The Efficacy of SNOMED, Read Codes, and UMLS in Coding Ambulatory Family Practice Clinical Records

H.C. Mullins, M.D., Patricia M. Scanland, M.M.T./M.S.,
Dean Collins, B.S., Leslie Treece, Peter Petruzzi, Jr.,
Allen Goodson, and Miriam Dickinson, Ph.D.
Department of Family Practice
University of South Alabama
Mobile, Alabama

This study was initially developed as a traditional quantitative study to determine the level of match of identified clinical terms in three (3) clinical vocabularies. To address concerns raised by a review of the literature and our own experience, a supplemental study to collect qualitative data was added.

Dictated progress notes from a stratified sample of patient visits over a period of four (4) years were used to obtain a representative sample of terms. A total of 144 progress notes were selected taking into consideration the usual demographics plus additional variables.

From the 144 clinical notes, 864 terms were extracted and evaluated by level of match. The within-term effect was highly significant ($F=58.69$, $p \leq .001$), indicating significant differences in the mean level of match for the three coding systems.

Qualitative findings suggest that this and other published studies may not answer questions about the "efficacy of available clinical vocabularies in coding ambulatory family practice clinical records", and additional studies are needed which must be carefully structured and utilize a standardized procedure.

INTRODUCTION

The future of primary health care in the United States may well depend on the development and use of an electronic ambulatory clinical medical record, which is expected to improve the quality of care and reduce costs (1). For recorded clinical data to be useful, it must be structured in such a way that it can be manipulated by computers, which will require a controlled clinical vocabulary. (2)

Practitioners, developers, vendors, and others are asking the question, "what clinical vocabularies are available and suitable to use in ambulatory electronic clinical records, and what are the standards?" There is considerable movement toward developing and adopting standards, but as yet, there is no general agreement, which has seriously hampered the development and use of ambulatory electronic medical records in the U.S.

This study was undertaken in an attempt to determine which of the leading clinical vocabulary candidates, i.e., Read Codes (Ver. 3.1, July 1995), SNOMED (Systematized Nomenclature of Medicine International, Ver. 3.1, Feb 1995) and UMLS (the National Library of Medicine's Unified Medical Language System, 6th Edition, April 95) is most useful for coding primary care clinical information.

Initially, the study was designed, similar to previously reported studies, as a traditional quantitative study to determine the level of match of identified clinical terms in each of the three (3) vocabularies. A review of the literature (3,4,5,6,7) and our own early experience with the study design, raised a number of concerns about the meaningfulness and transferability of such studies. To address our concerns, a parallel study component to collect qualitative data was added to the original design. (8)

METHODS

This study was conducted at the University of South Alabama Department of Family Practice in Mobile. The practice's patient population covers a broad geographic region including urban, suburban and rural areas. Dictated progress notes from patients seen in the clinic for a period of four years were used to obtain a sample which was representative of the broad spectrum of patients. Equal numbers of male and female patients from each of three age groups (0-17; 18-54; and 55+) were randomly selected. This resulted in a sample of 144 progress notes. Ethnicity and gender of both patients and physicians and physician length of practice were also taken into consideration.

0195—4210/96/$5.00

When the final sample had been chosen, two experienced family physicians were each given copies of progress notes and asked to note: 1) basic terms and concepts; 2) any significant synonyms; and 3) the modifiers or qualifiers. Each physician had a unique set of progress notes with no overlap. The physicians did, however, collaborate frequently throughout the process. This step was taken to control for inconsistencies within and among coders and among search procedures for each of the three coding systems.

Developers of each of the systems supplied copies of their browsers for the study. These browsers are computer programs on CD-ROM which demonstrate the structure and capability of each coding system and provide the ability to search for term codes.

Three first year medical students, (two males and one female) were hired for the summer months to work with research staff in evaluating the three systems. These students were given the opportunity to review the existing literature and to use each of the computerized browsers for one week so as to become more familiar with them. A two week pilot study was carried out to evaluate the level of consistency among coders. During that time, students received instruction regarding the format for progress notes and guidelines for using the systems to search for terms. Each of the students was given the same set of progress notes. Every effort was made to control for diversity in training among students and differences in search techniques in order to minimize the inconsistencies. In spite of efforts to control for inconsistencies, the pilot study revealed a low level of agreement among coders. Following the pilot study, additional training was provided to the student coders regarding procedures for searching for terms and coding sheets were developed in an effort to standardize the search protocol.

The actual study began with each of the three student coders being given one third of the progress notes with no overlap. Each coder began searching for all basic terms/concepts and their modifiers in all three coding systems. A coding sheet was completed for each term and its modifiers reflecting the results of the search in each of the three systems. Coding was based on the level of coverage of the term or concept. Five levels of coverage were possible ranging from unmatched to exact one-to-one match ***(Figure 1)***. This format had been used in prior studies reported in the literature and provided for greater standardization. (9)

Data entry and analysis were carried out by research staff. Statistical analysis was performed using SPSS, version 5, for personal computer.

CLASSIFICATION OF MATCHES

- **Good Match**
  - **M4:** exact one-to-one match
  - **M3:** many-to-one match
- **Partial Match**
  - **M2:** main match concept, but modifiers missing
  - **M1:** partial match, main concept
- **No Match**
  - **U:** unmatched concept

Figure 1

RESULTS

The 864 terms from 144 patient records were coded by level of match for each system as described above. A repeated measures analysis of variance was used to compare the mean match level for the three systems and assess possible differences among the three coders in assigning level of match. The five coded match levels were collapsed into three categories for purposes of the remaining analyses: *Unmatched* - Level 0; *Partial Match* - levels 1 and 2; and *Good Match* - levels 3 and 4. Chi square tests were used to assess relationships between the level of match and certain variables of interest for each coding system. Descriptive statistics are given in ***Table 1.***

Table 1

Demographics by Terms	N (%)
Males	355 (41)
Females	509 (59)
Well Visits	300 (35)
Ill Visits	564 (65)
White Patients	392 (45)
Non-White Patients	472 (55)
White Physicians	700 (82)
Non-White Physicians	160 (18)
Male Physicians	728 (84)
Female Physicians	134 (16)
Children or Adolescents (0-17)	253 (29)
Adults (18-54)	264 (31)
Older Adults (55+)	347 (40)
Total Number of Terms	864

Initially, we wished to know if there were differences in the level of match across coding systems and whether the three students coded terms similarly. A repeated measures analysis of variance was used to determine this. Each term was coded by the same student for all three systems. Coding system constitutes the within-terms factor for a total of 2580 observations. Each student received a

total of 2580 observations. Each student received a portion of the terms to code and, therefore, student is the between–terms factor. Data were incomplete for four terms, leaving a total of 860 terms divided among three coders with each term coded using all three systems. The dependent variable was level of match and values ranged from 0 (unmatched) to 4 (exact one-to-one match). The within-term effect was highly significant ($F_{(2, 1722df)}$ = 58.69, $p \leq .001$), indicating significant differences in the mean level of match for the three coding systems. The between-students effect, however, was not significant at the $p \leq .05$ level.

Since all subsequent analyses are based on the three collapsed categories described above, level of match for each coding system is provided according to these catagories in table 2. "Good matches" constituted 54% and 52% of the terms respectively in SNOMED and UMLS, whereas only 40% of the terms coded using Read Code were considered good matches. ***(Table 2)***

Table 2

Level of Match	Read N(%)	SNOMED N(%)	UMLS N(%)
Unmatched	233(27)	150(17)	137(16)
Partial Match	286(33)	245(28)	469(32)
Good Match	345(40)	469(54)	453(52)

We were also interested in determining if there were significant associations between level of match and the following variables: patient age, the part of the S.O.A.P. note from which the term was obtained, type of visit (illness vs. wellness), and ethnicity of physician. These analyses were performed separately for each coding system.

There was a significant association between level of match and patient age for all three coding systems. Percentages and counts are given only for the good matches for each system. The highest percentage of good matches was observed in the middle age group for all three coding systems. However, level of match was more strongly associated with age in SNOMED than in the other two systems. ***(Table 3)***

Table 3
Level of Match by Age
"Good Matches"

Age	Read N(%)	SNOMED N(%)	UMLS N(%)
0-17	93(37)	146(58)	134(53)
18-54	122(46)	170(64)	146(55)
55+	130(38)	153(44)	173(50)
χ^2 4df	9.59 *	29.67 **	10.52 *

Note: ** p<.01

Likewise, significant associations were observed between level of match and the part of the progress note from which the term was derived for all three coding systems . The highest percentage of good matches was observed for terms obtained from the assessment part of the progress note for all three systems. ***(Table 4)***

Table 4
Level of Match by Part of Progress Note
"Good Matches"

Part of Progress Note	Read N(%)	SNOMED N(%)	UMLS N(%)
Subjective	113(37)	157(51)	157(51)
Objective	131(39)	158(47)	158(47)
Assessment	41(52)	55(70)	55(70)
Problem list	43(49)	54(61)	54(61)
Plan	17(34)	29(58)	29(58_
χ^2 8df	30.46 **	23.72 **	37.58 **

Note:* p<.05, ** p<.01

Type of visit and physician ethnicity were significantly associated with level of match for Read Code, but not for UMLS or SNOMED. ***(Tables 5 and 6).***

Table 5
Level of Match by Type of Visit

Type of Visit	Read N(%)	SNOMED N(%)	UMLS N(%)
Ill	100(33)	150(50)	160(53)
Well	245(43)	319(57)	293(52)
χ^2 2df	13.46 **	n.s.	n.s.

Note: ** p<.01

Table 6
Level of Match by Physician Ethnicity

Physician Ethinicity	Read N(%)	SNOMED N(%)	UMLS N(%)
White	270(39)	370(53)	353(50)
Non-White	72(45)	96(60)	96(60)
χ^2 2df	6.44 *	n.s.	n.s.

Note:* p<.05

DISCUSSION

The results of the traditional quantitative component of this study, as reported above, are consistent with findings reported in other studies. None of the vocabularies

performed at a high level of match (90 – 95%) that might be necessary to properly structure clinical records. SNOMED performed at the highest level of good matches at 54%, UMLS at 52%, and Read at 40%. Level of match was significantly associated with age and part of the progress note for all three systems. In addition, level of match was significantly associated with the type of visit and physician ethyniciy for Read code, but not for UMLS or SNOMED.

In terms of good matches, the Read Codes lagged behind the other coding systems in all significant relationships, i.e., age, part of progress note, type of visit, and physician ethnicity, falling below 50% in every category for percentage of matching terms. The Read Codes also had more unmatched terms than the other coding systems in each category. In terms of partial matches overall, however, the Read Codes outranked both SnoMed and UMLS. An observation made by student coders may shed some light on why the Read Codes, which have been widely accepted and utilized throughout the United Kingdom for 10 years, seemed to perform so poorly in comparison to the other coding systems. A number of terms, when not found initially in the Read Codes, were searched by one coder using a different spelling (ex. edema/oedema) indicating a difference between British and American spellings of some words. This may have adversely affected our assessment of that system.

Arguably, the most valuable outcomes were the qualitative collection of questions raised during the course of the study, including:

1. Does the use of "browsers" not designed for actual clinical use, with variable search algorithms, allow valid study results?

2. How meaningful is this study and other studies without standardized methods?

3. Are studies using retrospective coding of previously dictated records transferable or meaningful to point of service use of controlled vocabularies?

4. What is the effect of a structured vocabulary on the choice of terms used as opposed to free form dictation without any concern for a vocabulary?

5, Are study results flawed by inconsistent source and selection of terms to be matched?

6. Can studies that fail to mimic or reflect actual application in a computerized medical record be meaningful?

7. How important is uncontrolled redundancy in the eventual use and/or application?

8. How valid and comparable are studies utilizing coders with varying levels of clinical acumen and training?

9. Does the version and release date of the coding scheme used effect the meaningfulness of study results?

10. Is the retrievability of data an important consideration?

11. How valid is the "degree of term matching" in determining the usefulness of a vocabulary in actual use in a computerized clinical record?

CONCLUSIONS

These qualitative findings suggest that this and other published studies do not answer questions about the "efficacy of available clinical vocabularies in coding ambulatory Family Practice clinical records", and additional studies are needed. We suggest that for future studies to be meaningful and useful:

1. Data must be collected at the point of service in a real life setting.

2. A fully functional automated medical record, pen-based or with a picking list, must be used so that the "coding" is invisible to the user.

3. Investigators must be primary care practitioners.

4. There must be standardization across studies, particularly in regard to study protocol, hardware specification, software specification, and the training of coders.

References

1. Dick RS, Steen EB, eds. The computer-based patient record: An essential technology for health care. Institute of Medicine, National Academy of Sciences. 1991
2. Rector AL, et al. Foundations for an electronic medical record. Yearbook of Medical Informatics. 1992
3. Campbell James R., M.D. et al. A comparison of four schemes for codification of problem lists. Proc 18th Ann Symp Comput Applic Med Care. 1994.
4. Henry Suzanne B., R.N. et al. Representation of nursing terms for the description of patient problems using SNOMED III. Proc 18th Ann Symp Comput Applic Med Care. 1994.
5. Hausam Robert R., M.D. et al. Representation of clinical problem assessment phrases in U.S. Family Practice using Read version 3.1 terms: A preliminary study.Proc 19th Ann Symp Comput Applic Med Care. 1995.

6. Duisterhout JS, et al. Implementation of ICPC coding in information systems for primary care. MEDINFO. 1992.
7. Payne, TH, et al. How well does ICD9 represent phrases used in the medical record problem list?. Proc 16th Ann Symp Comput Applic Med Care.1992.
8. Yin RK.Case study research, design and methods. Beverly Hills, CA. Sage,1984.
9. Bouhaddou Omar et al. Evaluating How the UML Meta1.1 covers disease information contained in a diagnostic expert system (Iliad). AMIA Spring Congress.1993.