Inter-rater Agreement in Physician-coded Problem Lists

Adam S. Rothschild, M.D., Harold P. Lehmann, M.D., Ph.D., George Hripcsak, M.D., M.S.
Department of Biomedical Informatics
Columbia University (ASR, GH), New York, NY
Division of Health Sciences Informatics
Johns Hopkins University School of Medicine (HPL), Baltimore, MD

Abstract

Coded problem lists will be increasingly used for many purposes in healthcare. The usefulness of coded problem lists may be limited by 1) how consistently clinicians enumerate patients' problems and 2) how consistently clinicians choose a given concept from a controlled terminology to represent a given problem. In this study, 10 physicians reviewed the same 5 clinical cases and created a coded problem list for each case using UMLS as a controlled terminology. We assessed inter-rater agreement for coded problem lists by computing the average pair-wise positive specific agreement for each case for all 10 reviewers. We also standardized problems to common terms across reviewers' lists for a given case, adjusting sequentially for synonymy, granularity, and general concept representation. Our results suggest that inter-rater agreement in unstandardized problem lists is moderate at best; standardization improves agreement, but much variability may be attributable to differences in clinicians' style and the inherent fuzziness of medical diagnosis.

Background

Transition to electronic health records (EHRs) will make coded electronic problem lists increasingly available. These problem lists will be used for many purposes including triggering problem-specific decision support, disease data reporting, clinical research, financial purposes, etc.[1, 2] Cross-clinician inconsistency in applying concepts from the controlled terminology may limit coded problem lists' optimal use.

Numerous terminologies have been designed specifically to code problem lists[3-6]. Ideal controlled terminologies avoid both redundancy and ambiguity, are concept-oriented, and have outstanding content coverage of the domain of interest[7]. SNOMED CT has been recommended by the Consolidated Health Informatics group to serve as the problem list terminology for the United States National Health Information Infrastructure (NHII).[8, 9] SNOMED CT was made freely available in early 2004 through the Unified Medical Language System (UMLS). UMLS itself has also been used as a controlled terminology for coding problem lists [10-12]even though it was not designed to be used as a controlled terminology. UMLS' limitations for this purpose include concept redundancy and ambiguity. Nonetheless, investigators have still used UMLS as a terminology in its own right because of its ready availability, ease of use, and extensive concept coverage[13].

When creating a coded problem list, the clinician must first decide what problems he believes that the patient has. He must then select the most relevant concept from the controlled terminology to code each problem. Variation may exist in both of these steps[14]. Two clinicians considering the exact same case might enumerate the patient's problems in conceptually different manners. Even where conceptual overlap does exist between a distinct problem on each of multiple clinicians' problem lists, the clinicians may have chosen to code the problems at different levels of granularity or with different nuances.

Our goal in this study was to quantify the level of inter-rater coded problem list agreement for a given case across multiple physicians. In our analysis we aimed to identify where using SNOMED CT might enable improved problem coding consistency and where it might not.

Methods

This study was conducted as a sub-study of a larger study, which has been previously described in detail[15]. Briefly, ten third-party physicians of various training levels and specialties each reviewed the same 5 inpatient cases from a general internal medicine service at Johns Hopkins Hospital and created coded problem lists using the unlimited (except to English) UMLS 2003 AC release as a controlled terminology. Reviewers could add one of four qualifiers in a post-coordinated fashion, although we actively discouraged use of qualifiers in the study instructions and by visual prompting in the user interface. Qualifier choices included "history of," "status post," "rule out," and "prevention of/prophylaxis."

Table 1. Examples of problem standardizations.

round of standardization	pre-standardized term	post-standardized term
raw → synonym standardized	Reflux, NOS	Gatroesophageal reflux disease
synonym standardized → granularity and synonym standardized	Complex partial seizures	Seizures
granularity and synonym standardized → general concept standardized	Acute coronary syndrome	Chest pain

Reviewers directly performed all study-related tasks using a Web-based application with a relational database backend that we designed especially for this study. Cases consisted of the scanned and de-identified hand-written intern admit note and an edited and standardized list of the patient's orders placed within the first 24 hours after admission. Orders were available from a query of the hospital's legacy computerized provider order entry system. As was required for the parent study, the reviewers familiarized themselves with the case, created a coded problem list, and linked each order with the problem for which it was most indicated. Reviewers were instructed that for the purpose of this study, the problem list would be considered complete if it was able to account for all of the listed orders, although they were permitted to leave an order unassigned to a problem if necessary. Only problems that were linked to at least one order were included in our analysis. Every problem list contained the universal problem "Hospital admission, NOS," which the reviewers were not permitted to delete from the problem list, although they were not required to link any orders to this problem.

The cases used in this study were selected as an exclusive sub-sample of the cases selected for the primary purposes of the parent study. Cases for the parent study were selected based on their ICD-9 admission diagnosis frequency. Cases for this sub-study were chosen by randomly selecting one case from each of the 5 most frequent ICD-9 admission diagnoses in the parent sample of cases.

Analysis was performed for each case by calculating the average pair-wise positive specific agreement (p_{pos}) for reviewer-stated problems (including qualifiers) for all reviewer pairs[16]. p_{pos} is a measure of inter-rater reliability that is appropriate for studies where the universe of terms to choose from is either poorly defined or very large, as in this study. p_{pos} is the average probability of a reviewer picking a term, given that another reviewer picked the same term; its scale and interpretation are similar to that of kappa. p_{pos} of the problem lists for a pair of reviewers was computed as the ratio of two times the number of overlapping problems over the sum of two times the number of overlapping problems and the total number of non-overlapping problems.

In addition to calculating the average pair-wise p_{pos} for the raw problem lists, we also recalculated it after standardizing problems to common terms, where possible, across reviewers' lists for a given case, adjusting sequentially for synonymy, granularity, and general concept representation. Each round of standardization was thus more permissive than the previous, and the final problem lists from each round of standardization became the starting problem lists for the next round. These standardizations were performed manually by one investigator (A.R.) using a custom-designed Web-based application that displayed all reviewers' problem lists for a given case

Table 2. Summary statistics by case for raw problem lists.

case	ICD-9 admission diagnosis	avg. # overlapping problems per reviewer	avg. # non-overlapping problems per reviewer	avg. # problems per reviewer
A	CHEST PAIN NOS	2.4	2.4	4.8
B	SYNCOPE AND COLLAPSE	3.4	1.7	5.1
C	CHEST PAIN NEC	3.4	3.9	7.3
D	CONGESTIVE HEART FAILURE	3.2	3.8	7
E	GASTROINTEST HEMORR NOS	2.0	3.6	5.6
all cases combined		**3.1**	**2.9**	**6**

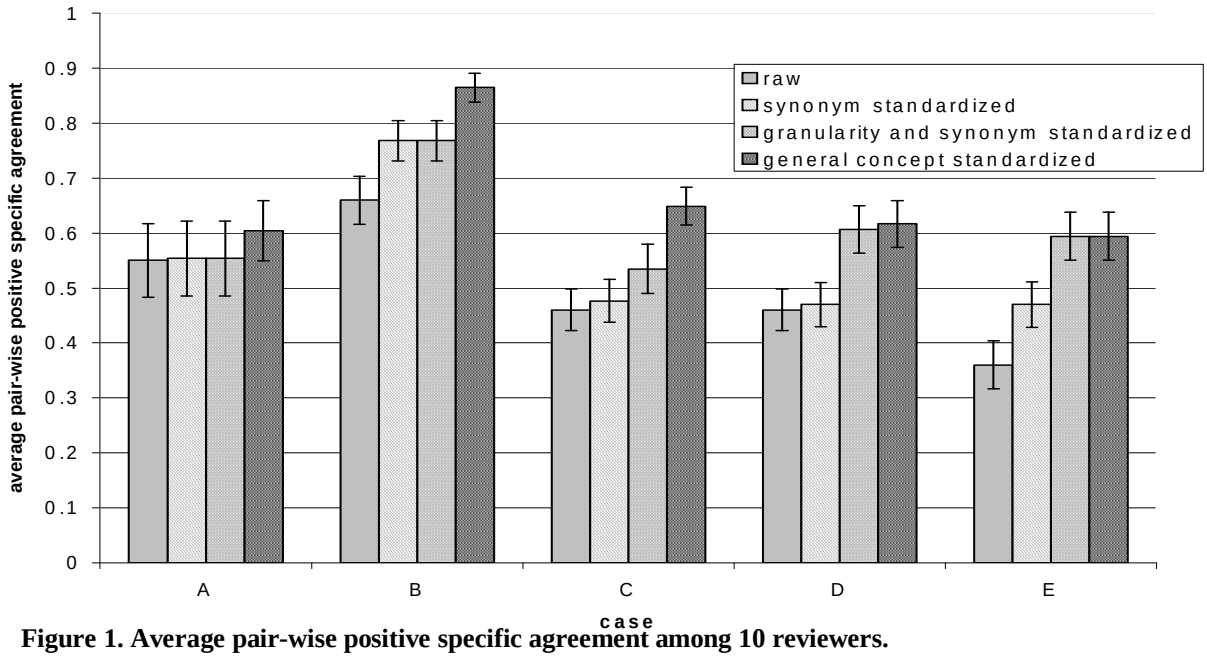

Figure 1. Average pair-wise positive specific agreement among 10 reviewers.

on a single screen and allowed iterative assignment of selected problems across reviewers' lists to a standard term. Table 1 lists examples of the problem standardizations that we performed. The "pre-standardized term" in Table 1 denotes the term before the indicated round of standardization but after any previous rounds of standardization. The "post-standardized term" denotes the term after the indicated round of standardization.

Results

The average pair-wise p_{pos} for raw problem lists by all reviewers across all 5 cases was .50. This means that for each case on average, assuming that the two reviewers had the same number of problems on their problem list, the pair of reviewers had the same number of overlapping problems as non-overlapping problems. An average pair-wise p_{pos} of 1 would indicate that there was complete agreement on problems among all of the reviewers' problem lists; a score of 0 would indicate that there was no agreement on problems among any of the reviewers' problem lists. Across all cases the average pair-wise p_{pos} for synonym standardized, granularity and synonym standardized, and general concept standardized problem lists was .55, .61, and .67, respectively. Of the 298 problems that were entered in this experiment, 297 were coded using UMLS.

Summary statistics are displayed in Table 2. Figure 1 shows the mean pair-wise p_{pos} on a case-by-case basis for both unstandardized (raw) and standardized problems at the three levels of standardization (i.e., synonym standardized, granularity and synonym standardized, and general concept standardized). Individual cases in Figure 1 are identified by capital letters, which reference the corresponding row in Table 2. Error bars in Figure 1 represent 95% confidence intervals.

The data suggest a trend towards improved inter-rater agreement with more permissive standardization, but perfect agreement was not achieved even with the most permissive standardization.

Discussion

Our results suggest that inter-rater agreement in unstandardized, physician-coded problem lists is moderate at best. They further suggest that standardizing synonymous terms may lead to modest improvement in measured agreement, as may additional standardizing by granularity and general concept.

Using UMLS as a controlled terminology may have contributed to this lackluster level of agreement. Its redundant representation of seemingly identical entities as distinct concepts may have led to some decreased performance in the raw problem lists. This is suggested by the modest improvement in performance in the synonym standardized data series in cases B and E in Figure 1. The synonym standardized series approximates what the measured similarity might have been if all reviewers had selected the same problem term where UMLS contained synonyms or near-synonyms as distinct concepts.

UMLS and many of its source terminologies represent entities at varying levels of granularity, and our reviewers sometimes chose to express a given problem at different levels of granularity. The granularity and synonym standardized data series in Figure 1 approximates what the measured similarity might have been if all reviewers had enumerated their problem lists using the same level of granularity where a similar problem existed across multiple reviewers' lists. The improvement in agreement seen from the synonym-standardized data series in cases D and E suggests that granularity differences may negatively affect inter-rater problem list agreement. The general concept standardized data series approximates what the measured similarity would be when taking into account only the highest-level unifying concept represented by the reviewers' enumerated problems. This also led to a small improvement in agreement in cases B and C.

When we conducted this experiment, SNOMED CT was not yet available in UMLS. As we mentioned above, SNOMED CT has recently been recommended to serve as the problem list terminology for the NHII. Although we performed the standardizations in this experiment manually, our results suggest that using SNOMED CT for coding problem lists might encourage and enable improved problem coding consistency. SNOMED CT largely circumvents the concept redundancy problem of UMLS since it is a true terminology and not a unifier of disparate terminologies. In addition, SNOMED CT contains a rich ontology that should allow for automated identification of common ancestors of related concepts. Thus, use of SNOMED CT has the potential to improve performance for problems coded at different levels of granularity and problems that represent different but related concepts.

Our results suggest that standardization of synonymy, granularity, and general concept may achieve some improvement in measured cross-reviewer problem list agreement, yet even under the most forgiving conditions of standardization, overall agreement was still only moderate. From this observation we conclude what might seem self-evident: Different physicians can view the same clinical situation in different ways, leading to different problem lists. This is one root cause of problem list variation that may be difficult for controlled terminologies to address, potential solutions falling instead in the domain of medical education.

Researchers and developers have devoted much work to the role of the problem list in the EHR. A major issue that has emerged is how and when electronic problem lists should be populated. Numerous published reports describe real-time problem list maintenance functionality for EHRs[17-23], but it has been noted that clinicians' maintenance of problem lists has been poor using these real-time tools. Others have reported automatically populating problem lists using post-facto approaches such as natural language processing of discharge summaries[24] and inclusion of coder-assigned ICD-9 diagnoses.

We suggest that the most useful problem lists are accurate, up-to-date, consistently coded, and available at the time of care. Post-facto approaches to problem list maintenance can effectively populate the problem list and may be useful for achieving compliance with JCAHO's "summary list" requirement[25]; however, these approaches raise concern about the trustworthiness of problem lists that are not directly entered by clinicians in real-time. In addition, post-facto approaches do not provide the EHR with knowledge of the patient's most current problems at the time of order entry, when they might be able to trigger decision support rules or otherwise drive the care process.

For data collection in this study, we simulated a "problem-driven" approach using third-party reviewers. The core premise of the problem-driven approach is that the clinician enters the problem list in real-time using a controlled terminology and subsequently performs other clinical information activities (e.g., order entry and assessment and plan documentation) in the context of an explicitly stated problem[15]. By requiring the problem list in order to perform other clinical activities, the problem-driven approach may lead to improved real-time problem list maintenance.

Limitations

This study required 10 physicians to create problem lists for the same 5 cases. This unnatural situation necessarily limited us to a small sample size. It also forced us to rely on third-party reviewers to create problem lists retrospectively rather than having on-duty clinicians create problem lists in real-time. This may have introduced some noise into the data, but this effect is somewhat mitigated by the fact that all reviewers faced the same retrospective conditions. Our manual standardization of problem lists by only a single investigator is also a limitation; performing standardizations by consensus might improve the reproducibility of our results.

Conclusions

Inter-rater agreement in unstandardized problem lists is moderate at best. Standardization improves

agreement, but much variability may be attributable to differences in clinicians' style and the inherent fuzziness of medical diagnosis. Use of SNOMED CT as a problem list terminology has the potential to improve problem coding consistency.

Acknowledgements
A.R. was supported by National Library of Medicine training grants 5T15LM007452 and 5T15LM07079.

References
1. Weed LL. Medical records that guide and teach. New England Journal of Medicine 1968;12:593-600, 652-657.
2. Safran C. Searching for Answers on a Clinical Information-System. Methods of Information in Medicine 1995;34(1-2):79-84.
3. Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, et al. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp 1997:500-4.
4. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Symp 1998:795-9.
5. Brown SH, Miller RA, Camp HN, Guise DA, Walker HK. Empirical derivation of an electronic clinically useful problem statement system. Ann Intern Med 1999;131(2):117-26.
6. Hales JW, Schoeffler KM, Kessler DP. Extracting medical knowledge for a coded problem list vocabulary from the UMLS Knowledge Sources. Proc AMIA Symp 1998:275-9.
7. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
8.http://www.whitehouse.gov/omb/egov/downloads/dxandprob_full_public.doc, accessed 1/5/2005.
9. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699-703.
10. Johnson KB, George EB. The rubber meets the road: integrating the Unified Medical Language System Knowledge Source Server into the computer-based patient record. Proc AMIA Annu Fall Symp 1997:17-21.
11. Elkin PL, Tuttle M, Keck K, Campbell K, Atkin G, Chute CG. The role of compositionality in standardized problem list generation. Medinfo 1998;9 Pt 1:660-4.
12. Goldberg H, Goldsmith D, Law V, Keck K, Tuttle M, Safran C. An evaluation of UMLS as a controlled terminology for the Problem List Toolkit. Medinfo 1998;9 Pt 1:609-12.
13. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc Annu Symp Comput Appl Med Care 1994:201-5.
14. Cimino JJ, Patel VL, Kushniruk AW. Studying the human- computer-terminology interface. J Am Med Inform Assoc 2001;8(2):163-173.
15. Rothschild AS, Lehmann HP. Information Retrieval Performance of Probabilistically Generated, Problem-Specific Computerized Provider Order Entry Pick-Lists: A Pilot Study. J Am Med Inform Assoc 2005;12(3):322-330.
16. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc 2005;12(3):296-298.
17. Campbell JR. Strategies for problem list implementation in a complex clinical enterprise. Proc AMIA Symp 1998:285-9.
18. Miller J, Driscoll C, Kilpatrick S, Quillen E, Jr. Management of prenatal care information: integration of the problem list and clinical comments. Top Health Inf Manage 2003;24(1):42-9.
19. Scherpbier HJ, Abrams RS, Roth DH, Hail JJ. A simple approach to physician entry of patient problem list. Proc Annu Symp Comput Appl Med Care 1994:206-10.
20. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. Am J Manag Care 2002;8(1):37-43.
21. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proc AMIA Symp 1998:280-4.
22. Zelingher J, Rind DM, Caraballo E, Tuttle MS, Olson NE, Safran C. Categorization of free-text problem lists: an effective method of capturing clinical data. Proc Annu Symp Comput Appl Med Care 1995:416-20.
23. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, et al. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. Int J Med Inf 2003;72(1-3):17-28.
24. Cao H, Chiang MF, Cimino JJ, Friedman C, Hripcsak G. Automatic summarization of patient discharge summaries to create problem lists using medical language processing. Medinfo 2004;2004(CD):1540.
25. Joint Commission on Accreditation of Healthcare Organizations. Comprehensive accreditation manual for hospitals : the official handbook. Oakbrook Terrace, Ill.: Joint Commission; 2000.