

# Comparative Analysis of the VA/Kaiser and NLM CORE Problem Subsets: An Empirical Study Based on Problem Frequency

Adam Wright, PhD<sup>1,2</sup>, Joshua Feblowitz, MS<sup>1</sup>,  
Allison B. McCoy, PhD<sup>3</sup>, Dean F. Sittig, PhD<sup>3</sup>

<sup>1</sup>Brigham and Women's Hospital, Boston, MA; <sup>2</sup>Harvard Medical School, Boston, MA;  
<sup>3</sup>University of Texas Health Science Center, Houston, TX

## Abstract

*The problem list is a critical component of the electronic medical record, with implications for clinical care, provider communication, clinical decision support, quality measurement and research. However, many of its benefits depend on the use of coded terminologies. Two standard terminologies (ICD-9 and SNOMED-CT) are available for problem documentation, and two SNOMED-CT subsets (VA/KP and CORE) are available for SNOMED-CT users. We set out to examine these subsets, characterize their overlap and measure their coverage. We applied the subsets to a random sample of 100,000 records from Brigham and Women's Hospital to determine the proportion of problems covered. Though CORE is smaller (5,814 terms vs. 17,761 terms for VA/KP), 94.8% of coded problem entries from BWH were in the CORE subset, while only 84.0% of entries had matches in VA/KP ( $p < 0.001$ ). Though both subsets had reasonable coverage, CORE was superior in our sample, and had fewer clinically significant gaps.*

## Introduction

The problem list, “a compilation of clinically relevant physical and diagnostic concerns, procedures, and psychosocial and cultural issues that may affect the health status and care of patients” (1), is a critical component of the problem-oriented medical record (2). However, many of its benefits depend on the use of coded problem terminologies. Two standard terminologies: International Statistical Classification of Diseases and Related Health Problems version 9 (ICD-9) and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) (ICD-9 and SNOMED-CT) are available for problem list documentation, and two subsets of SNOMED-CT (VA/KP and CORE) are available for SNOMED-CT users. We set out to examine the two subsets, characterize their overlap and objectively measure their coverage with a goal of making a recommendation to clinical system implementers and standards developers.

Our analysis builds on prior work that studied SNOMED-CT in its entirety, and also builds on the internal evaluation work done by the subset developers (3-6). This internal validation looked at the subsets' coverage of the datasets from which they were derived (6) – no previous validation of the subsets on an external dataset has been reported to this point. We provide such a validation on a sample of actual patient data from Brigham and Women's Hospital.

## Background

Problem lists are used for a variety of functions, including direct clinical care of patients, communication between healthcare providers (particularly in situations where one provider is covering for another) as well as clinical decision support (7), quality measurement (8) and clinical research (9).

One critical aspect of many electronic problem lists is structure and coding (1, 3, 10). Most modern (and indeed, most early) clinical information systems support coding of clinical problems using either proprietary code sets, ICD-9 or SNOMED-CT (3). Advantages of coded problems include greater standardization of problem descriptions and definitions, more computability, interoperability and, in many cases, the ability to use existing ontologies for subsumption and related operations to facilitate more efficient development of logic (11, 12).

There is presently some debate over the advantages of SNOMED-CT or ICD-9 as the best standard to represent problem lists, and both are currently acceptable for ONC-approved data exchange (13) and EHR certification (14). However, empirical data suggests that SNOMED-CT provides better coverage and clinical granularity (3), and ICD-9 is currently being phased out as a billing standard in the United States in favor of ICD-10 (15). ICD-10 may have

utility as a problem list vocabulary; however, its use and study in the United States has been limited as it will not be officially adopted for billing in the United States until Oct. 1, 2013.

SNOMED-CT CT is a clinical terminology originally developed by the College of American Pathologists (CAP) (16). The 2011 release of SNOMED-CT contains 393,072 terms across 40 concept types including disorders, findings, situations, procedures, events, morphologic abnormalities and navigational concepts. The terminology is very broad and includes entities ranging from medications to occupations and even physical objects such as tires and rocket fuel (SNOMED-CT code 77132009). In addition to concepts and descriptions, SNOMED-CT also contains numerous relationships between concepts such as “is a” relationships which define a hierarchy as well as association, ingredient and composition relationships such as “has dose form” and “associated finding”.

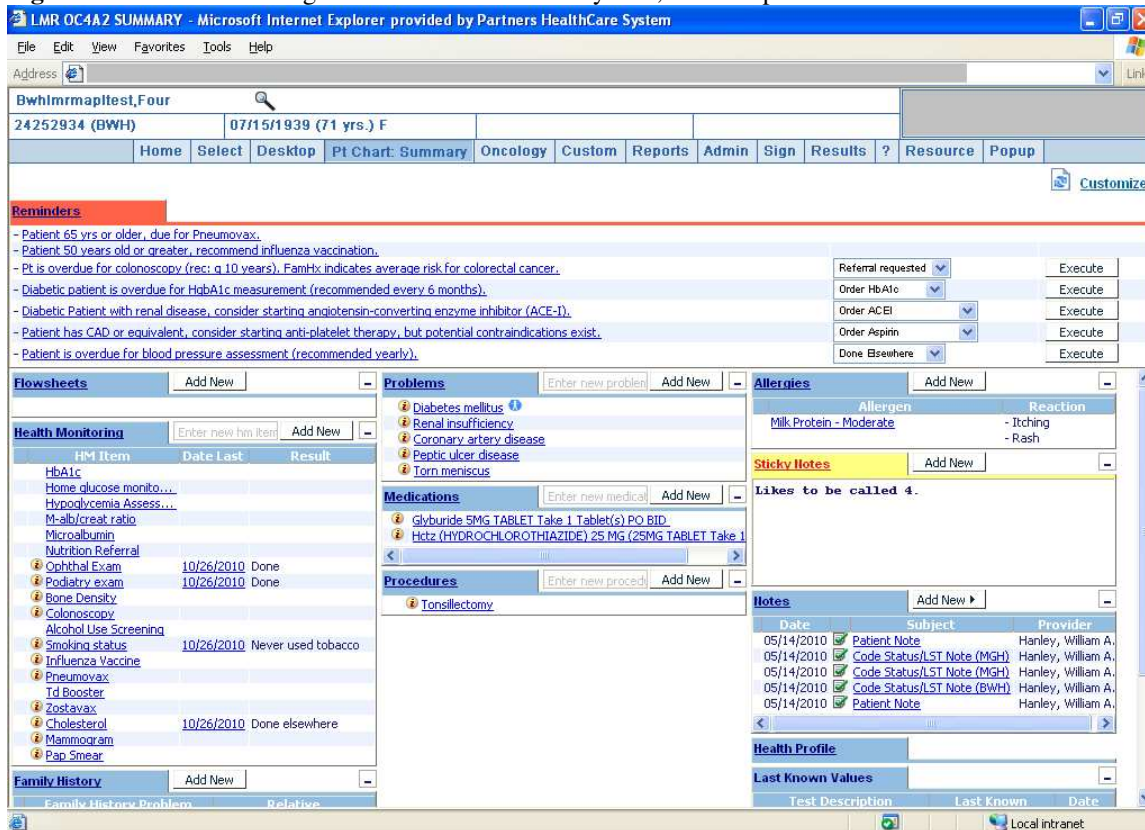
One challenge of using SNOMED-CT for problem lists is its immense size as well as the potential difficulty for end users attempting to locate appropriate concepts. With HIV, for example, there are 138 concepts related to HIV in the current version of SNOMED-CT. Some of these codes refer to HIV infection or seropositivity (and would be relevant to the problem list); however, other concepts describe the HIV virus itself or the fact that HIV status is an observable entity and would thus not be reasonable to include on the problem list, although without appropriate filtering by the EHR application a clinician could easily select one of these by mistake.

Appreciating this issue, most institutions which use SNOMED-CT for their problem list (either directly or via a mapping) use only a subset of SNOMED-CT. Developing and maintaining this subset can be a knowledge management challenge, and when different institutions use different subsets with non-overlapping terms, significant interoperability and decision support definitional challenges result.

In order to address these challenges, *two* standard subsets of SNOMED-CT for use in clinical problem lists are publicly available. The first one was developed by the Veterans Health Administration and Kaiser Permanente based on internal SNOMED-CT-coded problem lists (the VA/KP subset). The VA/KP subset is empirically driven based on requests from users for new problem terms (17) and is used in the United States Food and Drug Administration’s (FDA) Structured Product Label (SPL) initiative – an HL7 approved format for submitting electronic drug label information (18). The subset is available directly from FDA and a UMLS-enhanced version of the subset is available from the National Library of Medicine (NLM). In addition to the VA/KP subset, NLM also makes available the Clinical Observations Recording and Encoding (CORE) subset. CORE is a new initiative from NLM and is a frequency-based approach to problem list development. Seven institutions (Beth Israel Deaconess Medical Center, Intermountain Healthcare, Kaiser Permanente, Mayo Clinic, Nebraska University Medical Center, Regenstrief Institute and the Hong Kong Hospital Authority) (19) submitted usage frequency data for problems to NLM, and NLM identified a set of 14,000 terms which covered 95% of problems in use at each institution. Of those terms, 92% existed in the NLM’s Unified Medical Language System (UMLS), and 81% of the terms in the UMLS had associated SNOMED-CT codes.

The Partners HealthCare system is a large integrated delivery network in Boston, MA, founded by the Brigham and Women’s Hospital and Massachusetts General Hospital. Partners and its member institutions have had a structured, coded problem list in place since the COSTAR system (20), and Partners currently uses a self-developed subset of SNOMED-CT for its problem list terminology in the Partners Longitudinal Medical Record (LMR) system (**Figure 1**). In addition to coded entry, the LMR supports manual free-text entry, though clinicians are encouraged to use a coded problem whenever possible, and receive an alert whenever they add a free-text problem that the problem will not be used for clinical decision support. Use of the problem list is relatively high, and its composition has been described in previously published studies (21-23).

**Figure 1:** The Partners Longitudinal Medical Record system, with the problem list shown in the middle column.



## Methods

We began our analysis by retrieving the January 2011 SNOMED-CT CT International Release (the most recent available) from the NLM. SNOMED-CT CT is available pursuant to a license from the International Health Terminology Standards Development Organisation (IHTSDO), which became the license-holder for SNOMED-CT from CAP in 2007. We retrieved the files in their native format, though they are also available as part of the UMLS release. We also retrieved the NLM's SNOMED-CT CORE release based on the January 2011 SNOMED-CT release, as well as the latest version of the VA/KP problem list subset from the FDA. Finally, we retrieved the Partners HealthCare problem list dictionary from the internal Partners knowledge management system.

In addition to retrieving the terminology files, following IRB approval (Partners Human Subjects Committee Protocol #2008P000233), we also obtained problem list data for a sample of 100,000 randomly selected patients seen at least two times in the outpatient clinics of the Brigham and Women's Hospital. This data included the name of the problem, its SNOMED-CT code, its Partners Problem List (PPL) code (an internal legacy coding system), date added, date of onset, modifiers, comments and text for all free-text entries.

We began our analysis by assessing the concept type for each concept in the VA/KP subset, the CORE subset and the Partners problem dictionary to determine what concept types are used in each subset. We then did a comparison of the terms in each subset and their overlap. This comparison was performed in Microsoft Access by taking the union of all observed SNOMED-CT codes in the three subsets and then determining subset membership for each term, resulting in a Venn diagram. Next, we applied the VA/KP and CORE subsets to the sample of patient data from Brigham and Women's Hospital to assess coverage of the subsets and to determine any clinically significant gaps in either subset.

## Results

### *Descriptive Analysis and Concept Types*

We first reviewed the content of SNOMED-CT CT as well as the three problem subsets. The current release of SNOMED-CT CT contains 393,072 terms, of which 293,768 are active and 99,304 are in other statuses (e.g. limited use, moved elsewhere, outdated, erroneous, etc.). The VA/KP subset of SNOMED-CT contains 17,761 terms, with 16,704 active in the subset and 1,057 retired. Some of the inactive codes were retired because they were retired from SNOMED-CT – others were simply retired from the subset, though they remain active SNOMED-CT codes. The CORE subset is considerably smaller than the VA/KP subset, with 5,814 total terms, including 5,725 active terms and 89 retired terms. The Partners problem dictionary is smaller still, with 1,494 unique terms (the dictionary also contains 68 synonymous terms, where multiple problem names are mapped to the same SNOMED-CT code – these duplicates are not considered for this analysis). None of the Partners terms have been retired, though 14 are suppressed in the EHR’s problem editor or lookup tool, meaning they cannot be added to new patients. In addition, 15 of the terms have no direct SNOMED-CT equivalent (e.g. borderline hypertension or positive HLA-B27 antigen), so they have no SNOMED-CT mapping.

**Table 1** shows the number of terms with each concept type in SNOMED-CT CT itself, the VA/KP and CORE subsets and the Partners subset (dictionary). The table is sorted by number of occurrences in the subsets, and concept types with no terms in any of the three terminologies are not shown. Disorders predominate by frequency, though findings, situations, procedures and events are also common. A number of concept types appear in small numbers, and only in one subset (e.g. physical object in Partners or environment for the VA/KP subset). Notably, the VA/KP subset contains no procedure concepts, though procedures are common in the Partners and CORE subsets.

**Table 1:** Number of terms with each concept type in SNOMED-CT CT and the three subsets.

Concept Type *	SNOMED-CT CT	Partners Subset <sup>^</sup>	CORE Subset	VA/KP Subset
disorder	64,523	996	4,175	15,063
finding	33,015	218	863	1,683
situation	3,239	3	182	886
procedure	49,551	237	466	0
event	3,660	3	57	93
morphologic abnormality	4,372	3	46	1
navigational concept	638	0	2	33
regime/therapy	2,460	3	21	0
person	423	4	2	0
physical object	4,474	5	0	0
observable entity	8,219	4	0	0
environment	1,093	0	0	2
body structure	25,635	1	0	0
qualifier value	8,984	1	0	0
tumor staging	214	1	0	0

\* 25 additional concept types present in SNOMED-CT CT had no terms in any of the subsets, including organism, substance, product and occupation. These types are omitted from this table.

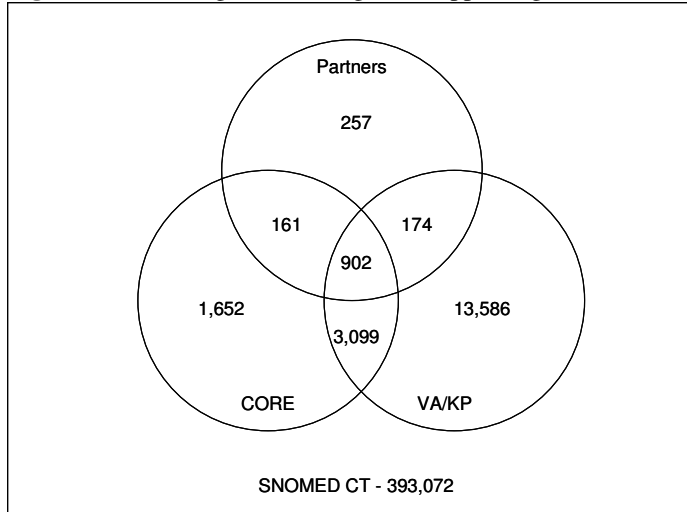
<sup>^</sup> 15 Partners terms do not exist in SNOMED-CT, so no concept type was available.

### *Term Analysis*

After completing the concept type comparison, we proceeded to compare the composition of each of the three subsets with a special focus on overlapping terms. The result of our analysis is shown in **Figure 2**. The diagram shows several important features. First, only 19,831 SNOMED-CT concepts appear in any subset, though there are 393,072 SNOMED-CT terms (for this analysis, both active and retired terms are included because retired terms can still appear in subsets). In other words, using any of the subsets (or even the union of them) eliminates at least 94.95% of SNOMED-CT concepts. Second, looking at the two standard subsets, 68.8% of CORE concepts also

appear in the VA/KP subset, though only 22.5% of VA/KP concepts appear in CORE (in some large part due to VA/KP's larger size). By term count, the two subsets are similar in their coverage of the Partners dictionary – 1,063 Partners problems appear in CORE and 1,076 appear in the VA/KP subset. That said, 257 Partners problem terms do not appear in either set.

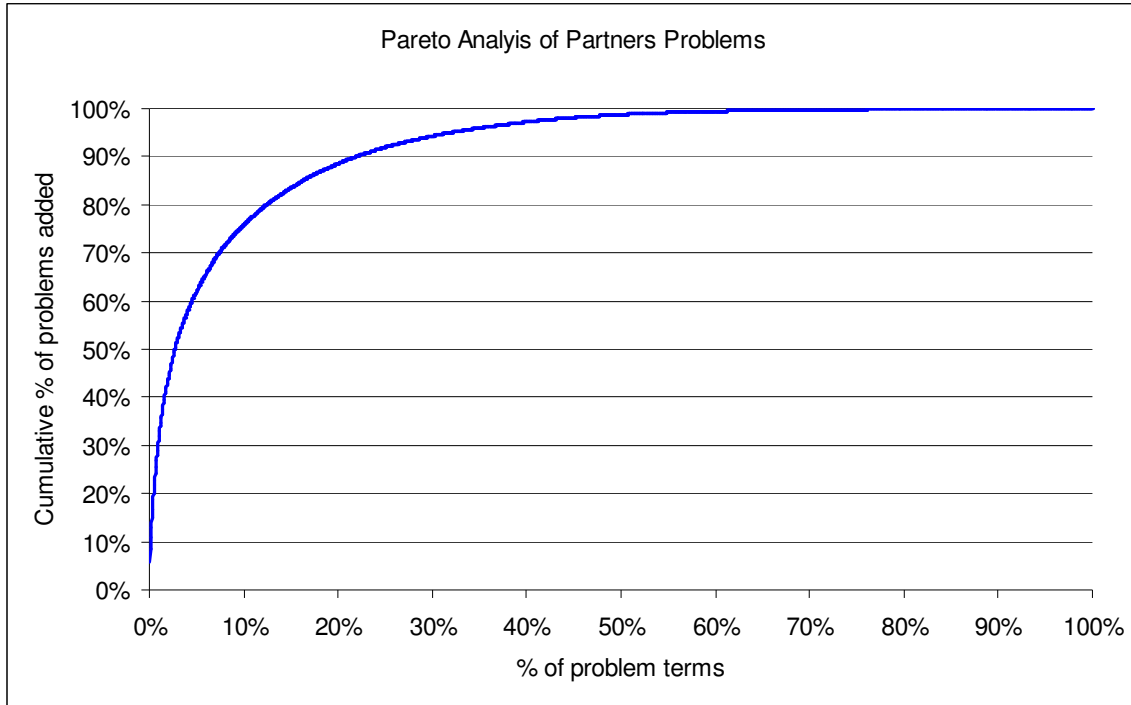
**Figure 2:** Venn diagram showing terms appearing in each subset and overlaps.



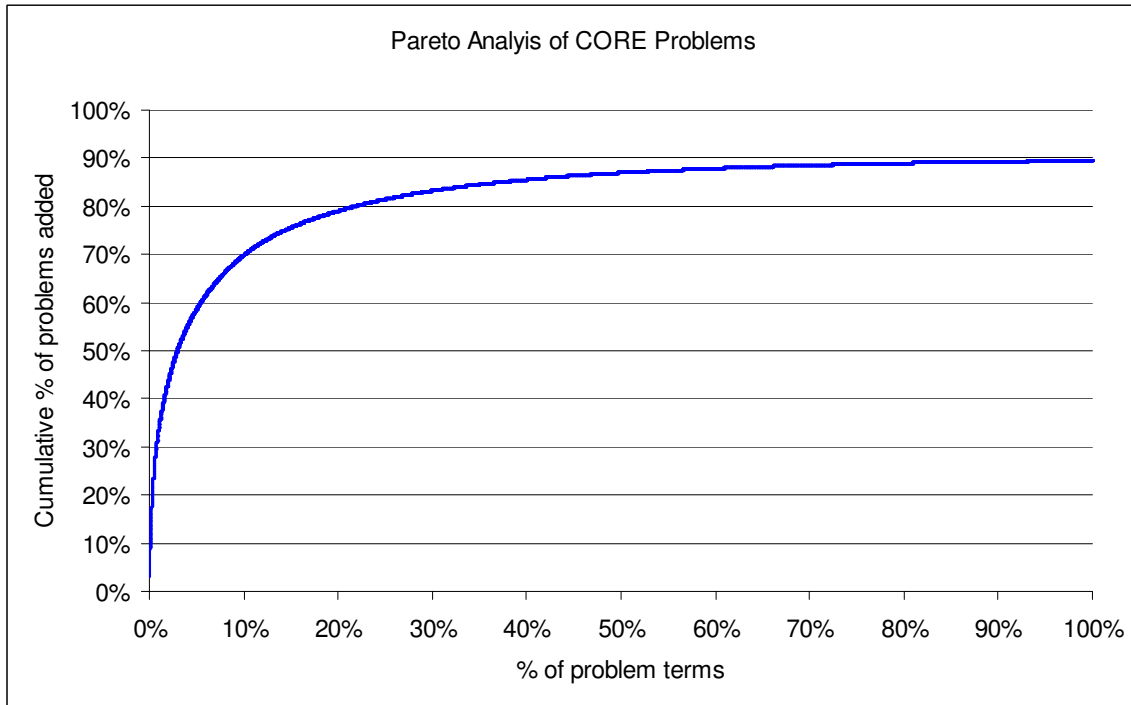
### **Frequency Analysis**

The term analysis, based on counts is instructive; however, it neglects an important property of problems (and other clinical data) – the non-uniformity of their use in clinical practice. In a prior study, we showed that medication, problem and laboratory result data exhibit a power-law distribution and that, in particular, 18.5% of all problem terms in the set account for 80% of all problem list additions (or terms actually used). **Figure 3** shows the cumulative distribution of problems added at Partners, and features rapid ascent but a long tail, suggesting that a small number of problems account for the bulk of usage. This property was the basis for the formation of the CORE subset, which was developed based on a similar frequency analysis of data from the seven CORE sites. **Figure 4** shows the cumulative distribution of usage of the problems in the CORE subset at the seven sites, and exhibits the same power law characteristic (only 89.43% of usage at the sites is covered by the subset).

**Figure 3:** Pareto analysis of problems added in a sample of 100,000 Brigham and Women’s Hospital Patients



**Figure 4:** Pareto Analysis of CORE Problems based on reported usage at 7 sites.



We hypothesized that the skewed nature of problem term usage might influence the coverage afforded by the two subsets, so we re-analyzed the coverage of the VA/KP and CORE subsets, this time informed by Partners frequency data. The Partners’ patient problem dataset contains 370,610 problems entered for a random sample of 100,000 patients. The most common problems are hypertension, elevated cholesterol, depression, coronary artery disease

and asthma. Of the problems, 270,165 are coded and mappable to SNOMED-CT (or in the set of 15 problems with no SNOMED-CT equivalent). An additional 98,861 were entered as free-text by the clinician, and have no underlying code. 2,584 problems (<1%) are coded but could not be matched to the dictionary – this is generally caused by imports from legacy systems. Our frequency analysis focuses on the 270,165 coded, mappable problems.

We found that 94.8% of all problem list entries in the extracted data sample used terms in the CORE subset, while only 84.0% of entries had matches in the VA/KP subset ( $p < 0.001$  for the comparison using both a binomial test and a chi square test). **Figures 5a and 5b** show the most frequently used Partners problems absent from the VA/KP and CORE subsets. All of the top ten problems absent from the VA/KP subset are present in the CORE subset, and five of the ten problems absent from CORE are in the VA/KP subset. Borderline hypertension is one of the 15 Partners problems with no matching SNOMED-CT code, so this problem cannot simply be added to the subset.

**Figure 5a:** 10 most frequent problems in Partners set absent from VA/KP subset

Rank	SNOMED-CT Code	Problem
8	254837009	breast cancer*
21	236886002	hysterectomy*
23	8517006	past smoking*
24	80146002	appendectomy*
26	38102005	cholecystectomy*
31	77543007	tubal ligation*
35	289903006	menopause*
37	173422009	tonsillectomy*
39	11466000	cesarean section*
43	428054006	adenomatous polyp*

\* Appears in CORE subset

**Figure 5b:** 10 most frequent problems in Partners set absent from CORE

Rank	SNOMED-CT Code	Problem
76	34436003	hematuria*
77	397881000	diverticulosis*
82	No code present^	borderline hypertension
104	57993004	benign breast disease*
151	59614000	guaiac positive stool*
155	165816005	HIV positive*
174	127348004	motor vehicle accident
175	129843006	health maintenance
188	313185002	not a problem#
199	14106009	Pacemaker

\* Appears in KP/VA subset, ^ this problem exists in the Partners dictionary, but does not have a SNOMED-CT code, # clinicians can use the “not a problem” problem to add a non-problem entry to the problem list, such as occupation, religion or contact information.

## Discussion

Our frequency-based analysis suggests that there both the VA/KP and CORE subsets have reasonable coverage of our problem list data; however, CORE was ultimately significantly more complete (both statistically and clinically), despite its smaller size.

Some of the gaps in VA/KP merit special discussion. Relatively common problems, including breast cancer, hysterectomy and adenomatous polyp were absent from the subset. In the case of hysterectomy (and the other procedures) this is because the VA/KP subset excludes procedures, though Partners and CORE include them. The case of breast cancer is more complex – the VA/KP subset actually includes many terms for specific types of breast

cancer (e.g. primary malignant neoplasm of areola of male breast or primary malignant neoplasm of central portion of female breast), but no general term. This decision may contribute to increased granularity in coding; however, it represents a significant departure from the CORE and Partners approaches, which used more general coded terms. In the LMR, it is possible for physicians to add additional coded or uncoded detail (e.g. status, laterality, severity and staging); however, this is done using separate fields rather than captured as part of the identifier for the problem.

Given our results, we believe it would be reasonable to standardize on either of the subsets, though CORE appears to perform better, at least on our data, and it has a long-term support and maintenance plan in place, overseen by the NLM. The most widely used version of the HITSP C80 Clinical Document and Message Terminology Component implementation guide specifies the VA/KP problem subset as the value set for a HITSP C32-compliant Continuity of Care Document (CCD) (24). In the most recent release, HITSP expanded the value set to include all SNOMED-CT codes in the Clinical Findings and Situation with Explicit Context hierarchies. HITSP should consider specifically allowing the NLM CORE subset as the C80 problem value set, and might consider ultimately specifying it as the preferred (or even required) value set.

Our study has several limitations. First, we present data from only a single site. However, we are only aware of a few sites with existing large databases of SNOMED-CT-coded problems, and most of those sites participated in the development of the CORE subset, so no other sites were included in our analysis. We considered including sites that used other terminologies (e.g. ICD-9), however, we found that their coverage depended heavily and somewhat arbitrarily on what techniques were used to map the codes from ICD-9 to SNOMED-CT, thus making the results difficult to interpret. We also did not look at inexact matches – in some cases, there are more or less specific concepts in a subset which would map to terms in another subset. We intentionally excluded such matches because granularity is often key to problem list development, and physicians at Partners are encouraged to include exactly as much detail as is known with confidence in their problems, but no more. Finally, we did not look at free-text problem entries. These entries often (though not always) represent gaps in the Partners problem dictionary, and it is possible that some of these problems have matches in SNOMED-CT CT, CORE and/or VA/KP, but they were not matched in our sample. We excluded them because it is difficult to match free-text terms exactly, and they are often ambiguous with multiple codings (23). We encourage other investigators to repeat our analysis on other samples and at other institutions in order to yield further generality.

## **Conclusion**

Both problem list subsets covered a reasonable proportion of problems used at the Brigham and Women's Hospital, though the CORE subset had superior overall performance. Clinical systems implementers and standards bodies should consider adopting CORE as their subset-of-choice for problem list development, and the NLM should continue to develop, improve and maintain the CORE subset.

## **Acknowledgements**

This project was supported in part by Grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects Program (SHARP) from the Office of the National Coordinator for Health Information Technology.

## **References**

1. Bonetti R, Castelli J, Childress JL, et al. Best practices for problem lists in an EHR. *Journal of AHIMA / American Health Information Management Association*. 2008 Jan;79(1):73-7.
2. Weed LL. Medical records that guide and teach. *The New England journal of medicine*. 1968 Mar 21;278(12):652-7 concl.
3. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care*. 1994:201-5.



4. Wasserman H, Wang J. An applied evaluation of SNOMED-CT CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annual Symposium proceedings / AMIA Symposium. 2003:699-703.
5. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED-CT CT: ability of SNOMED-CT clinical terms to represent clinical problem lists. Mayo Clinic proceedings. 2006 Jun;81(6):741-8.
6. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED-CT CT into the UMLS: An exploration of different views of synonymy and quality of editing. 2005. p. 486 - 94.
7. Wright A, Goldberg H, Hongsermeier T, Middleton B. A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):489-96.
8. Quality Data Set Model. 2010 [cited 2011 March 17, 2011]; Available from: <http://www.qualityforum.org/qds/>
9. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. Proceedings / AMIA Annual Symposium. 2002:552-6.
10. Kuperman G, Bates DW. Standardized coding of the medical problem list. J Am Med Inform Assoc. 1994 Sep-Oct;1(5):414-5.
11. Spackman KA, Campbell KE, Cote RA. SNOMED-CT RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997:640-4.
12. Alecu I, Bousquet C, Jaulent M-C. A case report: using SNOMED-CT CT for grouping Adverse Drug Reactions Terms. 2008. p. S4.
13. HITSP Summary Document Using HL7 Continuity of Care Document (CCD) Component: Healthcare Information Technology Standards Panel; 2009 July 8, 2009.
14. Test Method for Health Information Technology. Meaningful Use Test Method 2010 [cited 2011 March 14, 2011]; Available from: [http://healthcare.nist.gov/use\\_testing/index.html](http://healthcare.nist.gov/use_testing/index.html)
15. Preparing for ICD-10. ICD-10-CM/PCS 2011 [cited 2011 March 14, 2011]; Available from: <http://www.ahima.org/icd10/preparing.aspx>
16. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED-CT clinical terms: overview of the development process and project status. Proceedings / AMIA Annual Symposium. 2001:662-6.
17. Mantena S, Schadow G. Evaluation of the VA/KP problem list subset of SNOMED-CT as a clinical terminology for electronic prescription clinical decision support. AMIA Annual Symposium proceedings / AMIA Symposium. 2007:498-502.
18. UMLS Enhanced VA/KP Problem List Subset of SNOMED-CT CT. Unified Medical Language System 2008 [cited 2011 March 14, 2011]; Available from: [http://www.nlm.nih.gov/research/umls/SNOMED-CT/SNOMED-CT\\_problem\\_list.html](http://www.nlm.nih.gov/research/umls/SNOMED-CT/SNOMED-CT_problem_list.html)
19. The CORE Problem List Subset of SNOMED-CT CT. Unified Medical Language System 2011 [cited 2011 March 14, 2011]; Available from: [http://www.nlm.nih.gov/research/umls/SNOMED-CT/core\\_subset.html](http://www.nlm.nih.gov/research/umls/SNOMED-CT/core_subset.html)
20. Barnett GO, Justine NS, Somand ME, et al. COSTAR - A computer-based medical information system for ambulatory care. Proc IEEE. 1979 September 1979;67(9):1226-37.
21. Wright A, Bates DW. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. Appl Clin Inf. 2010;1:32-7.
22. Wright A, Pang J, Feblowitz J, et al. A Method and Knowledge Base for Automated Inference of Patient Problems from Structured Data in an Electronic Medical Record. J Am Med Inform Assoc. 2011 (under review).
23. Wang SJ, Bates DW, Chueh HC, et al. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. International Journal of Medical Informatics. 2003;72(1-3):17-28.
24. C80: Component Definition. 2010 [cited 2011 March 14, 2011]; Section 23311]. Available from: <http://wiki.hitsp.org/docs/C80/C80-3.html>