world" terminologies also plague ICD-9-CM, made it a good candidate for this study on the effects of properly handling terminology changes on reuse of healthcare data. ICD-9-CM-encoded data is ubiquitous, as it is currently a part of reimbursement and reporting requirements, therefore, the results of this study are applicable to a broad range of areas such as quality assurance and clinical research. One limitation of the study is that ICD did not change as drastically as it has in past years, and so we were limited to studying code changes, and the measured difference in performance was relatively small.

Code change	Dichotomization 1 (Positive=Y Negative=N,M,C)*		Dichotomization 2 (Positive = Y,M Negative=N,C)*		Dichotomization 3 (Positive = Y,M,C Negative=N)*	
	Retrieval Method 1	Retrieval Method 2	Retrieval Method 1	Retrieval Method 2	Retrieval Method 1	Retrieval Method 2
1	**54.55%** (n=11)	**54.55%** (n=11)	**46.15%** (n=13)	**46.15%** (n=13)	**42.86%** (n=14)	**42.86%** (n=14)
2	**42.86%** (n=14)	**64.29%** (n=14)	**42.86%** (n=14)	**64.29%** (n=14)	**42.86%** (n=14)	**64.29%** (n=14)
3	**49.34%*** (n=152)	**53.29%*** (n=152)	**49.67%*** (n=153)	**53.59%*** (n=153)	**50.00%*** (n=154)	**53.90%*** (n=154)
4	**37.50%** (n=24)	**37.50%** (n=24)	**37.50%** (n=24)	**37.50%** (n=24)	**38.46%** (n=26)	**38.46%** (n=26)
5	**35.50%** (n=16)	**56.25%** (n=16)	**33.33%** (n=18)	**50.00%** (n=18)	**33.33%** (n=18)	**50.00%** (n=18)
6	**20.00%** (n=15)	**20.00%** (n=15)	**20.00%** (n=15)	**20.00%** (n=15)	**18.75%** (n=16)	**18.75%** (n=16)
7	**14.29%** (n=14)	**21.43%** (n=14)	**13.33 %** (n=15)	**20.00%** (n=15)	**13.33%** (n=15)	**20.00%** (n=15)
8	**22.00%** (n=50)	**32.00%** (n=50)	**21.57%** (n=50)	**31.37%** (n=50)	**21.82%*** (n=55)	**32.72%*** (n=55)
9	**26.32%*** (n=266)	**39.85%*** (n=266)	**26.02%*** (n=269)	**39.41%*** (n=269)	**25.64%*** (n=273)	**39.19%*** (n=273)
10	**18.52%** (n=54)	**18.52%** (n=54)	**19.18%** (n=73)	**20.55%** (n=73)	**18.18%** (n=77)	**19.48%** (n=77)
Overall	**32.14%*** (n=616)	**40.91%*** (n=616)	**31.53%*** (n=647)	**40.03%*** (n=647)	**31.16%*** (n=661)	**39.67%*** (n=661)

Table 2. Recall performance of Method 1 and Method 2 on ten (10) code changes (reported in %; *n* is the number of positive cases found by the judges). Asterisks indicate a significant difference detected with the McNemar Test ($p < 0.05$). Method 2 performed significantly better for Code Changes 3 and 9, regardless of how judges' responses were dichotomized for the reference standard. Method 2 also performed significantly better for Code Change 8 under the third dichotomization scheme.
***Y = Yes, N = No, M=Maybe, C=Cannot tell**

Conclusion

When data are reused, the effects of terminology changes need to be taken into account. Our study shows that an ontology-based ICD-9-CM data retrieval method that does so performs better on recall than one that does not in the retrieval of data for terms whose codes had changed but which retained their original meaning.

References

1. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Methods Inf Med. 1996 Sep;35(3):202-10.
2. McGuinness DL, Harmelen Fv. OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/. 2004.
3. Peterson ED, Shaw LK, DeLong ER, Pryor DB, Califf RM, Mark DB. Racial variation in the use of coronary-revascularization procedures. Are the differences real? Do they matter? N Engl J Med. 1997 Feb 13;336(7):480-6.
4. Chapman WW, Dowling JN, Wagner MM. Generating a reliable reference standard set for syndromic case classification. J Am Med Inform Assoc. 2005 Nov-Dec;12(6):618-29.
5. Vickrey BG, Rector TS, Wickstrom SL, Guzy PM, Sloss EM, Gorelick PB, et al. Occurrence of secondary ischemic events among persons with atherosclerotic vascular disease. Stroke. 2002 Apr;33(4):901-6.
6. McNemar I. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12:153-7.
7. Shrout PE FJ. Intraclass correlations: Uses in assessing rater reliability. Psychol Bulletin. 1979;86:420-7.
8. SPSS I. 233 S. Wacker Drive, 11th floor, Chicago, Illinois 60606. 2006.