Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers

Michael F. Chiang, MD, MA[1,2], John C. Hwang, MD[2], Alexander C. Yu, MD, MPhil[1], Daniel S. Casper, MD, PhD[2], James J. Cimino, MD[1,3], Justin Starren, MD, PhD[1,4]

Departments of Biomedical Informatics[1], Ophthalmology[2], Medicine[3], and Radiology[4]
Columbia University College of Physicians and Surgeons, New York, NY

Abstract

SNOMED-CT has been promoted as a reference terminology for electronic health record (EHR) systems. Many important EHR functions are based on the assumption that medical concepts will be coded consistently by different users. This study is designed to measure agreement among three physicians using two SNOMED-CT terminology browsers to encode 242 concepts from five ophthalmology case presentations in a publicly-available clinical journal. Inter-coder reliability, based on exact coding match by each physician, was 44% using one browser and 53% using the other. Intra-coder reliability testing revealed that a different SNOMED-CT code was obtained up to 55% of the time when the two browsers were used by one user to encode the same concept. These results suggest that the reliability of SNOMED-CT coding is imperfect, and may be a function of browsing methodology. A combination of physician training, terminology refinement, and browser improvement may help increase the reproducibility of SNOMED-CT coding.

Introduction

Electronic health record (EHR) systems are currently used by approximately 15-20% of American medical institutions[1]. President George W. Bush established the goal of implementing a national network of computer-based medical records in his 2004 State of the Union address, and the United States Department of Health and Human Services subsequently presented a 10-year strategy for the widespread adoption of interoperable EHRs[2]. A critical challenge for implementation of these systems will be the structured representation of medical concepts using controlled medical terminologies, in order to provide infrastructure for functionalities such as data reuse and clinical decision support.

SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms) has been promoted as a reference terminology for EHR systems. In 2003, the National Library of Medicine signed a 5-year, $32 million agreement with the College of American Pathologists to make SNOMED-CT freely available to all American health care institutions and vendors. The goals of this agreement were to broaden the usage of interoperable systems, improve patient care, and improve patient safety[3].

The coverage of SNOMED-CT for representation of medical concepts has been studied in several clinical domains, and has been found to be higher than that of other controlled terminologies[4-8]. However, two additional issues must be addressed to determine whether a controlled terminology can adequately support EHR systems: (1) Agreement of coding by multiple physicians using the terminology must be sufficiently high. This is because data entry is often performed directly by physicians at the point of care, and consistent coding is essential for subsequent retrieval of aggregated information. Previous studies suggest that coding reliability may be relatively low among physicians and between physicians and professional coders[9-12]. (2) Adequate computer-based terminology browsers must be available to support coding by physicians and EHR vendors. No previous studies to our knowledge have examined coding reliability using multiple terminology browsers.

This study evaluates two aspects of coding reliability among three physicians using two SNOMED-CT browsers: inter-coder reliability (whether the same codes are assigned to one concept by multiple coders), and intra-coder reliability (whether a coder assigns the same code to one concept using different browsers). Coding agreement is analyzed for concepts from a set of clinical ophthalmology case presentations. This report builds on our previous work involving terminology coverage and reliability[5,12]. The present study focuses on coding agreement in SNOMED-CT, and is not intended to address content coverage, limitations of SNOMED-CT, or provider education in coding.

Methods

Source of Data

Five consecutive case presentations were selected from the "Grand Rounds" section of the Digital Journal of Ophthalmology, a publicly-available on-line journal (http://www.djo.harvard.edu). Each case

included seven sections: history, examination, laboratory tests, radiology results, pathology, differential diagnosis, and final diagnosis. Cases represented both outpatient and inpatient encounters. Although presentations were based on actual data, no identifying information was present. Therefore, Institutional Review Board approval was not required because this study involved only analysis of publicly-available data that were not individually identifiable.

SNOMED-CT Browsers and Coders

Two computer-based browsers were used for this study: (1) SNOMED-CT Browser 2.0 (http://snomed.vetmed.vt.edu; Virginia-Maryland Regional College of Veterinary Medicine, Blacksburg, VA); and (2) CLUE-5 (http://www.clininfo.co.uk; Clinical Information Consultancy, Reading, UK). Both browsers supported the July 2005 SNOMED-CT release.

Three physician coders participated in the study (JCH, ACY, MFC). One was a practicing ophthalmologist, and the other two were non-practicing general physicians. Two coders had extensive postdoctoral training in biomedical informatics and controlled terminologies, and the third received several months of focused training and experience in these areas. All concepts were coded initially with the first browser, and then with the second browser after a nine-month washout period.

Parsing, Coding, and Scoring of Cases

The text of each case presentation was parsed into discrete concepts by the three coders (JCH, ACY, MFC) using a uniform methodology. Multiple-word, pre-coordinated terms were considered to be a single concept when judged clinically appropriate. For example, phrases such as "essential hypertension" and "diabetic retinopathy" were parsed as single concepts. Concepts were integrated into a list of 242 unique terms, which were independently coded and scored by each coder using the two SNOMED-CT browsers based on previously published methods[4-7].

The adequacy of assignment for each concept was scored by each coder on a three-point scale: 0, if no match for the concept was found; 1, if a partial match was found; and 2, if an exact match was found. Match scoring was based on SNOMED-CT definitions, as well as coder judgment. SNOMED-CT permits generation of complex concepts through post-coordination of multiple simpler concepts. For example, "preauricular lymphadenopathy" did not exist as a concept in SNOMED-CT, but could be coded through post-coordination of the existing terms "preauricular" and "lymphadenopathy." These properly post-coordinated terms were accepted as complete matches for the purposes of this study.

Our previous work has demonstrated that ophthalmology content coverage by controlled terminologies is significantly higher for SNOMED-CT than other terminologies[5]. To measure coding reliability only among concepts for which SNOMED-CT had adequate coverage, coding agreement in this study was determined only for the subset of concepts with adequate coverage based on assignment of a match score of "2" by at least two of the three coders.

Inter-Coder Agreement

For each concept, the observed level of agreement among codes assigned by the three coders using each browser was grouped into one of three categories: *complete agreement*, *partial agreement*, and *no agreement*. This was done by two methods: (1) Based on automated determination of exact match of codes assigned by the three coders. In this method, coders were classified as having *complete agreement* when all coders assigned the same code, *partial agreement* when only two coders assigned the same code, and *no agreement* when all coders assigned different codes. (2) Based on manual review for semantic equivalence of all assigned codes by an independent practicing ophthalmologist (DSC). In this method, clinical judgment was used to classify inter-coder agreement for each concept as *complete agreement, partial agreement,* or *no agreement*.

Intra-Coder Agreement

For each coder, the intra-coder agreement for assignment of the same concepts using the two terminology browsers was compared. This was based on automated determination of exact code match, as well as on manual review for semantic equivalence by an independent ophthalmologist (DSC), as described above. This analysis was performed for the subset of concepts with adequate coverage, based on assignment of a match score of "2" by that particular coder using either browser.

Data analysis

Findings from all case presentations were combined, and categorical level of coding agreement (complete, partial, or none) among the three coders using two browsers was compared. Numerical computations were performed using a spreadsheet package (Excel 2003; Microsoft, Redmond, WA). Statistical comparison of categorical findings was performed using the chi-square test.

Table 1. *Inter-coder agreement among three coders using two browsers. Left columns show results from exact code matching, while right columns show results from manual review for semantic equivalence.*

Inter-coder Agreement	Browser #1* (n = 179)		Browser #2* (n = 210)	
	Exact†	Semantic†	Exact†	Semantic†
Complete agreement	79 (44%)	98 (55%)	112 (53%)	158 (75%)
Partial agreement	68 (38%)	65 (36%)	62 (30%)	40 (19%)
No agreement	32 (18%)	15 (8%)	36 (17%)	12 (6%)

*With exact code matching, difference in level of inter-coder agreement between the two browsers (i.e. "exact" column for browser #1 vs. "exact" column for browser #2) was not statistically significant. With manual review for semantic equivalence, difference in level of agreement between the two browsers was statistically significant ($p<0.0001$).

†When comparing analysis from exact code matching to results from manual review for semantic equivalence (i.e. "exact" vs. "semantic" columns for each browser), the differences in levels of inter-coder agreement using each browser were statistically significant ($p<0.01$ for each browser).

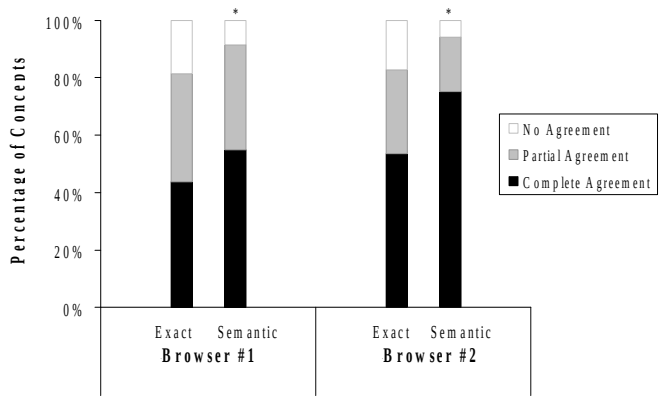
Results

Inter-Coder Agreement using Two Browsers

The overall data set consisted of 242 unique concepts from five case presentations. Based on coder match scores with each SNOMED-CT browser, the number of concepts with adequate coverage was 179 (74.0%) using browser # 1 and 210 (86.8%) using browser #2.

Inter-coder agreement for concepts judged to have adequate terminology coverage is displayed numerically in Table 1 and graphically in Figure 1. When comparing agreement among coders using manual review for semantic equivalence, the difference in level of agreement between browsers #1 and #2 was highly statistically significant ($p<0.0001$). When comparing agreement with exact code matching, there was no statistically significant difference between the two browsers. As an example of no agreement by either exact matching or semantic equivalence, the term "Horner syndrome" was coded using browser #2 as 12731000 (cervical sympathetic dystrophy) by coder 1, 271730003 (Horner syndrome of the pupil) by coder 2, and 164018003 (On examination – Horner syndrome) by coder 3.

Figure 1. *Inter-coder agreement among three coders using two browsers. Left bars show results of exact code matching, and right bars show results of manual review for semantic equivalence.*



*Statistically significant difference in level of agreement between two browsers ($p<0.0001$).

Intra-Coder Agreement using Two Browsers

Based on match scores using both browsers, adequate matches were identified for 224 (92.6%) of the 242 overall concepts by the first coder, 211 (87.2%) by the second, and 206 (85.1%) by the third.

The intra-coder agreement using the two browsers is shown in Table 2. Based on exact code matching, different codes were assigned for the same concept using the two browsers in 43% (Coder 1) to 55% (Coder 3) of cases. Based on manual review for semantic equivalence, different codes were assigned in 25% (Coder 1) to 37% (Coder 3) of cases. As an example of intra-coder disagreement, the concept "enlarged pupil" was coded by one coder as 188557000 (large pupil) in browser #1 and 405270006 (persistent mydriasis) in browser #2.

Exact Code Matching vs. Semantic Review

When comparing inter-coder reliability results based on exact code matching to results based on manual review for semantic equivalence (i.e. "exact" vs. "semantic" columns in Table 1), the differences in agreement using each browser were statistically significant ($p<0.01$ for each browser). When comparing intra-coder reliability results based on exact code matching to results based on manual review for semantic equivalence (i.e. "exact" vs. "semantic" columns in Table 2), the differences in

Table 2. *Intra-coder agreement for assignment of the same concept using two browsers. Left columns show results from exact code matching, while right columns show results from manual review for semantic equivalence.*

Agreement with Browsers	Coder 1 (n=224)		Coder 2 (n=211)		Coder 3 (n=206)	
	Exact*	Semantic*	Exact*	Semantic*	Exact*	Semantic*
Same Code	127 (57%)	167 (75%)	115 (55%)	165 (78%)	92 (45%)	129 (63%)
Different Code	97 (43%)	57 (25%)	96 (45%)	46 (22%)	114 (55%)	77 (37%)

*When comparing analysis from exact code matching to results from manual review for semantic equivalence (i.e. "exact" and "semantic" columns for each coder), the differences in levels of intra-coder agreement using Browsers #1 and #2 were statistically significant for each coder ($p<0.001$ for each coder).

agreement using each browser were statistically significant ($p<0.001$ for each coder).

Discussion

This study evaluates the reliability of concept coding among three physicians using two SNOMED-CT browsers. There are three main findings from this study: (1) Inter-coder agreement is imperfect, and is unequal when using two different SNOMED-CT browsers (Table 1 and Figure 1). (2) Intra-coder agreement for coding the same concepts using the two SNOMED-CT browsers is imperfect (Table 2). (3) Results obtained from exact code matching are different than those obtained from manual review for semantic equivalence (Tables 1 and 2).

Electronic health record systems not only support improved communication among clinicians, but also promote reuse of medical data for applications such as research, quality assurance, regulatory compliance, and public health. Structured coding of medical concepts using controlled terminologies provides the opportunity to avoid the ambiguities that are inherent in natural language[13]. However, the imperfect inter-coder agreement results from this study raise concerns about the reliability of coded medical data in these real-world situations in which codes are either assigned by physicians at the point of care, or mapped to terms used in an EHR system that have previously been coded in a controlled terminology. Our current study shows that the complete inter-coder agreement for assignment of SNOMED-CT concepts by exact code matching was only 44% when with one browser and 53% with the other (Table 1). This is consistent with the results of previous studies involving other controlled terminologies, which have suggested that coding reliability may be imperfect[10-12]. Of course, the validity of using coded medical data for purposes such as retrospective clinical research and quality assurance will be heavily dependent on the extent to which identical concepts are reproducibly represented by multiple coders.

Two SNOMED-CT browsers were compared in this study. Higher levels of inter-coder agreement were obtained when the three coders used browser #2 compared to browser #1 (Table 1). In addition, coding of the same concepts using the two browsers by the same physician resulted in different assignments up to 55% of the time based on exact code matching (Table 2). These findings show that results of SNOMED-CT coding may be significantly affected by the specific browser that is used for modeling. It is not surprising that the reliability of terminology coding appears to be dependent on the sophistication of the browser used, particularly given that the SNOMED-CT is a highly complex ontology including over 366,000 unique concepts organized into multiple hierarchies[14]. For example, computer-based browsers that are better able to recognize synonyms of the term being searched for (e.g., realizing that "hypertension" and "high blood pressure" have equivalent meanings), or to recognize "fully-specified" SNOMED-CT terms in addition to "preferred names" of concepts, may perform differently from browsers that do not perform these functions. These findings suggest that development of improved terminology browsing tools may allow physicians and professional coders to improve the reliability of SNOMED-CT coding, and that future research would be helpful to guide this process.

Agreement of coding was determined in this study by exact code matching, as well as by manual review of codes for semantic equivalence by an independent ophthalmologist. Although a full description of this topic is beyond the scope of this paper, it is instructive to examine several examples. In an analysis of discrepancies between exact matching and manual review for terms coded with browser #1, the majority were judged semantically equivalent despite discrepancies in the actual assigned codes because of:

(a) "No clinically significant difference in meaning" (e.g. "congenital ptosis of upper eyelid" [60938005] compared to "congenital ptosis" [268163008]), or (b) Post-coordination in SNOMED-CT[12]. There are important benefits to terminologies that permit expression of subtle differences in meaning. However, a disadvantage is that this may decrease inter-coder and intra-coder agreement because of difficulties in distinguishing among concepts that appear to be very similar, particularly if these decisions must be made quickly at the point of care.

Several limitations of this study should be noted: (1) Coding agreement was based only on results from three physician coders and one physician to manually assess for semantic equivalence. In addition, each coder had greater familiarity with SNOMED-CT than would be expected from an average physician. Although comparison of codes among the three physicians did not reveal any clear pattern of disagreements (data not shown), further studies involving a broader cross-section of coders may be helpful to determine the generalizability of findings. (2) The data set for this study was limited in size, and contained numerous terms from the limited domain of clinical ophthalmology. Future studies of larger data sets in additional domains will be informative. For example, studies designed to elucidate additional common characteristics of concepts that are coded either consistently or inconsistently by multiple physicians may be useful. (3) Coding of the entire set of 242 concepts in browser #2 was performed after coding in browser #1. Although there was a nine-month wash-out period between these processes, it is conceivable that this may have biased toward higher reliability with browser #2 because of increased coder familiarity with the concepts. The impact of varying the sequence and time of coding may warrant additional research. (4) This study was not designed to evaluate intra-coder agreement using the same browser at different times, or to develop strategies for improving browser design. Future research of this nature would be useful for comparison to results from the present study.

Many important EHR functions such as improved efficiency of clinical care, data analysis for retrospective research, support for prospective studies, and automated decision support are based on the assumption that medical concepts will be reproducibly coded by multiple users of electronic systems. This study raises questions about the inter-coder and intra-coder reliability of SNOMED-CT coding by three physicians using two different terminology browsers. Although additional studies are required, it is likely that a combination of physician training, terminology refinement, and development of more sophisticated browsers will help to improve coding reliability for maximizing potential benefits of EHR systems.

References

1. Grove AS. Efficiency in the health care industries: a view from the outside. JAMA 2005; 294: 490-2.
2. United States DHHS. Office of the National Coordinator for Health Information Technology (ONC). Goals of strategic framework. Available at: http://www.hhs.gov/healthit/goals.html.
3. United States DHHS. HHS launches new efforts to promote paperless health care system. Available at: http://hhs.gov/news/press/2003pres/20030701.html.
4. Chute CG, Cohn SP, Campbell KE, et al. The content of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996; 3: 224-33.
5. Chiang MF, Casper DS, Cimino JJ, Starren J. Representation of ophthalmology concepts by electronic systems: adequacy of controlled medical terminologies. Ophthalmology 2005; 112: 175-83.
6. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997; 4: 484-500.
7. Langlotz CP, Caldwell SA. The completeness of existing lexicons for representing radiology report information. J Digit Imaging 2002; 15: 201-5.
8. Campbell JR, Carpenter P, Sneiderman C, et al. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, clarity. J Am Med Inform Assoc 1997; 4: 238-51.
9. Lorence D. Regional variation in medical classification agreement: benchmarking the coding gap. J Med Syst 2003; 27: 435-43.
10. Yao P, Wiggs BR, Gregor C, et al. Discordance between physicians and coders in assignment of diagnoses. Int J Qual Health Care 1999; 11: 147-53.
11. King MS, Lipsky MS, Sharp L. Expert agreement in Current Procedural Terminology evaluation and management coding. Arch Intern Med 2002;162:316-20.
12. Hwang JC, Yu AC, Casper DS, et al. Representation of ophthalmology concepts by electronic systems: inter-coder agreement among physicians using controlled terminologies. Ophthalmology 2006; 113: 511-9.
13. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc 1998; 5: 421-31.
14. SNOMED International. SNOMED home page. Available at: http://www.snomed.org.

Acknowledgements

Supported by grants EY13972 (MFC) and LM07079 (ACY) from the National Institutes of Health, and by a Career Development Award from Research to Prevent Blindness (MFC). We thank SNOMED International for assistance with the CLUE-5 browser.