Comparing Clinical Vocabularies Using Coding System Fidelity

J. Craig Klimczak, D.V.M., M.S.[1,2], Allen W. Hahn, D.V.M., Ph.D.[3],
Mary Ellen Sievert, Ph.D.[4], Greg Petroski, M.S.[2], and John Hewett, Ph.D.[2]
[1]Health Services Management, [2]Medical Informatics Group, [3]College of Veterinary Medicine,
[4]School of Library and Informational Science, University of Missouri-Columbia, MO 65211

Much effort has been directed toward the development of an ideal multipurpose controlled medical vocabulary for use in human and veterinary medicine. SNOMED International is one effort that has resulted in a larger and more complex nomenclature system. Although it was able to code more concepts, SNOMED International failed to statistically improve vocabulary fidelity when compared with the 30+ year old SNVDO vocabulary. We found that SNOMED has a lower intercoder consistency than SNVDO and that a greater number of codes were necessary to represent an individual concept. Our study shows a significant Coder-Vocabulary interaction which suggests that more emphasis should be placed on coding guidelines and coder training. Clinician data entry and coding may be necessary for maximum vocabulary fidelity.

INTRODUCTION

Recently the Systematized Nomenclature of Human and Veterinary Medicine International third edition (SNOMED) has been introduced to serve as a nomenclature for coding terms in human and veterinary medicine.[1] Along with the developer, the College of American Pathologists, the American Veterinary Medical Association participated in the introduction of SNOMED as a standard for coding terms in both human and veterinary medicine. Since veterinary medicine is contemplating switching from the Public Health Service's Standard Nomenclature of Veterinary Diseases and Operations (SNVDO) to SNOMED, we are interested in which system best meets the needs of the veterinary profession.

Many authors have attempted to identify the qualities of a ideal multipurpose controlled medical vocabulary. For example Cimino et al. defined seven properties that improve query sensitivity, specificity, and reliability.[2] These properties are: 1) domain completeness, 2) unambiguous terms, 3) non-redundancy, 4) synonyms, 5) multiple classification of terms, 6) consistency of views, and 7) explicit relationships. The Canon Group defined 13 properties in their model: 1) parsimonious, 2) non-redundancy, 3) domain completeness, 4) nonvaguesness, 5) nonambiguity, 6) synonymy, 7) polysemyns, 8) lexically decomposable, 9) semantic typing, 10) compositionally extensible, 11) support of multiple hierarchies, 12) support of non-hierarchical relationships, and 13) multilingual characteristics.[3] While these properties define and tell us much about the ideal structure of medical vocabulary systems, they fail to provide a good performance measure to compare vocabularies as they are used in medical record systems. Further, coding a case or document for retrieval with high sensitivity and specificity is much different than coding an actual diagnostic finding where the decoded information must match the original input concept phrase.

The American Heritage Dictionary defines "fidelity" as 1) faithfulness; loyalty; 2)conformity to truth; accuracy; 3) the degree to which an electronic system reproduces sound without distortion.[4] We believe that a multipurpose controlled medical vocabulary also should have fidelity. We define this property as the degree to which a coding system reproduces concept phrases without distortion. In other words, vocabulary fidelity means how close the output concept phrase matches the original input concept.

Several papers have compared coding systems based on vocabulary output; which we have defined as fidelity. Payne, et al., used a five point Likert scale to assess clinician satisfaction with coding problem list.[5,6] The following labels were placed on the 5-point Likert scale: "1" which indicated "extremely dissatisfied"; "2" which indicated "quite dissatisfied"; "3" which indicated "neutral, not satisfied but no objection"; "4" which indicated "quite satisfied"; and "5" which indicated "extremely satisfied". While satisfaction is not the same as degree of match, it does provide a scale for how well a vocabulary output matches its input.

METHODS

In this study, we randomly selected 50 small animal patient records from the University's Veterinary Medical Teaching Hospital (VMTH) medical

0195-4210/95/$5.00

records and recorded the problems found on the patient master problem list. The problem list terms were entered into a relational database (Paradox®*) for later manipulation. The terms were entered exactly as they appeared in the chart without any attempts to correct spelling or clear up meaning. The 50 charts yielded 148 problem list entries of which 132 were unique. The number of entries per chart ranged from 0 to 21 with a mean of 2.94 and a median of 2 problems per chart. Chart review revealed that the master problem lists were not consistently maintained and the expected dynamic changes that were supposed to be represented in the problem lists were not present. We searched for working problem lists as described by Weed; however, none were found in the charts reviewed.[7,8]

Two professional veterinary medical record coders from the VMTH were asked to code the list of Problem terms in both vocabularies. Coders were familiar with the SNVDO nomenclature and represented a combined coding experience of 20 person years; however, neither coder was familiar with the SNOMED vocabulary (it had just been released). Coders were given forms containing the 132 unique problem concepts. They were asked to code each concept twice, once with SNVDO and once with SNOMED. The coders worked independently of each other. Prior to coding the lists, coders were provided with a short introduction to SNOMED. This introduction included a demonstration of the use of an electronic browser as well as an overview of the printed version of the nomenclature.[9,10,11] The electronic browser and the printed version of the vocabularies were made available to facilitate the coding process.

Table 1. Coder Response by Vocabulary

VOCABULARY		Coder 1	Coder 2	Total
SNOMED	No. Coded	128/132	115/132	243/264
	Codes Used	191	138	329
	Range	0-6	0-3	
	Codes/Prob.	1.49	1.20	
SNVDO	No. Coded	108/132	110/132	218/264
	Codes Used	133	116	249
	Range	0-4	0-2	
	Codes/Prob.	1.23	1.05	

*Paradox is a registered trademark of Borland International, Inc., Scotts Valley, CA.

The codes provided by the coders were entered into the database. Individual codings, between coders using the same vocabulary, were compared for consistency and scored according to agreement. Consistency scores were recorded as "agree", "partially agree", and "disagree". In this study, "partially agree" was defined as a common code used by each coder for a particular term.

Codes were then decoded for evaluation by three veterinary clinicians chosen from the clinical practice staff of the VMTH. Each of the selected clinicians was certified in a recognized veterinary specialty -- veterinary internal medicine, veterinary ophthalmology, and veterinary surgery. Each clinician evaluator was given forms containing the original problem concept and the decoded SNOMED and SNVDO terms for each coder. Evaluators were asked to score each problem/vocabulary/coder combination as to "exact match", "broader partial match", "narrower partial match", and "no match".

Outcomes were realized for each combination of coder, evaluator, and vocabulary which resulted in a three factor study design with repeated measures over problem list term, vocabulary, and coder for fidelity assessment. Originally, plans were to combine coder responses to obtain coder consensus and a single coder outcome for each problem statement. However, in reviewing the coder inconsistencies, we found equally valid codings. We believed that it would be important to evaluate the coder impact on vocabulary fidelity.

STATISTICAL METHODS

The Statistical Analysis System (SAS®†) was used to perform the statistical analysis of the data stored in the database. For each of the randomly selected problem statements, two experiments with two sets of outcomes were realized. The first experiment evaluated inter-coder consistency. The second experiment evaluated the factors that affect vocabulary fidelity.

For assessment of inter-coder consistency, the designations "partially agree" and "disagree" were combined to form a dichotomous outcome of "agree" and "disagree". McNemar's test for paired proportions was used to test if the proportion of

† SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.

agreeing codes was different for the two vocabularies.[12]

For assessment of fidelity, the two varieties of partial matching, "broader partial match" and "narrower partial match" were combined to form a three point scale of "no match", partial match", and "exact match". Since the outcomes for fidelity were realized for each combination of vocabulary/coder/evaluator, an appropriate statistical model for fidelity assessment is a three factor analysis of variance (ANOVA) with repeated measures on each factor. The repeated measures model for categorical data developed by Koch, et al. as implemented in SAS Proc Catmod was used for the analyses.[13,14] This model provides an ANOVA type analysis for the marginal probabilities associated with a contingency table.

Table 2: Coder Consistency

	Percent No.	SNVDO		
		Disagree	Agree	Total
SNOMED	Disagree	16% (21)	32% (43)	48% (64)
	Agree	12% (16)	40% (52)	52% (68)
	Total	28% (37)	72% (95)	100% n=132

RESULTS

The consistency evaluation revealed that the two coders behaved differently when coding with each vocabulary (See Table 1). Coder 1 used between 0 and 6 SNOMED codes and between 0 and 4 SNVDO codes per problem list term, whereas, Coder 2 used between 0 and 3 SNOMED codes and between 0 and 2 SNVDO codes per problem list term. The coders varied on the number of problem list terms that they coded both within a vocabulary and between vocabularies. Coder 1 assigned a code for 128/132 problem list terms with SNOMED, while coder 2 assigned a code for 115/132 problem list terms with SNOMED. With SNVDO coder 1 assigned a code for 108/132 problem list terms, while coder 2 assigned a code for 110/132 problem list terms.

There was a significant difference between vocabularies with respect to the proportion of agreeing codes (p=0.0003). The consistency scores for "agree" and "disagree" between coders revealed that the coders agreed 72% of the time with SNVDO and only 52% of the time with the newer SNOMED. With 39% of the problem list terms the coders chose the same codes from both vocabularies and for 16% of the terms the coders chose different codes from both vocabularies. Table 2 illustrates coder agreement between the two vocabularies over the problem list term set.

Table 3. Summary of Evaluator Responses by Coder

Coder	Vocabulary	No Match	Partial Match	Exact Match
C1	SNOMED	10%	32%	58%
	SNVDO	24%	27%	49%
C2	SNOMED	19%	33%	48%
	SNVDO	22%	29%	49%

The ANOVA type analysis of the fidelity data revealed a statistically significant coder by vocabulary interaction (p=0.01). As with the traditional analysis of variance, main effects are not interpretable in the presence of significant interaction among factors, thus, a separate ANOVA was performed for each coder.

From the separate ANOVA's, we find for coder 1 the vocabulary effect was statistically significant (p=0.0004) and not significant for coder 2 (p=0.52). Table 3 summarizes the results of the fidelity study by pooling evaluator responses across coders. With coder 1, SNOMED's fidelity was judged "exact match" 58% compared with 49% for SNVDO (significant). Whereas, with coder 2, SNOMED's fidelity was judged "exact match" 48% compared with 49% for SNVDO (not significant). One coder was able to perform better with one vocabulary over the other. In the separate ANOVA's, the evaluator by vocabulary interaction was not significant for either coder, thus, the main effects are interpretable.

Table 4: Evaluator response to SNOMED Coding

Coder	Eval.	No Match	Partial Match	Exact Match
C1	E1	11% (14)	21% (28)	68% (90)
	E2	14% (19)	36% (48)	50% (65)
	E3	5% (6)	40% (53)	55% (73)
C1 tot	n=396	10% (39)	32% (129)	58% (228)
C2	E1	21% (28)	20% (26)	59% (78)
	E2	21% (28)	37% (49)	42% (55)
	E3	14% (19)	42% (55)	44% (58)
C2 tot	n=396	19% (75)	33% (130)	48% (191)
Total	n=792	14% (114)	33% (259)	53% (419)

The within-coder analyses revealed a significant evaluator effect for each coder (p=0.0001 for both coders). This suggests that evaluators were not

consistent with respect to each other in their evaluations. The overall results of the fidelity study are summarized in Tables 4 and 5. These tables show that the clinician-evaluator fidelity responses to the codings varied across the coder using the SNOMED and SNVDO vocabularies. Using SNOMED, Coder 1 was judged "exact match" 68% by E1, 49% by E2, and 55% by E3 and coder 2 was judged "exact match" 59% by E1, 42% by E2, and 44% by E3. Using SNVDO, Coder 1 was judged "exact match" 54% by E1, 45% by E2, and 48% by E3 and coder 2 was judged "exact match" 57% by E1, 48% by E2, and 41% by E3.

Table 5: Evaluator Response to SNVDO Coding

Coder	Eval.	No Match	Partial Match	Exact Match
C1	E1	26% (34)	20% (27)	54% (71)
	E2	26% (34)	29% (39)	45% (59)
	E3	20% (26)	32% (42)	48% (64)
C1 tot	n=396	24% (94)	27% (108)	49% (194)
C2	E1	26% (34)	17% (22)	57% (76)
	E2	23% (31)	29% (38)	48% (63)
	E3	17% (23)	42% (55)	41% (54)
C2 tot	n=396	22% (88)	29% (115)	49% (193)
Totals	n=792	23% (182)	28% (223)	49% (387)

DISCUSSION

There was statistically significant difference in coder consistency between the two vocabularies (72% for SNVDO versus. 52% for SNOMED). Some of the variation can be accounted for by the fact the coders were experienced with SNVDO but not experienced with SNOMED. However, some variation was due to the ability to represent concepts in multiple ways with SNOMED. When comparing the coding choices made by the coders we found that one coder was more verbose in coding style than the other. This coder took advantage of SNOMED's expressive power and used many modifiers and linkage terms to construct a problem term coding. The other coder tended to be more terse and code only the root concept of the problem term.

The variation in coding style proved to be the most significant interaction discovered by this study. Neither SNOMED nor SNVDO have a published set of coding guidelines in use. This leads to many problems since both systems are multi-axial and allow combinations of codes to represent a concept. The ability to combine codes greatly improves the expressive power of each system, but, this creates problems for database designers who must specify field sizes in clinical databases. Further, these multi-code descriptions have ordinal/positional significance with respect to what a modifier modifies. No standards exist at the present time to determine code ordering, for example, "does a modifier precede the concept or follow it?"

There was very little agreement among the clinicians about the fidelity of the codings both across vocabularies and coders. The clinicians were equally dissatisfied with both vocabulary systems having a pooled "exact match" score of 53% for SNOMED and 48% for SNVDO. SNOMED's higher score hinted that SNOMED may have a better fidelity than SNVDO; however, no final judgment can be made about vocabulary fidelity due to significant interactions. Much of this variation was due to the fact that SNOMED had a much lower percentage of "no codes" (8%) than SNVDO (17%). One would expect SNOMED, which contains over 130,000 terms, to perform much better than SNVDO which contains only 26,000 terms. In spite of the five fold increase in terms with SNOMED, there was not a five fold reduction in "no codes".

For years clinicians have complained that they can't find their cases in the coded databases. Based on the fidelity percentages generated by this study we can understand why they feel this way since they scored the codings 50% "exact match". We believe that there are solutions to the problem. First, the clinicians need to take a more active role in the capture and storage of data in the medical databases. Their role could be expanded to include direct data entry or required attestation to coder entered data. Second, clinicians need to be informed about the organization and structure of the nomenclature systems used to capture and store medical findings. Lastly, an active, centrally managed, vocabulary-use-group needs to develop operational guidelines for vocabulary use. Continued maintenance and use of a vocabulary is necessary to develop a controlled vocabulary.

CONCLUSIONS

The work developing an ideal multipurpose controlled medical vocabulary is far from complete. Neither vocabulary was a clear winner in this comparison which leaves no compelling reason to switch from one vocabulary to the other. SNVDO provided the most consistent responses from the coders, while SNOMED provided slightly, but not statistically significant better, fidelity than SNVDO.

SNOMED was able to provide greater coverage of the domain and had fewer "no code" responses.

The effect of coder on the fidelity of a vocabulary is significant. Considering that this task is often relegated to non-clinical support staff, little attention is paid to the coder's effort by the clinicians. While these people do take their jobs seriously and expend much effort to provide accurate codings they are not at the patient's side and don't have first hand knowledge of that patient's problems. They must often guess at what the clinician meant by the comments in the chart and must repeatedly request clarification from them. For these reasons we believe that data collection must occur at the point of care if we are to obtain much better fidelity.

We are a long way from the goal of delivering fidelity measurements close to 100%. More work needs to be done improving the vocabularies and training the users. In no way do we intend to imply that SNOMED is not an appropriate vocabulary. We do however contend that a vocabulary without usage guidelines is of limited utility. Guidelines must be incorporated into the system so that use is consistent across institutions. The users of the system must be familiar and educated with the vocabularies to be successful using it. The authors believe that this is best accomplished by incorporating vocabulary and nomenclature concepts into the medical curriculum where its structure becomes part of a students' medical knowledge. Only when a critical mass of informatics training is in the curriculum will we see effective use of information technology in the daily practice of medicine.

ACKNOWLEDGMENTS

This work supported in part by National Library of Medicine grant LM07089. We would like to thank Ms. Florence Nelson and Ms. Jan Russell and Drs. Philip Johnson, Eric Pope, and Cecil Moore for their invaluable assistance in conducting this project. Without their help this project would not have been possible.

References

[1] Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L: eds. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. 1993, College of American Pathologists

[2] Cimino JJ, Hripcsak G, Johnson SB, Clayton PD: Designing an Introspective, Multipurpose, Controlled Medical Vocabulary, in Kingsland LC, ed.: *Proc. 13th SCAMC*; Wash., DC; Nov, 1989:513-8

[3] Huff SM, Rocha R: Vocabularies and Data Structure for the Computerized Patient Record, *AMIA Spring Congress,* St. Louis, MO, May, 1993

[4] The American Heritage Electronic Dictionary. 1990, Houghton Miffin Company, Cambridge, MA.

[5] Payne TH, Murphy GR, Salazar AA: How Well Does ICD9 Represent Phrases Used In the Medical Record Problem List?, in Safran C, ed.: *Proc. 17th SCAMC*; Wash., DC; Nov, 1993:654-7

[6] Campbell JR, Payne TH: A Comparison of Four Schemes for Codification of Problem Lists, in Ozbolt JG, ed.: *Proc. 18th SCAMC*; Wash., DC; Nov, 1994:201-5

[7] Weed L: Medical Records that Guide and Teach: NEJM 1968; 278:593-600

[8] Weed L: The Problem Orientated record as a Basic Tool in Medical Education, Patient Care and Research: Ann Clin Res 1971; 3(3):131-4

[9] Klimczak JC; Hahn AW, Hausam RR, Sievert ME, Mitchell JA: A System for Browsing the SNOMED International vocabulary. in Biomedical Sciences Instrumentation Dyer RA ed.: *Proc. 31st Annual Rocky Mountain Symposium*; April 1994, Manhattan, KS. 1994(30):127-32

[10] Klimczak JC, Hahn AW, Sievert ME, Mitchell JA: A Browser for SNOMED III, *in Proc. Spring AMIA*; San Francisco, CA. May, 1994:88

[11] Klimczak JC, Hahn AW, Sievert ME, Mitchell JA: Getting Around in a Large Nomenclature File: Browsing SNOMED International: in Ozbolt JG, ed., *Proc. 18th SCAMC*; Wash., DC, Nov. 1994:1023

[12] McNemar, Q: Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages, Psychometrika, 1947(12). 153-157.

[13] Koch GG, Landis JR, Freeman JL, Freeman DH, Lehnen RG: A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data, Biometrics 1977(33):133-158

[14] SAS Institute Inc., *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SAS Institute Inc., 1989:943