Research Paper ■

Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts

James E. Andrews, PhD, Rachel L. Richesson, PhD, MPH, Jeffrey Krischer, PhD

Abstract **Objective:** To compare consistency of coding among professional SNOMED CT coders representing three commercial providers of coding services when coding clinical research concepts with SNOMED CT.

Design: A sample of clinical research questions from case report forms (CRFs) generated by the NIH-funded Rare Disease Clinical Research Network (RDCRN) were sent to three coding companies with instructions to code the core concepts using SNOMED CT. The sample consisted of 319 question/answer pairs from 15 separate studies. The companies were asked to select SNOMED CT concepts (in any form, including post-coordinated) that capture the core concept(s) reflected in the question. Also, they were asked to state their level of certainty, as well as how precise they felt their coding was.

Measurements: Basic frequencies were calculated to determine raw level agreement among the companies and other descriptive information. Krippendorff's alpha was used to determine a statistical measure of agreement among the coding companies for several measures (semantic, certainty, and precision).

Results: No significant level of agreement among the experts was found.

Conclusion: There is little semantic agreement in coding of clinical research data items across coders from 3 professional coding services, even using a very liberal definition of agreement.

■ **J Am Med Inform Assoc.** 2007;14:497–506. DOI 10.1197/jamia.M2372.

Introduction

A major focus of clinical research informatics is the use of information technology to aid in the efficient translation and application of research findings into patient care and public health settings.[1–3] Data representation specifications are the backbone of any data collection system in clinical research, and standards for the representation of clinical data are integral to facilitating the speed and quality of clinical research, as well as meeting translational science goals.[4] While the U.S. has made strides toward identifying data standards in various areas of health care data,[5] no standards have been explicitly named in the U.S. for clinical research in general. The clinical research community has embraced the importance of data standards,[6] and is currently creating information model standards[6–8] with the intent of adopting terminological standards to use within these models. The myriad contexts and domains from which clinical research data are generated, coupled with the complexity and size of candidate terminology standards, have the potential to make implementation of data standards into the clinical research domain a challenge. Exploration of terminological implementation issues in clinical research is therefore both timely and warranted.

SNOMED CT has been recognized as a key terminology standard by various standards organizations[9–11] and is sanctioned by the Consolidated Health Informatics (CHI) initiative as the standard for diagnoses and problem lists, anatomy, and procedures.[5] Additionally, the U.S. Food and Drug Administration (FDA) recently named SNOMED CT as the standard for the highlights section of the Structured Product Labeling (SPL). The fact that SNOMED CT is clinically-rich, comprehensive, and is an emergent standard for healthcare data make it a strong candidate for clinical research contexts, as well. Practical and comprehensive guidance for constructing new and/or multi-faceted concept expressions using SNOMED CT post-coordination is forthcoming. In its absence, both new and experienced SNOMED CT users (including experts in professional companies that charge for providing coding services) have developed their own style and tools to cope, which may lead to variation in coding. Lack of coding consistency is an area of concern for advocates of data standards, because it can impede the

Affiliations of the authors: School of Library and Information Science (JEA), University of South Florida; Pediatrics Epidemiology Center (RLR, JK), University of South Florida, Tampa, FL.

The project described was supported by Grant Number RR019259 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH. The authors wish to thank Heather Guillette, MS of the Pediatrics Epidemiology Center (USF) for providing the test sample, and all of the investigators and researchers of the RDCRN, who collectively provided data collection forms that motivated this study. Also, we wish to thank the NIH Office of Rare Diseases, and Drs. Kent Spackman and Asif Syed (SNOMED, Intl., College of American Pathologists) for their helpful suggestions on the coding instructions.

Correspondence and reprint requests to: James E. Andrews, PhD, School of Library and Information Science, University of South Florida, 4202 E. Fowler Ave., CIS 1040 , Tampa FL 33620; e-mail: <jandrews@cas.usf.edu>.

Received for review: 1/10/2007; accepted for publication: 4/09/2007.

ultimate goals of implementing a data standard (i.e., variation in the use of a standard implies the lack of a standard).

This study examines the consistency of SNOMED CT coding of clinical research questions by experts from three such professional coding companies. We have generated a data set from 15 clinical studies being conducted by the NIH-funded Rare Disease Clinical Research Network (RDCRN). By determining the coding agreement among these experts we can learn more about the overall utility of SNOMED CT for coding clinical research data, as well as potential coding problem areas where more attention may be needed. The results of this study could provide a minimum estimate of the difficulties and variation in SNOMED CT when utilizing for clinical research purposes.

Background

Clinical Research Data

Representing the breadth, depth, and overall variety of data collected in clinical research, is a key challenge to identifying and properly utilizing existing data standards.[12] Clinical research encompasses a variety of data constructs, including clinical observations and findings. Typically, these data are recorded on paper data collection forms (called Case Report Forms, or CRFs) and later entered into electronic systems for storage and analysis. The data items on CRFs are generally focused to the clinical researcher to record their objective (e.g., "Pulse: ____ bpm") or subjective (e.g., "Agitated behaviors? Present/absent") findings from observations or interviews with human subjects. These CRF data items can be thought of as questions, as they are often worded as such (e.g., "Does patient show *xxx*?", "How many episodes of *xxx* did subject experience in past 6 months?"). The CRF data items differ from questions on patient self-assessment instruments, however, in that the representation of exact wording and construction of the data item, while important, does not necessarily bear the same importance or weight in clinical research as it does in areas such as psychometrics (where the item is itself an instrument and can have impact on the results).[13,14] There are no current standards for the design of CRFs or for modeling the items they contain, although there are suggestions,[15] and clinical research standards groups, such as Health Level 7 (HL7) and the Clinical Data Standards Interchange Consortium (CDISC), are beginning to discuss their importance.

Standards that capture the clinical content as represented on CRFs in clinical research will be required for interoperability of data and systems within and outside the clinical research domain. Brandt et al.[16] stress the importance of standards for representing the content of questions and questionnaires for the maintenance and curation of data libraries that support the clinical research process. They also speculate that such standards could allow intelligent aggregation and analysis of multiple question formats that attempt to measure the same construct in different settings. Although SNOMED CT does not claim to represent clinical questions per se, it may be flexible and comprehensive enough to accommodate this unmet need in clinical research, and seems to the leading candidate for this.

SNOMED CT for Clinical Research Data

Arguably, SNOMED CT is well-suited to clinical research data insofar as it offers broad coverage and is clinically rich, which is needed for use by multiple disciplines.[12,17–28] It also has the potential to represent complex clinical concepts, including time, subject, and negation, within its terminology model. As noted, SNOMED CT already is the recommended data standard in three CHI-defined areas (procedures, anatomy, problem lists and diagnoses)[5] and has been identified as the standard terminology for these same constructs in the Rare Disease Clinical Research Network (RDCRN; the context of this study—See Methods Section). Moreover, it is experiencing a period of renewed growth with an increase in access generated by the National Library of Medicine public license agreement in 2000.[10,29] While SNOMED CT is often considered the most comprehensive vocabulary,[20,27,28,30–33] widespread adoption has not been achieved in either clinical medicine or research, and there has been little exploration into consistency and reliability of SNOMED CT coding across persons and institutions, especially since the expansion of the SNOMED CT terminology model in 1999.[17,34,35]

Important to implementation in research context, SNOMED CT has a robust conceptual model (called the 'Clinical Context Model') that allows for post-coordination (i.e., the creation of new concept expressions using the logical combinations of other concepts). A recent study showed that clinical research data items tended to require post-coordination,[12] since post-coordination by coders is required to allow for a high level of expressiveness, sufficient granularity, and to facilitate the particular concept representation needs for specific contexts. The use of post-coordination is relatively straightforward in many areas where the SNOMED CT (conceptual) terminology model is complete and intuitive. In some important cases, however, such as for the use of context-dependent concept qualifiers including negation and subject of observation, the use of post-coordination is novel and complex.[12]

Using a terminology model that allows for the post-coordination of new or complex concepts normalizes a terminology and eases terminology maintenance. However, in practice, there is an inherent tension between terminology management and navigation. That is, the needs for overall efficiency of the terminology conflict with users' desire for ease of coding, which can be enhanced by offering "pre-coordinated" terms and the flexibility for users to quickly create needed or missing concepts.[36–40] A recent review by Rosenbloom et al.[41] summarizes the problems with post-coordination succinctly as: a) the need for mechanisms or syntax to restrict post-coordination to meaningful concepts; b) the creation of duplicate concepts (or "undetected synonymy"); and c) the potential for inefficiency in creating concept expressions. These three consequences imply a need for guidance and structural features to ensure "correct" use of post-coordination. The ability to create duplicate concepts in terminologies supporting post-coordination has been noted.[42–45] Still, the increased likelihood for duplicate concepts directly implies variation in coding across coders, and has implications for data standardization.

It is likely that post-coordination of concepts is an area where greater inconsistency might occur. If clinical research data utilizes post-coordination more than health care delivery data, then clinical research is perhaps more vulnerable to variations in implementation of the data standard. Formal evaluation of SNOMED CT inter-rater variation, or agree-

ment, is therefore necessary to provide metrics on the usability of SNOMED CT as a standard in clinical research, the feasibility of its implementation, and the quality and integrity of SNOMED CT-encoded research data.

Coding Consistency

Coding consistency is an area that has not yet received much attention in the medical informatics literature. The few studies that have been done examined the consistency of coding within an organization or among coders and clinicians. For instance, Chiang et al.[34] measured reliability of SNOMED CT coding among three physicians using two different terminology browsers. While the focus of this small study was the browsers' effects on coding, consistency among the three physicians was shown to be around 50% for "exact matches", but increased a bit when a manual review of reliability, or "semantic matching" was conducted. Other studies have examined coding in the context of data quality, such as concordance between information managers and physicians in coding patient data,[46] or to check the quality of ICD-9-CM coding across a large healthcare costs and utilization project.[47]

Other studies have touched on consistency in the course of examining the coverage or performance of certain terminologies. For instance, one study from Great Britain conducted an in-depth trial of two terminologies (Clinical Terms Version 3 and Read Codes 5 byte set) to identify which had the best coverage for patient records in general practices, and which seemed to support the greatest consistency among practitioners.[48] Clinical Terms outperformed Read Codes, yet the relevance of these results to our study is limited. That is, the researchers examined terminologies to be used in a primary care electronic healthcare record, which certainly would require a broad and clinically rich, yet usable, terminology. However, the context of interest here is clinical research data which requires a finer level of granularity and oftentimes more complex post-coordination of terms to represent less than common concepts, as discussed earlier and in the literature.[47]

Analogous studies have been conducted to examine indexing of the medical literature. While not directly equivalent to coding of clinical research data, indexing of the literature is an important part of the health information infrastructure, particularly in this era of evidence-based medicine, and inter-indexer consistency studies highlight inherent difficulties in the utilization of controlled vocabularies in expressing knowledge of various sorts. Funk and Reid's[49] now classic study highlighted the importance of highly trained indexers when using the particularly well-designed MeSH vocabulary to index articles for MEDLINE. The issues raised in their study reveal some interesting foresight into the critical challenges that information retrieval systems would have to face. This was reflected in a subsequent study comparing indexing in CINAHL and MEDLINE during the 1980s,[50] as well. Also, inter-rater agreement became a relevant issue during NLM/AHCPR's Large Scale Vocabulary test.[51]

To the best of our knowledge, no studies have been conducted that examine the inter-coder agreement or consistency among experts working for professional coding companies. Additionally, we are not aware of any studies examining the coding consistency across trained individuals applying standardized terminologies to represent clinical research data in pure research settings.

Methods

The Rare Disease Clinical Research Network (RDCRN)

The data for this study were generated from physical examination and clinical assessment CRFs from 15 studies in the NIH-funded* RDCRN.[52] The RDCRN consists of ten clinical research consortia, each focused on several related rare diseases. The ten RDCRN consortia focus on research activities in the areas of: urea cycle disorders, neurological channelopathies, bone marrow failure diseases, cholestatic liver diseases, vasculitis, genetic steroid disorders, rare thrombotic diseases, rare lung diseases, genetic diseases of mucociliary clearance, and Angelman, Rett, and Prader-Willi syndromes. Collectively, these ten consortia research over 50 rare diseases across 46 research sites. One goal is to accelerate the development of diagnostics and treatments across a variety of rare diseases by encouraging cooperative partnerships and data sharing among the investigators at these centers.

The RDCRN is committed to the use of data standards, and is storing all data related to clinical findings, procedures, and anatomy using SNOMED CT, as recommended by CHI. A centralized Data and Technology Coordinating Center (DTCC; located at the University of South Florida, College of Medicine) promotes standards and tools for efficient study implementation. The DTCC's charge to implement SNOMED CT across the variety of study designs, settings, and medical specialties represented in RDCRN studies motivated this research.

Study Goal

We sought to compare the consistency or degree of variation among experts from professional coding services when coding concepts represented in data items on clinical research CRFs using SNOMED CT. Agreement across the three separate companies, each using the same terminology and with the same instructions, was measured, as well as other relevant information to inform a better understanding of coding consistency. Our interest was in how experts in SNOMED CT might vary or concur on the coding of the types of items in this special clinical research context.

As discussed earlier, CRFs provide an excellent source for clinical research concept harvesting since they represent the actual data collected in the clinical research context in which they are collected. We chose a random sample of questions culled from CRFs from 15 observational and interventional studies currently being conducted by consortia participating in the RCDRN. CRFs with content covered by other CHI data standards, such as Logical Observation Identifiers Names and Codes (LOINC) for lab test names, and RxNorm

*The network is supported by several NIH components, including the Office of Rare Diseases (ORD), National Center for Research Resources (NCRR), National Institute of Neurological Disorders and Stroke (NINDS), National Institute of Child Health and Human Development (NICHD), National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

for clinical drugs names, were not considered for this study. To include a maximum number of current clinical findings and observations (as opposed to medical history and patient self-report items, which contain varying and complex temporal aspects), question selection for this study was restricted to physical exam forms or their equivalents (e.g., clinical assessment forms).

Once the 15 active studies and relevant CRFs were selected, an SQL query was executed to retrieve all qualifying data items (questions and their corresponding answers) from a database of CRF data items. This resulted in an initial set of over 1800 question/answer sets. From this population, a random sample of approximately 20% ($n = 377$) was derived using the sampling function in SPSS v.14. This is a recommended sample percentage for studies similar to this one, wherein inter-coder reliability has been measured.[53] Some duplicates and difficult question constructions were removed at the discretion of the authors (JA,RR), resulting in a final data set of 319 question/answer pairs. The assumption was that, collectively, the sampled data items represented particular CRF items from a variety of clinical research domains.

Data Collection

Upon approval by the University of South Florida, College of Medicine IRB, a cover letter was sent to each of the participating companies explaining the study and providing instructions. Each company was provided with a spreadsheet containing the question set described above. The instructions included an explanation of the different kinds of questions they were to code, including unique identifiers for each question, the disease name under study, and the name of the form that the terms came from in order to provide further context to assist in coding. The companies were directed to capture the complete *meaning* of each data item (question/answer pair), but not to model the question structure or format per se. The companies were asked to add the SNOMED CT code(s) in specially marked columns, and told to use either pre-coordinated expressions or to post-coordinated concepts, whichever they deemed appropriate for capturing the fullest meaning of the question. Furthermore, they were made aware that, because these terms come from research projects in rare diseases, some of these terms might not exist in SNOMED CT and so post-coordination might be difficult or not possible. For such cases, they were asked to provide comments in a separate column on the spreadsheet. Participants were informed that their company's identity would remain confidential, and only aggregated results would be published. Moreover, we informed participants that we would not use a "gold standard," but were only interested in exploring the variations in coding among three professional coding companies with expertise in SNOMED CT.

Recognizing there may be ambiguity about how to model concept expressions to fit into either *Clinical Findings* or *Observable Entities* axes, we added a column for "Preferred hierarchy," and indicated the SNOMED CT hierarchy we would like these coded into, and asked companies to model concepts into the indicated preferred hierarchy whenever possible in order to reduce this type of variation. Moreover, for post-coordinated expressions, we requested the use of the "SNOMED International Post-coordination Syntax," recommended by the SNOMED Concept Modeling Working Group, and found in the SNOMED International document, "Abstract Logical Models and Representation Forms".[54]

In addition to SNOMED CT codes, other information was requested that we felt would provide a richer comparison of consistency or variation in coding. First, we requested each company to report their certainty of the coding selections. The companies were given three choices: *certain, somewhat certain,* or *uncertain*. Also, we requested that coders record the precision of each code using the following choices: *exact match, broader than, narrower than, or related to* (in some other relationship than hierarchical), given the presumed meaning of the original term that we provided.

Given the data set and the instructions for coding, we expected that there would be at least moderate levels of agreement. We also expected that areas showing greater variation would occur when either more complex post-coordination was required, or if multiple choices of descriptors were selected.

Data Analysis

Once each company had returned the coded set of questions, including the additional information described above, the data were processed for further analysis. Frequency data were calculated to provide an overall picture of the data, including overall percentages of agreement among coders and similar descriptive information. Further analyses conducted are described below.

Measuring Agreement

Inter-rater agreement, or reliability, among human observers or judges is a critical component in much social science research. There are a number of techniques and statistical tools used in various contexts to provide researchers with some confidence that certain measures, coding or indexing procedures, and other research tools are useful, and that results are likely to be trustworthy or reproducible. For instance, content analysis usually explores written or otherwise recorded information artifacts by use of coders (or judges, raters, or other synonymous terms) who must interpret such communication using some predetermined guidelines (e.g., from how many times certain concepts are mentioned in a text, to interpreting meaning using some coding procedure, or other related methods). Clearly, for such studies to be useful requires some evidence that the coding is being carried out with reasonable consistency, or for it to be otherwise deemed reliable. This would not be dissimilar to coding of clinical information, be it a clinical narrative, or concepts from a clinical trial questionnaire. Coding such concepts using a standard terminology means the rules for that terminology act as the guidelines, as well as the context for the coding (e.g., for a hospital's EHR, or clinical trial data collection, such as is the case for the RDCRN). In this study, the coding generated by three experts from professional coding companies are the variables of interest, and techniques similar to those used in content analysis when determining coding agreement are utilized. Specifically, we wished to explore the agreement among the experts in their use of SNOMED CT for coding clinical research concepts from clinical trial questionnaires.

Simply figuring out the percentage of times that two or more coders agree provides only limited information. Mostly, this

Table 1 ■ Example of Semantic Coding of Data

Concepts from CRF Questions	Company 1	Company 2	Company 3
Disease: Rett Syndrome	20573003	65740005	20573003
Form Name: Clinical Assessment Form	Ineffective breathing	Increased forced expiratory	Ineffective breathing
Question/Answer: *Breathing pattern-Forced air/saliva expulsion/Yes*	pattern (finding)	volume (finding)	pattern (finding)
Agreement coding by researchers (nominal)	1	2	1

is due to the fact that simple agreement percentages do not take into account chance occurrences. A number of statistical methods are available that seek to offer more robust measures of agreement. These usually are dependent on the nature of the data being analyzed. For instance, Cohen's kappa is a commonly applied statistic for measuring inter-rater agreement or reliability, and offers a stronger measure of agreement since it takes into account agreement that may have occurred by chance. However, this statistic measures the differences between two coders who make some single selection from a small set of choices or categories.[55] The nature of this study is such that coders can make more than one choice from many choices, and there is no gold standard to compare against. Therefore, a more flexible and, arguably, more robust method was chosen. Specifically, we employed Krippendorff's alpha.[56] The need and criteria for a standard reliability measure are outlined by Hayes and Krippendorff,[57] who designed a measure (Krippendorff's alpha) which addresses many of the shortcomings of other statistics used in the social sciences to measure agreement. The strength of this statistic (which makes it appropriate for use in analyzing the data for this study) is that it is applicable in various contexts and, importantly, it allows for more than two raters using nominal data. An SPSS macro was created to calculate this statistic, which is available freely at: http://www.comm.ohio-state.edu/ahayes/SPSS%20programs/kalpha.htm.

Measuring Semantic Agreement

The final data sets returned by each company showed variation in how each approached the coding task. While this is understandable given that each may use tools and certain approaches that differ slightly, this made a simple comparison of the SNOMED CT concepts more difficult. Thus, we initially examined only whether the core concept IDs matched. That is, if the same core concept ID was used by all three experts, despite the fact that some other SNOMED IDs might be present (as qualifiers), these were considered to be in agreement. Table 1 provides an example of this:

In the above example, experts from Company 1 and Company 3 used the same SNOMED CT core concept ID to express the question. The coder for Company 2, however, chose to express this using a different SNOMED CT concept. You will notice that the terms are likely synonymous, or at least quasi-synonymous; however, we chose not to make such judgments and to recognize these as being a different term. More about this will follow in the Discussion section. Also note that agreement data were recorded as nominal. The 1, 2, or 3 entered are not quantitative at all, but simply reflect which companies shared the same concept choices.

Comparing Syntax

The flexible nature of SNOMED CT leads to potential syntactic variation that might or might not include semantic variation. For many concepts, precoodinated concepts are present in SNOMED CT and the concept model also allows for the construction of equivalent meaning via post-coordinated expressions. We did not provide guidelines on our preference for the format of the returned coded data, but we did want to look at the variety and proportion of strategies used by the different company experts. In order to better identify the ways in which experts were constructing concepts in SNOMED CT at their discretion, authors (JA,RR) looked at the coded data returned by each expert and classified the type of SNOMED CT construction was used -specifically, whether each code was a: 1) single concept, 2) post-coordinated concept with clinical qualifiers, or, 3) post-coordinated with both clinical and non-clinical qualifiers. These categories were determined by the authors.

We also were interested in examining agreement related to the syntax of the experts' coding. In particular, we examined whether a single concept (including pre-coordinated concepts) was utilized, whether post-coordination of concepts using clinical qualifiers was done, and whether post-coordination with clinical *and* non-clinical (e.g., time, date, etc.) qualifiers was done. Agreement here also was coded using nominal data (1 = single concept, 2 = post-coordinated concept with clinical qualifiers, or, 3 = post-coordinated with both clinical and non-clinical qualifiers). The data were analyzed in the same manner by examining basic frequencies and utilizing Krippendorff's alpha.

Level of Certainty

The self-expressed certainty reported by each company was also examined. Essentially, this helped us in identifying particularly tricky terms, and generally how often experts agreed that they were either *certain, somewhat certain,* or *uncertain* of how they coded these concepts. Moreover, an overall comparison of the average level of certainty among the companies was analyzed since these data were ordinal.

Precision

Requesting coders to report what they felt the precision of their coding was enabled other comparison analyses. The data were recorded as nominal, and allowed the researchers an insight into the level of granularity the experts felt they were able to achieve when coding. Frequencies were calculated, and Krippendorff's alpha was also conducted.

Results

The basic coding agreement percentages among the three experts from professional coding services were as follows: All three agreed on the same core concept, 33% of the time; two of the three coders selected the same core concept ID,

Table 2 ■ Krippendorf's Alpha (α) Reliability Estimate

	All Three Companies	Companies 1 and 2	Companies 1 and 3	Companies 2 and 3
Semantic Agreement	α −.0625	α −.2844	α −.2535	α .1406
(*Krippendorf's alpha (α); 95% CI: LL/UL)	−.1376 to .0068	−.4564 to −.1125	−.3859 to −.1212	.0377 to .2435
Certainty Agreement	α −.0215	α −.0534	α −.0887	α −.0534
	−.1014 to .0564	−.0781 to .1820	−.2241 to .0513	−.0788 to .1827
Syntactic Agreement	α −.1603	α −.5079	α −.4205	α .1074
	−.2132 to −.1091	−.5667 to −.4990	−.5252 to −.3158	−.0042 to .2255
Precision Agreement	α .1276	α −.0910	α .2787	α .1086
	.0648 to .1926	−.2007 to −.0188	.1806 to .3710	−.0222 to .2395

**Upper and Lower confidence intervals. 95% confidence that reliability should fall between these two values if the entire population were coded by these companies.
Agreement based on semantic, certainty, syntactic, and precision among experts from three coding companies.
*N = 319 with uncoded elements treated as missing; bootstraps = 8000 (see Hayes and Krippendorff for fuller explanation on how this helps in statistically determining sample distribution for such data).

44% of the time; and, no agreement among all three, 23% of the time. These frequencies are somewhat consistent with an analogous study by Chiang et al.,[34] cited earlier, who examined reliability of SNOMED CT coding by three physicians using two different browsers. They found "exact coding" among these physicians to be 44% when using one browser, and 53% using a second browser. It is likely, however, in the case where physicians each were using the same browser for coding resulted in an increased percentage of agreement (which is still, arguably, not ideal). In the case of our study, the companies likely had their own, independent tools and methods for searching for SNOMED CT codes. The examination of these proprietary tools, however, was outside the scope of this study and not considered in order to protect the identity of the study participants. While interesting, we felt that assessing and exploring any variation of the coding tools used by the coding companies would contribute little to our end results, since there is no gold standard in this study.

Krippendorff's alpha was then figured using the SPSS macro designed by Hayes and Krippendorff.[57] Table 2 shows the results of this analysis, as well as those for the measurement of agreement between pairs of coders (the table also includes results for the other analyses described below). For the latter, we considered the possibility that any pairing of two companies could show a stronger level of agreement, and so explored the different combinations. (Note: Each of these combinations were run on their respective sets of data, with appropriate modifications made to the macro—i.e., variable names).

The results of this analysis are not inconsistent with what one would expect given the raw percentages shown earlier. That is, if there was full agreement only 33% of the time, then statistically, given the probably of chance agreement, there is a low chance for a higher estimate of reliability. The negative numbers in all but one of the above cells (that is, in the Semantic Agreement row) indicates a lack of agreement among these experts. The one that is not a negative number still suggests lack of agreement; one would hope for a number much closer to 1.0. The alphas are consistent with the lower-level (LL) and upper-level (UL) confidence interval numbers displayed. These numbers reflect the range that, if the entire population of these data were coded, the measure of reliability would be between them.[56] Moreover, more detailed statistics generated by running the Krippendorff's alpha macro show that all of these data, for all analyses, have an almost certain chance of *not* achieving the minimum alpha needed to report reliable agreement. Krippendorff represents this as the probability (q) that the α_{min} will not be achieved. A table is generated as part of the overall results, and shows a relaxing of the reliability standard (α_{min}) in ten per cent intervals, matched with the respective q probability (that the minimum would not be met). This information is not presented here since every case, for every degree of α_{min}, q equaled 1.0. Lastly, Krippendorff also notes that nominal data tends to result in a higher alpha, further supporting our conclusions that there is a lack of agreement.

Frequency data for agreement among the coding experts regarding their self-reported level of certainty were as follows: no agreement, 20%; two out of the three companies reported same certainty level, 55%; and, all three reported same certainty level, 25%. This indicates little agreement in certainty ratings, showing that the companies independently had confidence in coding on different items. The results for the Krippendorff's alpha analysis are found in Table 2, and again show lack of agreement for each combination of companies regarding how often they felt the same level of certainty for each term.

Frequencies and Krippendorff's alpha also were calculated to identify agreement on syntactical approaches to coding, as well as for precision. Frequencies for these are shown in Table 3, which reveals the variation in the approach each expert took in the construction of terms. Reliability estimates that support this lack of agreement are shown in Table 2.

Discussion

The central issue in this study was whether we could identify measurable consistency among experts employed by professional coding companies that utilized SNOMED CT to code clinical research concepts derived from CRFs. A key finding, therefore, was that we were unable to discern such agreement among these companies on any of the areas that we examined. We hope these results encourage discussion on the need for more directed efforts to better enable users of SNOMED CT, in various health contexts, to make efficacious use of this important terminological standard.

Table 3 ■ Comparison of Frequencies for Syntactic Approaches and Precision by Coding Experts

	Company 1	Company 2	Company 3
Syntactic			
Single or Pre-Coordinated	1%	70%	67%
Post-Coordinated with Clinical Qualifiers	95%	12%	33%
Post-Coordinated with both Clinical and Non-clinical Qualifiers	2%	11%	0%
Missing or Uncoded	2%	7%	0%
Precision			
Exact Match	41%	85%	65%
Broader Than	37%	9%	24%
Narrower Than	4%	5%	3%
Related	18%	1%	8%

These should be relevant to both users of the terminology as well as system developers.

The lack of agreement among the coding experts regarding semantic (how each actually coded concepts) seems critical, but also must be examined in light of the limitations of this study. First, as noted, the companies approached this project in ways that differed enough to make a simple matching of concept IDs unfeasible. This meant we were required to find another, more simplified approach to measuring semantic agreement. Yet, even when only examining the core concept ID from each of the coders, there still was a notable lack of consistency. Moreover, comparing the companies in pairs still failed to yield any level of agreement.

The aggregated data set comparing all the coding reveals a more qualitative assessment of these results, including some potential reasons for coding inconsistencies. First, looking at cases wherein all the experts chose the same core SNOMED CT ID, these often occurred when there was a specific condition or disorder to be coded. For instance, each chose *248650006|Cardiac murmur, intensity grade IV/VI (finding)* for an item on a Bone Marrow Failures physical examination form. Similarly, each chose *36760000|Hepatosplenomegaly (disorder)* for the question/answer pair from a Rett Syndrome clinical assessment form, "Abdomen-Hepatosplenomegaly/no". In this case, however, one coder chose only the concept ID and expression itself (as just shown), yet two of the three added more contextual qualifiers, as follows:

1. 243796009|situation with explicit context|: 246090004|associated finding|=36760000|Hepatosplenomegaly (disorder)|, 408729009|finding context|=410516002|known absent|, 408731000|temporal context|=410512000|current or specified|, 408732007|subject relationship context|=410604004|subject of record
2. 116680003|Is a (attribute)|=36760000|Hepatosplenomegaly (disorder), 408729009|Finding context (attribute)|=410516002|Known absent (qualifier value)

Since we were interested only in whether or not the experts captured the same core concepts the same way, we recorded this as each being in agreement, despite clear syntactical differences illustrated above. We recorded lack of agreement when different core concept IDs were used, even when the multiple core concept IDs were likely to be synonymous or quasi-synonymous. For instance, the concept from a question on one physical examination form was *Eyes Hemorrhage;* the three coders coded the core concept as follows: a) *246680008|Bleeding eye (finding); b) 246681007|Blood in eye (finding)*; and, c) *9347800|Intraocular hemorrhage (disorder)*. Arguably, each of these SNOMED CT codes expresses the core concept of the question. If so, cases such as this one mean that data management or information retrieval systems must be able to facilitate mapping among synonymous codes or risk a loss of data integrity, meaning that the robustness of information retrieval could be compromised. We did not explore the SNOMED CT description logic underlying these codes, but a goal of SNOMED CT is to facilitate the determination of equivalence (i.e., synonymy) of multiple post-coordinated expressions. Seeing the kind of variability in construction of post-coordinated expressions that we did might indicate that SNOMED CT has a significant challenge ahead. Description logics are the key,[45,58] but we suspect that they are incomplete in the current SNOMED CT version. Post-coding synonymy mapping would need to be organization or context specific to accommodate the level of synonymy (and information management requirements) appropriate for a given context. That is, a general practice office might only require a strict interpretation of synonymy, while a system dealing with clinical trials data aggregated from various study sites might require a looser definition to facilitate data-mining and the like.

Other inconsistencies were identified due to different post-coordination of concepts. Here, we are not referring to differences in the clinical or contextual qualifier choices where the core concept IDs are the same, but cases where coding of a concept was done through a coordination of two or more concepts. For instance, the core concept of an item on a Vasculitis physical examination form was "Vascular exams: Carotid Right/Tender." This was coded by the three companies as follows:

1. 309655006|On examination-artery (finding)|: 69105007|Carotid artery structure (body structure)|: 24028007|Right (qualifier)
2. 401050002|Carotid artery finding (finding)|: 363698007|finding site|= 69105007|Carotid artery structure (body structure)|:272741003|laterality|=24028007|Right (qualifier value|)
3. 116680003|Is a (attribute)|=301390006|Tenderness of cardiovascular structure (finding),363698007|Finding site (attribute)|=38917008|Structure of right internal carotid artery (body structure)

Each of these varies in ways that might preclude efficient retrieval, in a manner similar to that discussed previously in regards to issues of synonymy. And, as with synonymy, systems will be required that can deal with multiple, post-coordinated expressions representing similar concepts.

Our findings from examining the self-reported levels of certainty should also raise some questions. More than half of the time, all three companies agreed that they were only "somewhat certain" about the concept choice. It is possible that the complexity of SNOMED CT contributed to this, but clearly there are other factors not directly related to the vocabulary itself. For instance, a possibly confounding factor is that the data for this study (questions from RDCRN CRFs)

may be not be as clear (both syntactically and semantically) as other types of data these companies might traditionally code. If true, then more work may be needed to investigate and better understand the attributes of clinical research information and how these might be addressed in both terminologies and systems and by clinical research enterprises. The assumption is that the data items sampled for this study are representative of a broad range of CRF physical exam data, but this assumption needs to be tested. Although the data items come from rare disease research, we assume that the questions, and therefore the needed clinical research concepts, are typical within various clinical research domains, and that it is the clinical profile – or answer combinations - that are rare. We hope the clinical research community is inspired to characterize the nature and spectrum of CRF data items, as well as issues of SNOMED CT coverage and coding agreement, outside of the rare diseases that were targeted here.

The other areas studied here that revealed a lack of consistency came from our investigation of syntax and precision. For the former, we found that companies varied with one another, but were consistent in how they approached construction of concepts. For instance, one company had very many cases where a single concept code was chosen, another had no instances of this, and the other had a balance of single, pre-coordinated, and shorter post-coordinated concepts. The data shown for precision revealed that while there was not agreement among the three companies on this issue, each company stated "exact match" for the vast majority of coded items. Thus, despite the variation in selected SNOMED CT codes, each coder frequently felt that they had good matches and were certain of those matches. The high level of confidence (i.e., reported certainty as "very certain") in light of the high variation in actual coding is alarming to see. The nature of the questions and concepts from the data sample might explain some of this. SNOMED CT is a terminology to represent clinical concepts, and not necessarily CRF questions. Aside from determining the capacity of SNOMED CT to represent desired concepts used in clinical research (i.e., characterizing the breadth and depth of needed concepts), future studies need to assess the nature of CRF data collection and question modeling.

The results of this study, coupled with our experiences conducting it, do not change our opinion that SNOMED CT is a viable and appropriate data standard for clinical research, albeit some modification and clarification is needed. SNOMED CT is the front-running standard for clinical practice and healthcare delivery, so it is sensible, particularly with the goals of translational research, that it emerge concurrently for research. As noted earlier, we found that it showed good coverage for a data set of concepts from a vasculitis research consortium within the RCDRN. And, coverage per se did not appear to be a problem here; rather, the coding experts seemed to differ in many cases regarding structuring concepts. Thus, concepts that were arguably synonymous were recorded, but post-coordination and other syntactical differences emerged. We encourage future study to determine how much synonymy and quasi-synonymy can be formally detected (e.g., computationally) by the defining relationships within the SNOMED CT knowledge base. Anecdotally, we saw examples of variation across coders that were quasi-synonymous but the existing SNOMED CT description logic was not sufficient to determine this, especially in instances where the experts chose different axes (e.g., observables versus findings). A separate study would be needed to study the extent of this.

We also believe that the variability we saw underscores the need for consensus and communication regarding this standard's use. Because there are unique features of clinical research data, we think that guidance should be developed cooperatively between the SNOMED CT development community and the clinical research community. Additionally, because the context of clinical research data collection is typically items on a (case report) data collection form, supplemental standards that more easily represent structural, time, and contextual aspects of the clinical research questions might be used in conjunction with SNOMED CT to reduce coding variability. This approach has been recommended for structured assessment items, using SNOMED CT for the clinical content and clinical LOINC for the structural aspects.[14,59,60]

Conclusion

This study illustrates the issues that emerge with attempts to use standard terminologies, such as SNOMED CT, in several studies of rare disease. A grand challenge in medical informatics has been the development and implementation of standardized terminologies meant to represent the vast number of concepts in the information-intensive fields within healthcare. As SNOMED CT continues to emerge as a key terminology standard addressing this challenge, one which has the potential to be adopted in a variety of health care contexts, attention also must be focused on the consistent application of the terminology, as well as its structure and completeness. A clinically rich, flexible terminology is needed, albeit one that will enable suitably consistent use.

The fact that there was inconsistency among experts from three professional coding services when utilizing SNOMED CT for clinical research concepts may suggest that this context has some special needs that will require further, more focused attention; and it may be that the concepts in clinical research are not as similar to health care concepts, generally, as one might assume. Our findings help give an appreciation for the complexities involved in applying data standards into the clinical research domain, and should inspire future implementation and evaluation activities in this area. Since there is no widespread use of SNOMED CT in clinical research, there is little known about whether this data standard is adequate to represent clinical research concepts, and whether SNOMED CT can be applied consistently. Determining the extent and consistency to which concepts embedded in clinical research can be represented by SNOMED CT will help illuminate unmet needs and inherent complexities that may impede semantic interoperability and effective clinical research data management.

References ■

1. Yasnoff WA, Overhage JM, Humphreys BL, LaVenture M. A national agenda for public health informatics. J Am Med Inform Assoc 2001;8:535–45.
2. Altman RB. The interactions between clinical informatics and bioinformatics: a case study. J Am Med Inform Assoc 2000;7: 439–43.

3. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001;8:527–34.
4. American Medical Informatics Association and American Health Information Management Association Terminology and Classification Policy Task Force. Healthcare terminologies and classifications: an action agenda for the United States. Available at: http://www.library.ahima.org/xpedio/groups/public/documents/ahima/bok1_032401.html/. Accessed: June 11, 2007.
5. CHI. CHI executive summaries. Consolidated Health Informatics; 2004.
6. NCI. caBIG: cancer biomedical informatics grid: data standards. National Cancer Institute; 2006.
7. CDISC. Clinical data interchange standards consortium. CDISC, 2005.
8. HL7. Health level seven. Health Level Seven, Inc.; 2005.
9. Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthc Inform 2004;21:54, 56.
10. CAP. News release: HHS Secretary Tommy G. Thompson announces access to SNOMED CT through National Library of Medicine. In: S. International, ed. Northfield, IL: SNOMED International; 2004.
11. FDA. FDA news: FDA advances federal e-health effort, In: U.S.D.o.H.a.H. Services, ed. U.S. Food and Drug Administration; 2006.
12. Richesson RL, Andrews JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. J Am Med Inform Assoc 2006;13:536–46.
13. Aday LA. Designing and conducting health surveys, 2nd ed. San Francisco, CA: Jossey-Bass; 1996,560.
14. White TM, Hauan MJ. Extending the LOINC conceptual schema to support standardized assessment instruments. J Am Med Inform Assoc 2002;9:586–99.
15. Pocock SJ. Clinical trials: a practical approach. New York: Wiley and Sons; 1983,247.
16. Brandt CA, Cohen DB, Shifman MA, Miller PL, Nadkarni PM, Frawley SJ. Approaches and informatics tools to assist in the integration of similar clinical research questionnaires. Meth Inform Med 2004;43:156–62.
17. Rothschild AS, Lehmann HP, Hripcsak G. Inter-rater agreement in physician-coded problem list. Proc Am Med Inform Assoc 2005:644–8.
18. Giannangelo K, Berkowitz L. SNOMED CT helps drive EHR success. J Ahima 2005;76:66–7.
19. Burkhart L, Konicek R, Moorhead S, Androwich I. Mapping parish nurse documentation into the nursing interventions classification: a research method. Comput Inform Nurs 2005;23:220–9.
20. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699–703.
21. Van Berkum MM. SNOMED CT encoded cancer protocols. AMIA Annu Symp Proc 2003:1039.
22. Shamoun D, Livesay L. Organizing the animal hierarchy into a Linnean Taxonomy in SNOMED CT. AMIA Annu Symp Proc 2003:1005.
23. Cantor MN, Lussier YA. Putting data integration into practice: using biomedical terminologies to add structure to existing data sources. AMIA Annu Symp Proc 2003:125–9.
24. Dolin RH, Spackman KA, Markwell D. Selective retrieval of pre- and post-coordinated SNOMED concepts. AMIA Annu Symp Proc 2002:210–4.
25. Bakken S, Warren JJ, Lundberg C, Casey A, Correia C, Konicek D, et al. An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED clinical terms. Int J Med Inform 2002;68:71–7.
26. Burgun A, Bodenreider O, Mougin F. Classifying diseases with respect to anatomy: a study in SNOMED CT. AMIA Annu Symp Proc 2005:91–5.
27. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clinic Proc 2006;81:741–8.
28. Warren JJ, Wilson RP. Representing cardiovascular concepts in an electronic health record using SNOMED CT®. AMIA Annu Symp Proc 2006:1135.
29. NIH. SNOMED clinical terms® to be added to UMLS® Metathesaurus®, NLM, ed. 2003.
30. McClay JC, Campbell J. Improved coding of the primary reason for visit to the emergency department using SNOMED. AMIA Annu Symp Proc 2002:499–503.
31. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity: CPRI work group on codes and structures. J Am Med Inform Assoc 1997;4:238–51.
32. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4:484–500.
33. Dykes PC, Currie LM, Cimino JJ. Adequacy of evolving national standardized terminologies for interdisciplinary coded concepts in an automated clinical pathway. J Biomed Inform, 2003;36: 313–25.
34. Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ, Starren JB. Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers. AMIA Annu Symp Proc. 2006:131–5.
35. Hasman A, de Bruijn LM, Arends JW. Evaluation of a method that supports pathology report coding. Methods Inf Med 2001; 40:293–7.
36. Cimino JJ. Terminology tools: state of the art and practical lessons. Methods Inf Med 2001;40:298–306.
37. Spackman K. Rates of change in a large clinical terminology: three years experience with SNOMED clinical terms. AMIA Annu Symp Proc 2005:714–8.
38. Tuttle MS, Cole WG, Sheretz DD, Nelson SJ. Navigating to knowledge. Methods Inf Med 1995;34:214–31.
39. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7:298–303.
40. Cimino J. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394–403.
41. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13:277–88.
42. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Cote RA. Toward representations for medical concepts. Med Dec Making 1991;11(Suppl):S102–8.
43. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc 2005;12:486–94.
44. McKnight LK, Elkin PL, Ogren PV, Chute CG. Barriers to the clinical implementation of compositionality. AMIA Annu Symp Proc 1999:320–4.
45. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med 1997;9:139–71.
46. Lorence D. Regional variation in medical classification agreement: benchmarking the coding gap. J Med Syst 2003;27:435–43.

47. Berthelsen CL. Evaluation of coding data quality of the HCUP national inpatient sample. Top Health Inf Manage 2000;21:10–23.
48. Brown PJ, Warmington V, Laurence M, Prevost AT. Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice. Br Med J 2003;326:1127.
49. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc 1983;71:176–83.
50. Sievert MC, Andrews MJ. Indexing consistency in information science abstracts. J Am Soc Inform Sci 1991;42:1–6.
51. Ortiz E, Eagon C, Lincoln MJ. Methods for evaluating inter-rater agreement during the NLM/AHCPR large scale vocabulary test. AMIA Annu Symp Proc 1997:883.
52. NIH Press Release. NIH establishes rare diseases clinical research network. NCRR, ed. 2003.
53. Nuendorf KA. The content analysis guidebook. Thousand Oaks, CA: Sage: 2002.
54. CAP, SNOMED Clinical Terms® Guide: abstract logical models and representational forms. January 2006 CMWG revision, version 5; 2006, 2006.
55. Stein C, Devore RB, Wojcik BE. Calculation of the kappa statistic for inter-rater reliability: the case where raters can select multiple responses from a large number of categories. In: SAS® Users Group International Conference. Cary, NC, USA: SAS Institute Inc; 2005.
56. Krippendorf K. Content analysis: an introduction to its methodology. Thousand Oaks, CA: Sage; 2004.
57. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Commun Meth Meas 2007; 1:77–89.
58. Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. AMIA Annu Symp Proc 1998;740–4.
59. CHI. Consolidated health informatics: Standards Adoption Recommendation. Functioning and Disability. Consolidated Health Informatics, 2006.
60. Bakken S, Cimino JJ, Haskell R, Kukafka R, Matsumoto C, Chan GK, Huff SM. Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. J Am Med Inform Assoc 2000;7(6):529–38.