Comparing SNOMED and ICPC Retrieval Accuracies Using Relational Database Models

Yves A. Lussier, B. Eng., M.D.[1], Michel Bourque, Ph.D. [2]

[1] SNOMED International-French Secretariat, Faculty of Medicine, Université de Sherbrooke
Sherbrooke, P.Q., CANADA, J1H 5N4
[2] Clinidata, 2 Place Alexis-Nihon, suite 1600, Montréal, P.Q., CANADA H3Z 3C1

While SNOMED International has been generally accepted by the international community of pathologists, its use for primary and secondary care remains limited. This can probably be attributed to the coding complexity of clinical concepts into this multiaxial postcoordinated nomenclature. The SNOMED editors propose the use of multiple codes (aggregates) for any nuanced clinical concept, thus allowing alternative rigorous representations of the concept with SNOMED codes. Some classification critics argue whether such redundant coding precludes precise retrieval of data.

This research was initiated to compare the retrieval accuracies of a relational database using a simplified model of SNOMED against a classification-based model.

SNOMED-based queries showed improvement over ICPC-based queries, regardless of the use of SNOMED cross-references. The addition of the latter significantly improved the queries sensitivity and false negative rate.

In conclusion, the authors recommend using aggregates of SNOMED codes in relational database designs over classification-based designs in order to improve retrieval accuracy.

INTRODUCTION

Wingert previously published a detailed indexing methodology[1] for the *International Systematized Nomenclature of Human and Veterinary Medicine* (**SNOMED**)[2], that uses the multiple hierarchies for every SNOMED concept and which supports a concept coded with SNOMED code aggregates. This published algorithm applies a recursive query to the multiple level of hierarchies inherited by a SNOMED code from its axis or from its *cross-references* (**Xref**). Although SNOMED indexing strategies have been evaluated previously[3,4,5]; to our knowledge, the retrieval of SNOMED concepts from a relational database supporting Xref and code aggregates has not been objectively demonstrated. The absence of evaluation for SNOMED Xref retrieval capacities resides in the incompleteness of the Xref. We estimate that over 300,000 Xref are missing from the diagnosis axis (D) alone, while it already contains 50,000 such Xref to the system/organ anatomy (T), the morphology (M), the function (F), the physical agents (A), the chemicals and drugs (C), and the general modifiers (G) axes.

The expressiveness of SNOMED might theoretically limit its capacity to retrieve equivalent but unrecognizable code aggregates. The necessity for formal SNOMED coding rules to improve its power to recognize computational equivalences has already been described[1,6,7].

OBJECTIVES AND HYPOTHESES

The main **objective** was to demonstrate that the simplifying hypotheses, involved in a previously published relational database[8] model of SNOMED, provided better retrieval properties than the same clinical data encoded into a legacy classification. The secondary goal of this research was to validate the design of an improved, SNOMED-based, drug-disease contraindication alarm software for a Computerized Patient Record.

This preliminary evaluation consisted of comparing SNOMED to a biaxial classification: *the International Classification of Primary Care* (**ICPC**)[9]. While SNOMED contains approximately 140,000 codes, ICPC was purposely developed into a simple classification of less than 1,000 codes. ICPC is a well evaluated multilingual primary care classification[9].

The research **hypotheses** were the following:

1- SNOMED contains more concepts and more granularity than ICPC. *SNOMED-based queries not using SNOMED cross-references* (**S-Xref**) should therefore present a better *positive predictive value* (**PPV**) than ICPC.

2- The cross-references of SNOMED provide multiple hierarchies for every SNOMED code, therefore S*NOMED-based queries using the cross-references* (**S+Xref**) should show more sensitivity than the ones based on ICPC.

1091—8280/97/$5.00 © 1997 AMIA, Inc.

3- The proposed model compares multiple SNOMED codes for a clinical concept to one ICPC code, thus every clinical concept can inherit multiple SNOMED hierarchies. Hence, the *false negative rate* (**FNR**) of SNOMED should be lower than the ICPC FNR.

METHODOLOGY

The clinical data was stored into a previously published relational database model of SNOMED[8]. The database was developed under Access 7.0 for Windows 95. The data model supports multiple SNOMED codes for one clinical data concept. The following definition represents a clinical concept in the database:

$$\sum_{i=1}^{n} G_{Di}Di+G_{Ti}Ti+G_{Mi}Mi+G_{Fi}Fi+G_{Li}Li+G_{Ai}Ai+G_{Ci}Ci$$

The definitions of the axes letters are found in the introduction section.

This equation provides a simplification of a general declarative semantic representation of clinical data[6]. The adapted model for a relational database uses native SQL queries on SNOMED codes without an external semantic analyzer but it does not support explicit relationship between SNOMED codes.

The clinical data was provided by Dr. Robert Bernstein from a Primary Care Dictionary (PCdic) of diagnoses, symptoms, and chief complaints. All 9,297 PCdic terms were already precoded in ICPC. A subset of 2,739 codes was selected for economical reasons.

The clinical concepts of this subset of PCdic codes were precoded into SNOMED version 3.3 by the SeeSNO software[10] using a previously presented methodology[11]. Each clinical data entity was encoded by four trained encoders, into one or more SNOMED codes according to the previous model. Every encoded term was revised by the author, instructed in SNOMED encoding by the SNOMED editor, Dr. Roger A. Côté. Finally a third revision was done, which consisted in comparing the SNOMED label that was typed beside every SNOMED code to the SNOMED term from the original SNOMED database. This last revision confirmed that the intended SNOMED codes were not mistyped.

Here is an example of a coded PCdic term: "Toxoplasmosis in pregnancy" was provided with the ICPC code W84 "Pregnancy high risk" and was coded as SNOMED DE-51200 "Toxoplasmosis" and SNOMED F-84000 "Pregnancy".

Fifty *drug-disease contraindication* queries (**DDC**) were selected from an existing drug-disease contraindication alarm software[12]. The DDC comprise a large scope of general to very specific subsets of clinical data classes. The selected DDCs were encoded into ICPC or SNOMED queries by coding specialists and served as retrieval criteria in the database.

Every DDC query was performed according to three strategies:

- on ICPC encoded data
- on SNOMED encoded data without Xref
- on SNOMED encoded data with Xref

The number of false negative (missed records by the query) were manually assessed by the author, who revised the PCdic for terms omitted by the query. The evaluation of the DDC query results was similarly done to determine the false positive rate. Furthermore, we have previously stated that many SNOMED Xref were missing. The analysis of a SNOMED query using the Xref was manually completed by the author, when necessary, thus creating a possible bias.

The DDC query results were analyzed as follows:

- the true positive number of records (a)
- the true negative number of records (b)
- the false positive number of records (c)
- the false negative number of records (d)

The calculated accuracy tests:

• sensitivity =	a/(a+c)
• specificity =	d/(b+d)
• PPV =	a/(a+b)
• FNR =	c/(a+c)

The G-test of independence for multiple comparisons[13] was used to compare the observed frequencies of the previous accuracy criterias. Compared accuracy tests were considered significant when $p < 0.05$.

RESULTS AND ANALYSIS

The SNOMED coding specialists took more than 250 hours to encode 2,739 codes and the author took over 100 hours to revise them. This laborious process is congruent with similar observations[14].

Three of the 50 DDC queries generated an empty data set with no false positive nor false negative results and were consequently rejected from the analysis.

Table 1 shows specific query results for two instances of the 47 selected queries.

Table 1: Examples of query results

Query	Query strategy	Sensi-tivity	PPV	FNR
Mela-noma	ICPC	57 %	31 %	43 %
	S-Xref	100 %	100 %	0 %
	S+Xref	100 %	100 %	0 %
Infec-tion	ICPC	78 %	100 %	22 %
	S-Xref	59 %	100 %	41 %
	S+Xref	100 %	100 %	0 %

S = SNOMED, Xref = cross-references
PPV = positive predictive value
FNR = false negative rate

The designs for the ICPC and SNOMED queries of table 1 are shown below:

"Melanoma" query

ICPC

- S77 Malignant neoplasm of skin

SNOMED

- the prefixes M-872** to M-879** Nevi and Melanomas
- and SNOMED fifth-digit behavior code for neoplasm
  2 Carcinoma in situ
  3 Malignant, primary site
  Example: M-872*2 or M-872*3 or ...

"Infection" query

ICPC

- 71different codes!
  Example: A70 or A71 or A72 or ...

SNOMED

- DE-***** Infectious and parasitic diseases
- or L-1**** or L-2**** Bacteria or Rickettsiae
- or L-3**** Viruses
- or L-4**** Fungi - Mycetae
- or L-5**** Parasites - Protozoa and Helminthes

Table 2 provides a summary of the sensitivity, specificity, PPV, and FNR. Significant differences can be observed between any SNOMED and ICPC queries for all accuracy criteria with the exception of the specificity. Significant differences between S-Xref and S+Xref were only observed for sensitivity and FNR.

Table 2: Comparing ICPC to SNOMED query accuracies

Criteria	# Queries based on ICPC	# Queries based on S-Xref	# Queries based on S+Xref
Sensitivity > 85 %	17/47	35/47	47/47
Sensitivity > 95 %	11/47	35/47	47/47
Specificity > 99 %	46/47	46/47	47/47
PPV > 80 %	34/47	45/47	46/47
PPV > 95 %	27/47	44/47	45/47
FNR < 20 %	24/47	40/47	47/47
FNR < 5 %	20/47	39/47	47/47

S = SNOMED, Xref = cross-references
PPV = positive predictive value
FNR = false negative rate

DISCUSSION

Both SNOMED and ICPC query designs rely on a comprehensive knowledge of the nomenclature or the classification. SNOMED queries were generally simple to write, on the other hand, SNOMED encoding of clinical concepts is more time consuming and complex.

Table 1 shows a meager performance of S-Xref sensitivity compared to ICPC for the query on "Infection". Since many infections are not classified in the DE-***** section of SNOMED, the S+Xref query relies on the Xref to the infectious living organism to provide a better sensitivity than S-Xref and ICPC. For very specific queries, both S-Xref and S+Xref performed better than ICPC because of the granularity of SNOMED (refer to the "Melanoma" query). On very specific queries, the PPV shows poorer results for ICPC than for SNOMED because of the increased number of false positive occurences in the ICPC query. For example, the ICPC-based "Melanoma" query S-77 "Malignant neoplasm of skin", uses the non specific search. It is nevertheless the most precise ICPC class available for "Melanoma". The poor FNR observed for ICPC and S-Xref can be attributed to their use of only one hierarchy. ICPC recall of congenital, childhood, and pregnancy diseases was generally low.

Table 2 shows a progressive improvement of query accuracy according to the coding strategy. The results validated the presumptions previously cited with each of the three hypotheses.

A comprehensive study using S+Xref queries without a manual revision should be performed when the SNOMED editors have finished writing the Xref, and would provide an unbiased comparison. Unfortunately this analysis is not presently possible.

Queries based on S-Xref performed better than expected. The false positive rate was sufficient to justify an improvement over ICPC-based queries for a drug-disease contraindication alarm software. S-Xref lack of sensitivity compared to the queries based on S+Xref were attributed to the same problems as those observed with ICPC but to a lower extent. Clinical concepts were classified in one hierarchy for both S-Xref and ICPC, thus lowering the sensitivity of the query compared to S+Xref.

The lack of significant differences between SNOMED and ICPC specificity is due to the denominator of the calculation: the "true negative + the true positive" records for the query. Considering that the queried data set contained 2,739 records and that the average query result contains 12 records, the overall non discriminative specificity can be explained.

CONCLUSION

The expressiveness of SNOMED code aggregates has been criticized for its redundancy[15]. However the use of the Xref probably allows a sound query to extract equivalent, but different, codings for the same clinical concept. While the results presented do not provide a formal proof, they give an objective insight on the performance of queries according to the use of SNOMED or of a classification. The consequent *theoretical risk* of unretrievable equivalent expressions of SNOMED codes outperformed a classification most probably because of the latter *proven lumping* of codes into classes. An analysis comparing ICD-10 to SNOMED would provide complementary evidence and is being designed by the authors.

The overall accuracy of SNOMED over ICPC for querying the relational database has been demonstrated, thus allowing a simple solution to improve legacy classification-based Computerized Patient Records.

Meaningful research is being conducted on Natural Language Processing and indexing of clinical text using SNOMED, nonetheless more effort should be devoted to the analysis of the new retrieval features emerging from the use of a multiaxial, compositional, cross-referenced, nomenclature.

Acknowlegments

This research was partly supported by a grant from the Medical Research Council of Canada (HEALNet) and by Clinidata Inc.

The authors are indebted to both Dr. Roger A. Côté and the College of American Pathologists for the use of SNOMED nomenclature for research purposes and also to Dr. Robert M. Bernstein, who provided the clinical data dictionary for primary care.

References

[1] Wingert F. An Indexing System for SNOMED. Methods of Information in Medicine 1986;25 (1):22-30.

[2] Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, editors. The Systematized Nomenclature of Human and Veterinary Medicine. College of American Pathologists; 1993.

[3] Oliver DE, Altman RB. Extraction of SNOMED Concepts from Medical Record Texts. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care 1994, Washington, DC, Nov 1994:179-83.

[4] Brigl B, Mieth M, Haux R, Glück E. The LBI-method for automated indexing of diagnoses by using SNOMED. Part 2. Evaluation. International Journal of Bio-Medical Computing 1995;38:101-8.

[5] Do Amaral MB, Satomura Y. Associating Semantic Grammars with the SNOMED: Processing Medical Language and Representing Clinical Facts into a Language-Independent Frame. Proceedings of the Eighth World Congress on Medical Informatics 1995:18-22.

[6] Campbell KE, Das AK, Musen MA. A Logical Foundation for Representation of Clinical Data. JAMIA May/Jun 1994;1 (3):218-32.

[7] Cimino JJ. Review Paper: Coding Systems in Health Care. Methods of Information in Medicine 1996;35 (4/5):273-84.

[8] Lussier YA, Côté RA. Harnessing SNOMED: A Relational Database Design for the

Computerized Patient Record (CPR). Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care, New Orleans, La., Oct-Nov 1995:1019.

[9] Lamberts H, Wood M, Hofmans-Okkes I, editors. The International Classification of Primary Care in the European Community with a multi-language layer. New York: Oxford University Press Inc.; 1993.

[10] MedSight Informatique Inc., 1801 McTavish, St-Bruno, P.Q., CANADA J3V 4G2.

[11] Lussier YA, Côté RA. A Comprehensive Analysis of SNOMED International Encoding of Symptoms, Signs and the Patient Record Structure. AMiA Spring Congress, Boston, June 24-28, 1995.

[12] Clinidata, 2 Place Alexis-Nihon, suite 1600, Montréal, P.Q., CANADA H3Z 3C1.

[13] Sokal RR, Rohlf FJ. Biometry. 3rd ed. Freeman; 1995. P.729.

[14] Pole PM, Rector AL. Mapping the GALEN CORE Model to SNOMED-III: Initial Experiments. Proceedings of the Annual Fall Symposium AMIA 1996, Washington, DC, Oct 1996:100-4.

[15] Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a Medical-concept Representation Language. JAMIA May/Jun 1994;1 (3):207-17.