A RANDOMIZED DOUBLE-BLIND CONTROLLED TRIAL OF AUTOMATED TERM DISSECTION

P. L. Elkin, MD; K. R. Bailey, PhD; P. V. Ogren; B. A. Bauer, MD; C. G. Chute, MD, DrPH
Mayo Foundation, Rochester, MN

Abstract

Objective: To compare the accuracy of an automated mechanism for term dissection to represent the semantic dependencies within a compositional expression, with the accuracy of a practicing Internist to perform this same task. We also compare the results of four evaluators to determine the inter-observer variability and the variance between term sets, with respect to the accuracy of the mappings and the consistency of the failure analysis.

Methods: 500 terms, which required a compositional expression to effect an exact match, were randomly distributed into two sets of 250 terms (Set A and Set B). Set A was dissected using the Automated Term Dissection (ATD) Algorithm. A physician specializing in Internal Medicine dissected set B. He had no prior knowledge of the dissection algorithm or how it functioned. In this manuscript, the authors use Human Term Dissection (HTD) to refer to this method.

Set A was randomized to two sets of 125 terms (Set A_1 and Set A_2). Set B was randomized to two sets of 125 terms (Set B_1 and Set B_2). A new set of 250 terms Set C was created from Set A_1 and Set B_2. A second new set of 250 terms Set D was created from Set A_2 and Set B_1.

Two expert Indexers reviewed Set C and another two expert Indexers reviewed Set D. They were blinded to which terms were dissected by the clinician and which terms were dissected by the automated term dissection algorithm. The person providing the files for review to the Indexers was also unaware of which terms were dissected by ATD vs. the HTD method.

The Indexers recorded whether or not the dissection was the best possible representation of the input concept. If not, a failure analysis was conducted. They recorded whether or not the dissection was in error and if so was a modifier not subsumed or was a Kernel concept subsumed when it should not have been. If a concept was missing, the Indexers recorded whether it was a Kernel concept, a modifier, a qualifier or a negative qualifier.

Results: The ATD method was judged to be accurate and readable in 265 out of the 424 terms with adequate content (62.7%). The HTD method was judged to be accurate in 272 out of 414 terms with adequate content (65.7%). There was no statistically significant difference between the rates of acceptability of the ATD and HTD methods ($p=0.33$).

There was a non-significant trend toward greater acceptability of the ATD method in the subgroup of terms with three or more compositional elements. ATD was acceptable in 53.6% of the terms where the HTD was only acceptable in 43.6% ($p=0.11$). The failure analysis showed that both methods misrepresented kernel concepts and modifiers much more commonly than qualifiers ($p<0.001$).

Conclusions: There is no statistically significant difference in the accuracy and readability of terms dissected using the automated term dissection method when compared with human term dissection, as judged by four expert medical indexers. There is a non-significant trend toward improved performance of the ATD method in the subset of more complex terms. The authors submit that this may be due to a tendency for users to be less compulsive when the time to complete the task is long. Automated term dissection is a useful and perhaps preferable method for representing readable and accurate compound terminological expressions.

Introduction

As we move toward compositional terminologies, the need to organize the terms within a compositional expression becomes important for both the readability and understanding of these composite terms.[1,2] This trial evaluates a mechanism for automated term dissection using the semantic types available from within the Unified Medical Language System (UMLS). The system uses the Metaphrase™ search engine to retrieve a list of suggested terms, which serve as the substrate for compositional expressions. These terms are analyzed to find the best controlled terms to construct the compositional expression. We call this automated term composition, and reported this at the 1998 fall AMIA symposium.[3] For each term selected we determine its

1091—8280/99/$5.00

semantic type. We divide the semantic types into those that represent Kernel concepts, Modifiers, Qualifiers or Negative Qualifiers.[4] A rule base is then applied which organizes the Modifiers, Qualifiers and Negative Qualifiers around the Kernel concepts. These are represented in a hierarchical structure with the degree of indentation being representative of semantic dependency. To date the accuracy of this automated technique has not been evaluated. Many individuals have evaluated the accuracy of manual term composition.[5,6] The clinical coding center of the NHS has reported limited success with their own algorithm for automated term dissection in the past.[7,8] The method described in this manuscript is designed to evaluate this automated approach to term dissection.

To illustrate term dissection we reference a case of a 62 year old female who presents with erythema over the dorsum of the left foot with exquisite tenderness over a wound situated over the mid foot. After a comprehensive clinical work up she was found to have a Cellulitis of the left foot with Osteomyelitis of the third metatarsal without signs of lymphangitic spread of her infection.

The Automated term dissection method organizes the controlled representation of this patient's diagnosis (using UMLS CUIs) in an XML dtd and provides the following graphical representation to the user:

"Cellulitis, NOS
Left Foot
Osteomyelitis
Third Metarsal Bone, NOS
Without
Lymphangitis."

History of Classification

The present coding practices rely on data methods and principles for terminology maintenance that have changed little since the adoption of the statistical bills of mortality in the mid-17th century. The most widely accepted standard for representing patient conditions is ICD9-CM[9] and is an intellectual descendent of this tradition. ICD9-CM relies overwhelmingly on a tabular data structure with limited concept hierarchies and no explicit mechanism for synonymy, value restrictions, inheritance or semantic and non-semantic linkages. The maintenance environment for this healthcare classification is a word processor and its distribution is nearly exclusively paper-based.

Significant cognitive advances in disease and procedure representation took place in 1928 at the New York Academy of Medicine, which results in industry-wide support for what became the Standard Nomenclature of Diseases and Operations. The profound technical innovation was the adoption of a multiaxial classification scheme. Now a pathologic process (e.g. Inflammation) could be combined with an anatomic sight (e.g. Oropharynx Component: Tonsil) to form a diagnosis (e.g. Tonsillitis). The expressive power afforded by the compositional nature of multiaxial terminological coding system, tremendously increased the scope of tractable terminology and additionally the level of granularity that diagnosis could be encoded about our patients.[10]

The College of American Pathology (CAP) carried the torch further by creating the Systematized Nomenclature of Pathology (SNOP), and subsequently the Systemized Nomenclature of Medicine (SNOMED). In these systems, the number, scope, and size of the compositional structures has increased to the point where an astronomical number of terms can be synthesized from SNOMED atoms. One well-recognized limitation of this expressive power is the lack of syntactic grammar, compositional rules, and normalization of both the concepts and the semantics. Normalization is the process by which the system knows that two compositional constructs with the same meaning are indeed the same (e.g. that the term "Colon Cancer" is equivalent to the composition of "Malignant Neoplasm" and the site "Large Bowel"). These are issues addressed by CAP in their efforts to make SNOMED a robust reference terminology for medicine.[10,11]

Other initiatives of importance are the Clinical Terms v3 (Read Codes) which are maintained and disseminated by the National Health Service in the United Kingdom, and the Galen effort which expresses a formalism for term description which is very detailed.[12,13] The Read Codes are a large corpus of terms, which is now in its third revision, that is hierarchically designed and is slated for use throughout Great Britain.

Methods

Study Design

The most common 5,000 diagnoses were obtained from the Master Sheet Index and the Impression section of the Clinical Notes system at the Mayo Clinic. The Master Sheet Index consists of final diagnoses assigned by the primary care physician after each episode of care for a patient. For instance if a patient presents complaining of "Chest Pain" and after an extensive work up it was found that the cause of the chest pain was "Esophageal Spasm," then the master sheet entry would just be "Esophageal Spasm" and not chest pain which might imply a cardiac condition.

From these 5,000 terms, 500 were selected. They were specifically chosen from a subset of the 5,000 terms, which required a compositional match to claim equivalence on a conceptual level. Preference was given to those terms whose equivalent compositional match required more than two controlled terms. There were 1,328 terms, which required a compositional expression to effect an exact match. Of these 1,328 compositional expressions 1,173 were matched with two concepts from the UMLS and 155 expressions were matched with more than two concepts from the UMLS. From the 1,173 two concept compositional expressions 345 compound concepts were randomly selected. These were added to the 155 expressions formed with more than two UMLS concepts to create the set of 500 compositional matches. (See Figure #1)

These 500 terms were randomly distributed into two sets of 250 terms (Set A and Set B). Set A was dissected using the Automated Term Dissection Algorithm. A physician specializing in Internal Medicine dissected set B. The physician was a full-time, practicing, board certified academic General Internist at the Mayo Clinic. He had no prior knowledge of the dissection algorithm or how it functioned. All controlled terms were presented to him on one level and using a drag-and-drop motif he was asked to create the dependency structure which he felt was the most accurate and readable representation.

Set A was randomized to two sets of 125 terms (Set A_1 and Set A_2). Set B was randomized to two sets of 125 terms (Set B_1 and Set B_2). A new set of 250 terms Set C was created from Set A_1 and Set B_2. A second new set of 250 terms Set D was created from Set A_2 and Set B_1.

Two expert Indexers reviewed Set C and another two expert Indexers reviewed Set D. They were blinded to which terms were dissected by the clinician and which terms were dissected by the automated term dissection algorithm. The person providing the files for review to the Indexers was also unaware of the mechanism used for the dissection.

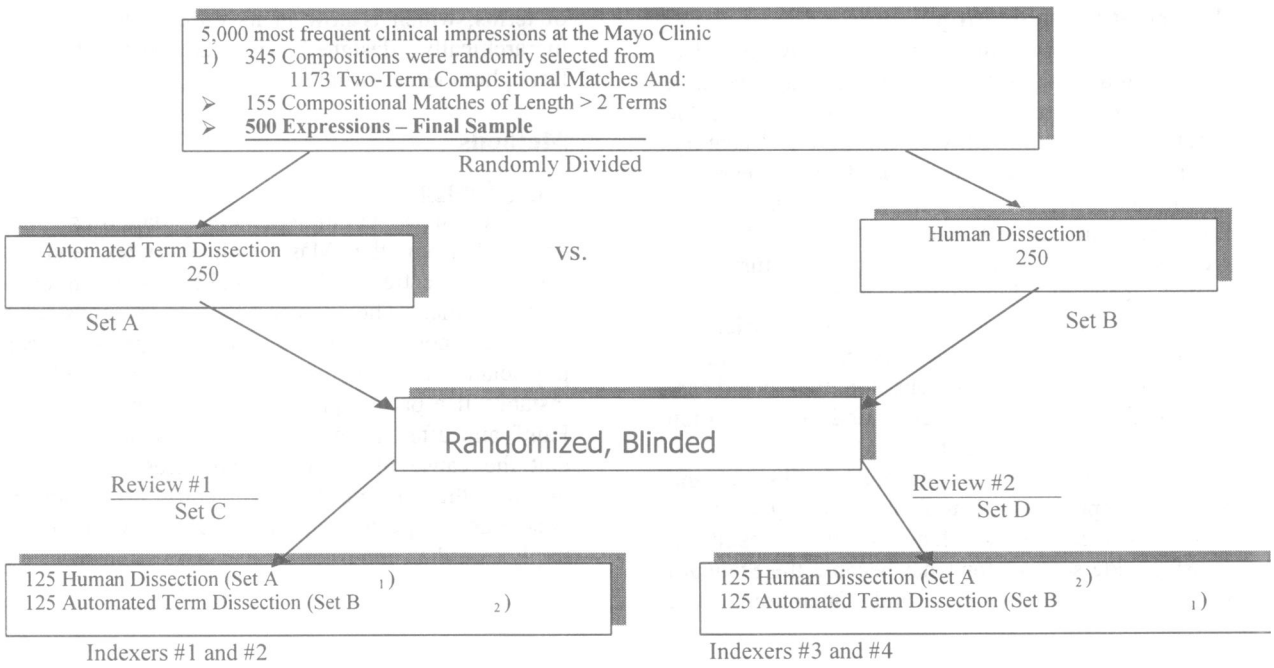
The Indexers recorded whether or not the dissection was the best possible representation of the input concept. If not, a failure analysis was conducted. They recorded whether or not the dissection was in error and if so was a modifier not subsumed or was a Kernel concept subsumed when it should not have been. If a concept was missing, the Indexers recorded whether it was a Kernel concept, a modifier, a qualifier or a negative qualifier.

The results were compared as to the rate that the reviewers judged the dissections from the clinician and the automated term dissection to be accurate. The variability between reviewers will also be analyzed. The failure analysis was analyzed to see if the reasons for poor dissections were different when the clinician's work was compared with the automated term composition routine.

Statistical Analysis

In addition to descriptive comparisons of the accuracy rates for each type of dissection, a few formal statistical comparisons were performed. We tested for equality of accuracy rates between ATD

Figure 1. Automated Term Dissection Trial – Study Design



and HTD systems, within each reviewer by the Pearson Chi-square statistic for equality of proportions. The results were pooled for an overall estimate of the difference in accuracy rates. Finally, an assessment of inter-reviewer variability was performed, by comparing accuracy rates between reviewers who reviewed the same dataset, using the McNemar test.

Results

The Automated term dissection (ATD) algorithm was deemed accurate and readable by the reviewers for 265 terms out of 424 terms deemed to have accurate compositional content (76 terms were deemed inaccurate content matches), making 62.7% of the dissections judged to be accurate. The Human term dissection (HTD) was deemed accurate and readable by the reviewers for 272 terms out of the 414 terms with accurate content (86 terms were felt to be inaccurate content matches), making 65.7% of the dissections judged to be accurate by the reviewers. Inaccurate content matches were excluded from the analysis, as their results did not reflect on the quality of the dissections. The rates of acceptable dissection using the ATD and HTD methods were not statistically significantly different, with a two-sided p-value = 0.33. (See Table 1.)

Special attention was paid to the subset of terms with three or more compositional elements. When we looked at this subset of the results there was a non-significant trend toward improved acceptability of the ATD algorithm as compared with the HTD method (p=0.11). In this subgroup analysis the ATD algorithm was judged acceptable in 75 terms out of 140 terms (53.6%) and the HTD was found to be acceptable in 51 out of 117 terms (43.6%). (See Table 2.)

Failure analysis showed that of the terms not judged to be completely accurate, 60% of the ATD and 70% of the HTD terms had subsumptive problems with a Kernel concept. Modifiers were incorrectly subsumed in 70% of the ATD and 68% of the HTD terms. All types of qualifiers were incorrectly subsumed in 9.4% of the ATD and only 2.1% of the HTD terms (See Table #3). Qualifiers were misrepresented much less commonly than either kernel concepts or modifiers by both methods ($p<0.001$).

Table 1. All Compositional Expressions: Summary Data Table

Reviewer	A	B	C	D	E	F	G	H	I	J
R_1 SetD_ATD	43	101	13	19	12	31	3	1	4	1
R_1 SetD_HTD	42	98	9	23	14	31	1	0	0	0
R_2 SetD_ATD	78	106	4	1	12	12	0	3	0	1
R_2 SetD_HTD	100	102	1	0	1	0	0	0	0	0
R_3 SetD_ATD	59	110	1	42	1	37	0	2	0	0
R_3 SetD_HTD	54	105	7	41	6	34	0	0	0	0
R_4 SetD_ATD	85	107	13	2	6	0	0	0	0	0
R_4 SetD_HTD	76	109	17	2	11	0	2	0	0	0

Table 2. Three Term Compositional Expressions: Summary Data Table

Reviewer	A	B	C	D	E	F	G	H	I	J
R_1 SetD_ATD	15	34	8	2	5	9	1	1	1	1
R_1 SetD_HTD	5	26	3	7	7	8	1	0	0	0
R_2 SetD_ATD	15	32	2	1	12	4	0	2	0	1
R_2 SetD_HTD	23	25	1	0	1	0	0	0	0	0
R_3 SetD_ATD	19	38	0	14	1	12	0	2	0	0
R_3 SetD_HTD	11	32	6	16	6	10	0	0		
R_4 SetD_ATD	26	36	4	2	4	0	0	0	0	0
R_4 SetD_HTD	12	34	11	1	8	0	2	0	0	0

Legend for Tables 1 & 2: **A**=Exact match; **B**=No. of terms with accurate content; **C**=Kernel indented inappropriately; **D**=Kernel not indented appropriately; **E**=Modifier indented inappropriately; **F**=Modifier not appropriately indented; **G**=Qualifier indented inappropriately; **H**=Qualifier not appropriately indented; **I**=Negative Qualifier indented inappropriately; **J**=Negative Qualifier not appropriately indented

There was significant inter-reviewer variability, with three out of the four reviewers showing trends in favor of the ATD algorithm over the HTD. Dissection accuracy rates varied considerably

between the four reviewers (R1 = 43%, R2 = 86%, R3 = 53%, R4 = 75%). In pairwise comparisons these differences were all statistically significant (R1 vs. R2 → $p<0.001$, R1 vs. R3 → $p=0.045$, R1 vs. R4 → $p<0.001$, R2 vs. R3 → $p<0.001$, R2 vs. R4 → $p=0.005$, R3 vs. R4 → $p<0.001$). The overall match rate was 64% ± 20%.

Table 3. Failure Analysis

	ATD	HTD	Total
Kernel Concepts	60%	70%	64.6%
Modifiers	70%	68%	69.0%
Qualifiers	9.4%	2.1%	6.0%

Conclusions

There is no statistically significant difference in the adequacy and readability of terms dissected using the automated term dissection method when compared with human term dissection, as judged by four expert medical indexers. There is a non-significant trend toward improved performance of the ATD method in the subset of more complex terms. The authors submit that this may be due to a tendency for users to be less compulsive when the time to complete the task is long. The validity of this finding will need to be verified in a larger prospective randomized trial.

The ATD and HTD methods were not significantly different with respect to the type of error recorded by our reviewers. The failure analysis showed that kernel concepts and modifiers were misrepresented with equal frequency. However the qualifiers were misrepresented much less often that kernel concepts or modifiers.

Compositional terminologies are one promising answer to the problem of clinical content completeness.[14] High quality controlled health vocabularies provide a gateway to better clinical data being available for outcomes research, utilization review and improved management of the electronic medical record.[15] This promise is contingent upon data entry mechanisms, which do not disrupt the flow of a busy practice.[16]

Creating well formed compositional expressions using a controlled health vocabulary can be labor intensive and time consuming. Given the ever-increasing demands on clinicians' time, we must work to create mechanisms, which aid the busy clinician as we migrate toward, an electronic clinical environment. Automated tools designed to assist clinicians with the formulation of compositional expressions are necessary if we are to make use of powerful compositional terminologies.

Automated term dissection is a useful and perhaps preferable method for representing readable and accurate compound terminological expressions.

Acknowledgements

The authors wish to thank James Buntrock for programming support, and Karen Elias for secretarial and formatting assistance. This work is supported in part by NLM/AHCPR grant U01 HS/LM08751.

References

[1] Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Compositional and enumerative designs for medical language representation. *JAMIA* 1997;SympSuppl: 620-4.

[2] Schulz EB, Price C, Brown PJ. Symbolic anatomic knowledge representation in the Read Codes version 3: structure and application. *JAMIA* 1997;4(1):38-48.

[3] Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. *JAMIA* 1998;Symp Suppl:765-9.

[4] Chute CG, Elkin PL. A Clinically Derived Terminology: Qualification to Reduction. *JAMIA* 1997;Symp Suppl:570-574.

[5] Rector AL. Thesauri and Formal Classifications: Terminologies for People and Machines. In: CG Chute, RH Baud, JJ Cimino, VL Patel AL Rector. Special Issue on Coding and Language Processing. *Meth Inform Med* 1998;37(4/5):501-509.

[6] Bernauer J, Schoop FM, Schoop D, Pretschner DP. The compositional approach for representing medical concept systems. *MEDINFO'95* 1995;1:70-74.

[7] Price C, Bentley TE, Brown PJ.,Schulz EB, O'Neil M. Anatomical characterisation of surgical procedures in the Read Thesaurus. *JAMIA* 1996;SympSuppl:110-4.

[8] Schulz EB, Barrett JW, Price C. Read code quality assurance: from simple syntax to semantic stability. *JAMIA* 1998;5(4):337-46.

[9] Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS, for the Canon Group. Toward a Medical-Concept Representation Language. *J Am JAMIA* 1994;1:207-217.

[10] Campbell KE, Musen MA. Representation of Clinical Data Using SNOMED III and Conceptual Graphs. *Proc.16th Ann Symp Comput Appl Med Care (SCAMC)* 1992:354-358.

[11] Musen MA, Wieckert KE, Miller ET, Campbell KE, Fagan LM. Development of a controlled medical terminology: knowledge acquisition and knowledge representation. *Meth Inform Med* 1995;34(1-2): 85-95.

[12] Rector AL. Nowlan WA. The GALEN project. *Comput Meth Prog Biomed* 1994;45(1-2):75-8.

[13] Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN project. *Proc 17th Ann Symp Comput Appl Med Care* 1993:414-8.

[14] Rogers JE, Rector AL. Terminological systems: bridging the generation gap. *JAMIA* 1997;Symp Suppl:610-4.

[15] Musen MA, Wieckert KE, Miller ET, Campbell KE, Fagan LM. Development of a controlled medical terminology: Knowledge acquisition and knowledge representation. *Meth Inform Med* 1995;4(1-2):85-95.

[16] Rector AL, Glowinski AJ, Nowlan WA, Rossi-Mori A. Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. *JAMIA* 1995;2(1):19-35.