A Large-Scale Evaluation of Terminology Integration Characteristics

FS McDonald, MD; CG Chute, MD; PV Ogren; D Wahner-Roedler, MD; PL Elkin, MD
Mayo Foundation, Rochester, MN

Objective: To describe terminology integration characteristics of local specialty specific and general vocabularies in order to facilitate the appropriate inclusion and mapping of these terms into a large-scale terminology.

Methods: We compared the sensitivity, specificity, positive predictive value, and positive likelihood ratios for Automated Term Composition to correctly map 9050 local specialty specific (dermatology) terms and 4994 local general terms to UMLS using Metaphrase™. Results were systematically combined among exact matches, semantic type filtered matches, and non-filtered matches. For the general set, an analysis of semantic type filtering was performed.

Results: Dermatology exact matches defined a sensitivity of 51% (57% for general terms) and a specificity of 86% (92% general terms). Including semantic type filtered matches increased sensitivity (75% dermatology; 88% general); as did inclusion of non-filtered matches (98% and 99%). These inclusions correspondingly decreased specificity (filtered: 82% and 74%; non-filtered: 52% and 32%). Positive predictive values for exact matches (93.0% dermatology, 97.6% general) were improved by small but significant ($p<0.001$) margins by including filtered matches (95.1% dermatology, 98.4% general) but decreased with non-filtered matches (89.2% dermatology, 87.8% general). Adding additional semantic types to the filtering algorithm failed to improve the positive predictive value or the positive likelihood ratio of term mapping, in spite of a 2.3% improvement in sensitivity.

Conclusions: Automated methods for mapping local "colloquial" terminologies to large-scale controlled health vocabulary systems are practical (ppv 95% dermatology, 98% general). Semantic type filtering improves specificity without sacrificing sensitivity and yields high positive predictive values in every set analyzed.

Introduction

Large-scale controlled health vocabularies are becoming commercially available. One of the barriers to implementation of these systems is the perception that they will change the way clinicians must represent data. Clinicians fear that the individual flavor of their institution may be compromised by this mandate. In order to move toward standard representation of our patient's conditions, we will need to provide a mechanism for mapping local parlance into large-scale concept based controlled health vocabularies.

Although these colloquial local terminologies are appropriate for integration into a controlled representation for a particular organization they may not be appropriate to disseminate to all users of the terminological system. For example at our institution it is common to list the name of the surgeon who performed the CABG along with the fact that the patient had heart surgery. (This helps clinicians to know the technique used and the accessibility of records such as an Operation Note). Nevertheless, this activity of local colloquial terminology integration will likely be important for most, if not all, efforts to disseminate large-scale controlled health vocabularies.

For a vocabulary to be useful it must evolve and its content must grow. The UMLS contains over 480,000 concepts but still does not cover all clinically useful terminology. Local additions of colloquial terminology fall into one of two categories. One area of needed evolutionary capacity is specialty specific terminology. The other is colloquial additions of a general nature. For example, at Mayo it is common to refer to uncomplicated "low back pain" as "mechanical low back pain." General and specialty specific local terminologies will likely continue to be used and, indeed, add richness to medical vocabularies.

As Cimino states, "...a formal methodology is needed for expanding content."[1] Chute reinforces this statement with the argument that "in the absence of a single, all-embracing health care terminology there needs to be coordination and organizing support for interrelated terminologies..." and that "developers of clinical classifications must consider ways they can develop their systems to become part of an integrated set of terminology systems."[2]

However, if one adds terms to a vocabulary indiscriminately one risks redundancy and combinatorial explosion making the vocabulary unwieldy and difficult to search in a timely fashion. "An alternative approach is to enumerate all the atoms of a terminology and allow users to combine

1091—8280/99/$5.00

them into necessary coded terms, allowing compositional extensibility."[1, 3, 4] One risk of this approach is that it has the potential of making the use of the vocabulary more complex.

We hypothesized that Automated Term Composition (ATC) as developed and recently tested in a randomized controlled trial[5] would allow large-scale coverage of specialty specific and general local vocabularies. This automated process will facilitate the appropriate inclusion of such terms into a larger vocabulary, e.g. UMLS, without creating redundancy within the current vocabulary.

Methods

The Mayo Clinic Department of Dermatology independently developed a lexicon of 9813 terms describing lesions photographed within their practice. This corpus was chosen as a local specialty specific terminology set as described above. A set of 5345 of the most frequently referenced terms were chosen from 1,000,000 terms randomly extracted from the general Mayo Clinic Master Sheet Index and the Impression/Report/Plan section of the Mayo Clinic clinical notes system forming a local general health care terminology.[6]

Each set was examined for common abbreviations, which were mapped to their respective full term descriptions. Designations such as "NEC/NOS", "NEC", and "NOS" were deleted from each set since they did not clarify term meaning. Obvious misspellings were corrected. Exact duplicates were then deleted from each set so that the final sets contained unique terms. This resulted in 9050 dermatology terms and 4994 general terms.

An automated process was then applied to each term from each set in the following cascade fashion (See Figure 1). The term, or search string, was sent to the Metaphrase™ server.[7] The Metaphrase™ server returned one or more result strings. If an exact, one-to-one, match was available and recognized by the ATC algorithm, the term was designated an Exact Match (EM). Result strings were filtered using the semantic type information contained within the UMLS. If ATC could compose a match from this semantic type filtered set, the term was designated a Filtered Composed Match (FCM). If the term was still unmatched, all of the Metaphrase™ result strings, regardless of their associated semantic types, were made available to the ATC algorithm in an attempt to compose a match. If a match could be composed using the non-filtered set of result strings it was designated a Non-Filtered Composed Match (NFCM). If a match could not be composed, the term was designated a Non-Match (NM).

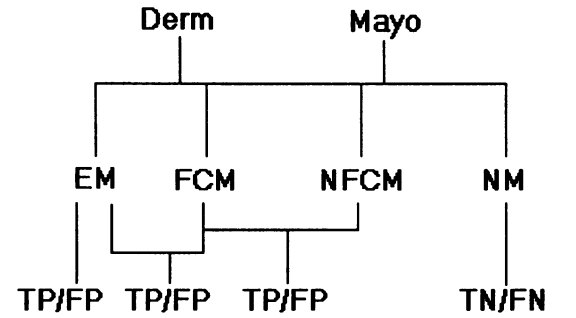


Figure 1: Study Design. The algorithm was applied to the general set and the dermatology set separately and independently. EM = exact, one-to-one term, match. FCM = filtered composed match, i.e. a match constructed from Metaphrase™ result strings which had a specifically chosen set of semantic types. NFCM = non filtered composed match, i.e. a match constructed from the Metaphrase™ result strings regardless of semantic type. NM = non match, i.e. ATC was unable to compose or recognize a match. TP = True Positive. FP = False Positive. TN = True Negative. FN = False Negative.

Each search string and the resulting match string was examined by a practicing Internist to determine if a correct match had been obtained. Having determined the actual matches in each set, the process was analyzed to determine a sensitivity, specificity, positive predictive value, and positive likelihood ratio for the various combinations of sets as follows:

Case 1: EM (Positive Results) – FCM, NFCM, NM (Negative Results).

Case 2: EM, FCM (Positive Results) – NFCM, NM (Negative Results).

Case 3: EM, FCM, NFCM (Positive Results) – NM (Negative Results).

The "true positive", "false positive", "true negative", and "false negative" designations were thus different depending on which sets composed the "positive results" for each automated algorithm. (See Figure 1) Both the dermatology and the general health care sets were examined in similar manners allowing comparisons of the mapping characteristics of these two types of colloquial terminologies.

For the true positive matches within the general NFCM set, the semantic types associated with each true match were examined. This evaluation was performed to determine if the addition of one or more semantic types to the semantic type filtering algorithm would result in increased match rates without unduly increasing false positive rates. A fourth case was then created:

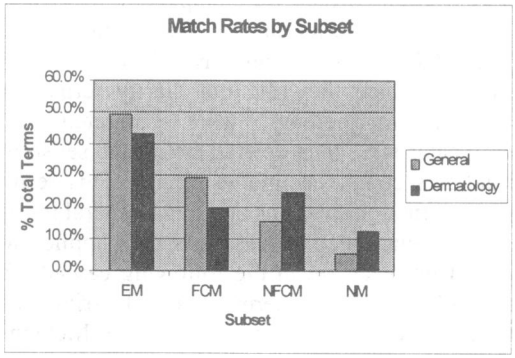
Case 4: EM, FCM with new semantic types (Positive Results) – NFCM, NM (Negative Results).

Table 1: Data by set and subset. EM = exact match. FCM = filtered composed match. NFCM = non-filtered composed match. NM = non-match. See text for details.

	Genl Set	% Total	% Subset	Derm Set	% Total	% Subset	Match Rates
Total Terms	4994	100%		9050	100%		Pearson Chi Square Test
EM	2469	49.4%	100%	3908	43.2%	100%	p<0.001
True Match	2409	48.2%	97.6%	3634	40.2%	93.0%	p<0.001
False Match	60	1.2%	2.4%	274	3.0%	7.0%	p<0.001
FCM	1466	29.4%	100%	1757	19.4%	100%	p<0.001
True Match	1328	26.6%	90.6%	1676	18.5%	95.4%	p<0.001
False Match	138	2.8%	9.4%	81	0.9%	4.6%	p<0.001
NFCM	787	15.8%	100%	2246	24.8%	100%	p<0.001
True Match	476	9.5%	60.5%	1637	18.1%	72.9%	p<0.001
False Match	311	6.2%	39.5%	609	6.7%	27.1%	p<0.001
NM	272	5.4%	100%	1139	12.6%	100%	p<0.001
True Negative	245	4.9%	90.1%	1029	11.4%	90.3%	p<0.001
False Negative	27	0.5%	9.9%	110	1.2%	9.7%	p<0.001

Table 2. Comparison of General and Specialty Specific (Dermatology) Case Specific Characteristics. Case 1, Case 2, Case 3 as defined in text. Sensitivity = True Positives / (True Positives + False Negatives). Specificity = True Negatives / (True Negatives + False Positives). PPV = positive predictive value = True Positives / (True Positives + False Positives). PLR = positive likelihood ratio = Sensitivity / (1 – Specificity).

	Sensitivity	Specificity	PPV	PLR
General **Case 1**	56.8%	92.0%	97.6%	7.14
Dermatology **Case 1**	51.4%	86.3%	93.0%	3.75
Pearson Chi Square p value	p<0.001	p<0.001	p<0.001	p<0.001
General **Case 2**	88.1%	73.7%	98.4%	3.34
Dermatology **Case 2**	75.2%	82.2%	95.1%	4.22
Pearson Chi Square p value	p<0.001	p<0.001	p<0.001	p<0.001
General **Case 3**	99.4%	32.5%	89.2%	1.47
Dermatology **Case 3**	98.4%	51.6%	87.8%	2.04
Pearson Chi Square p value	p<0.001	p<0.001	p<0.001	p<0.001



Results

Of the 4994 general terms, 2469 (49%) were EM, 1466 (29%) were FCM, 787 (16%) were NFCM, and 272 (5.5%) were NM. Of the general EM, 2409 (97%) were actual matches. Of the general FCM 1328 (91%) were actual matches. Of the general NFCM 476 (61%) were actual matches. Of the general NM, 27 (10%) were actual matches. Using EM only as the positive set resulted in a sensitivity of 57%, a specificity of 92%, a positive predictive value of 98%, and a positive likelihood ratio of 7.14. Combining EM and FCM as the positive set resulted in a sensitivity of 88%, a specificity of 74%, a positive predictive value of 98%, and a positive likelihood ratio of 3.34. Combining EM, FCM, and NFCM as the positive set resulted in a sensitivity of 99%, a specificity of 32%, a positive predictive value of 89%, and a positive likelihood ratio of 1.47. (See Tables #1 and #2)

Of the 9050 dermatology specific terms, 3908 (43%) were EM, 1757 (19%) were FCM, 2246 (25%) were NFCM, and 1139 (13%) were NM. Of the dermatology EM, 3634 (93%) were actual matches. Of the dermatology FCM 1676 (95%) were actual matches. Of the dermatology NFCM 1637 (73%) were actual matches. Of the dermatology NM, 110 (10%) were actual matches. Using EM only as the positive set resulted in a sensitivity of 51%, a specificity of 86%, a positive predictive value of 93%, and a positive likelihood ratio of 3.75. Combining EM and FCM as the positive set resulted in a sensitivity of 75%, a specificity of 82%, a positive predictive value of 95%, and a positive likelihood ratio of 4.22. Combining EM, FCM, and NFCM as the positive set resulted in a sensitivity of 98%, a specificity of 52%, a positive predictive value of 87%, and a positive likelihood ratio of 2.04. (See Tables #1 and #2)

Of the semantic types present in the general NFCM, adding the semantic type designated by Type Unique Identifier (TUI) 169, "Functional Concept", to the semantic type filtering added 131 true positives and 31 false positives to the FCM. These changes increased the positive match rate for FCM from 29.4% to 31.9% while decreasing the positive match rate of the NFCM set from 15.8% to 13.2%. For this new set (Case 4 of the Methods section) sensitivity increased from 88.1% to 90.4%, the specificity decreased from 73.7% to 69.6%, the positive predictive value remained virtually unchanged at 98.4% vs. 98.5%, and the positive likelihood ratio decreased from 3.36 to 2.98.

Discussion

Using an automated process we have demonstrated that large local terminology sets can be covered with very high positive predictive values. The use of semantic type specific filtering improved specificity without sacrificing sensitivity thus maintaining a high (greater than 95%) positive predictive value for exact and composed matches, for both the specialty specific (dermatology) and general data sets. We conclude that automated methods for mapping local "colloquial" terminologies to large-scale controlled health vocabulary systems are practical. They result in a high positive predictive value of matching when using exact matches and semantic type filtered composed matches. We did note a significant decrement in mapping capabilities to a very highly specialized local terminology (the dermatology term set) ($p<0.001$). Therefore attempts to integrate a specialized set of terminology will require more human review and possibly more additions to the terminology than a more general set.

Future directions include expanding semantic type analysis by examining the current filtering system for both the general and specialty specific term sets allowing ever more precise filtering and better compositional matches. Failure analysis of all false negatives will be done to determine if true matches were not recognized and, if not, why. User directed composition may allow salvage of many of the false positive and true negative matches thus significantly increasing the incorporation rate for the local terminologies.[8] The true negative terms which do not yield to user directed composition to form a positive match will form a set of local terms which can be considered for incorporation into larger vocabularies without the onus of redundancy.

Given the large size of both the specialty specific and local general terminological corpi utilized in our study, our methods should be generalizable to other local specialty specific and general terminology sets. These results should help to establish a method of developing searchable local lexicons for organization specific purposes and/or incorporating such terminologies into the UMLS without overburdening the system with redundancies or with every organization's list of colloquial terms.

Acknowledgements: The authors would like to thank Kent Bailey, PhD, for help with the statistical analysis; James Buntrock for programming support; and Karen Elias for secretarial and formatting support. This work was partially supported by NIH grants: UO1-LM08751 and R01-LM05416.

References

1. Cimino, JJ. Desiderata for controlled medical vocabularies in the Twenty-First Century. *Meth Inform Med* 1998; 37:394-403.
2. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: Development guidelines, criteria for selection, and public policy implications. *JAMIA* 1998; 5(6):503-10.
3. Cote RA, Robboy S. Progress in medical information management – Systematized Nomenclature of Medicine (SNOMED). *JAMA* 1980; 243:756-62.
4. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Cote RA. Towards representations for medical concepts. *Med Decis Making* 1991; 11: S102-8.
5. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. *JAMIA* 1998;SympSuppl:765-9.
6. Chute CG, Elkin PL. A clinically derived terminology: Qualification to reduction. *JAMIA* 1997; Symp Suppl:570-4.
7. Tuttle MS, Olson NE, Keck KD, et al. Metaphrase: An aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. *Meth Inform Med* 1998;37(4/5):373-83.
8. Elkin PL, Mohr DN, Tuttle MS, et al. Standardized problem list generation, Utilizing the Mayo Canonical vocabulary Embedded within the Unified Medical Language System. *JAMIA* 1997; SympSuppl: 500-4.