AN EFFICIENT APPROACH TO THE DEVELOPMENT AND VALIDATION OF A SNOMED CLINICAL TERMS SUBSET FOR USE AS A PROBLEM LIST DICTIONARY

by

Gregory M. Fraser, MD

A CAPSTONE PROJECT

Presented to the Department of Medical Informatics & Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Master of Biomedical Informatics
May 2004

APPROVALS

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Capstone Project of

Gregory M. Fraser

"An Efficient Approach to the Development and Validation of a SNOMED Clinical Terms Subset for Use as a Problem List Dictionary"

has been approved

[redacted]

Thomas R. Yackel, MD, MPH, MS, Advisor in charge of project

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Tom Yackel, whose assistance and support made this project possible. Dr. Kent Spackman was invaluable for his SNOMED expertise. He provided a great deal of assistance with SNOMED materials, strategies, and scripting code. I'd also like to acknowledge Dr. Ken Guappone, whose work laid the foundation for this project.

Thank you also to Andrea Ilg and Dr. Bill Hersh for their support of distance learning in general and myself in particular.

I would not have been able to participate in this program without the support of my colleague, Dr. Matt Degner, who made it possible for me to be away from my medical practice.

I am humbled by and grateful for the love and encouragement of my family, who give meaning and purpose to my life and work.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A problem list is an essential component of an electronic health record (EHR). OHSU is in the process of implementing an electronic ambulatory summary list, consisting of medications, allergies, and a problem list. A coded standard vocabulary, or dictionary, is required to allow implementation of the problem list within the OHSU EHR. *Objective*: To develop and validate a subset of SNOMED CT that can function as a problem list dictionary in this environment, meeting the needs of the clinicians and the limitations of the software. *Methods*: The problem list dictionary was developed in less than two months with the creative use of a variety of SNOMED-related tools, database querying, and inclusion of known common diagnoses. A sample of over 1400 problem list entries was obtained from existing problem lists and compared with the resultant problem list dictionary. *Results*: The problem list dictionary, consisting of over 35,000 unique concepts, representing disorders, surgical procedures, and partial diagnoses, and including cross mapping to ICD-9-CM, was found to match 85.3% of problem list entries and 81.0% of discrete problems. Omissions will be identified by clinician testing and feedback, and necessary additions will be made on an ongoing basis. *Conclusion*: The rapidly-produced problem list dictionary can be expected to perform well when fully implemented. Next steps, continuous improvement, and future research are discussed.

AN EFFICIENT APPROACH TO THE DEVELOPMENT AND VALIDATION OF A SNOMED CLINICAL TERMS SUBSET FOR USE AS A PROBLEM LIST DICTIONARY

Introduction

The idea that a "complete and accurate" problem list is an essential component of the medical record was introduced in the 1960's [1] and has been almost universally adopted since that time. The Institute of Medicine, in its report on the computer-based patient record, reiterated that "an easily reviewed and updated problem list" is an important part of an electronic health record.[2] Additionally, the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) requires that an "Ambulatory Summary List" is "initiated for each patient by the third visit and maintained thereafter".[3]

The electronic health record (HER) at Oregon Health & Science University (OHSU) is the Siemens Lifetime Clinical Record (LCR) [Siemens Medical Solutions, Malvern, PA]. LCR has the capability of keeping an electronic problem list. The medical staff at the outpatient OHSU clinics would like to have an electronic health summary which contains a list of medical conditions, current medications, and allergies. I set out to develop a coded problem list dictionary that clinicians can use to add medical conditions to the problem list component of the health summary.

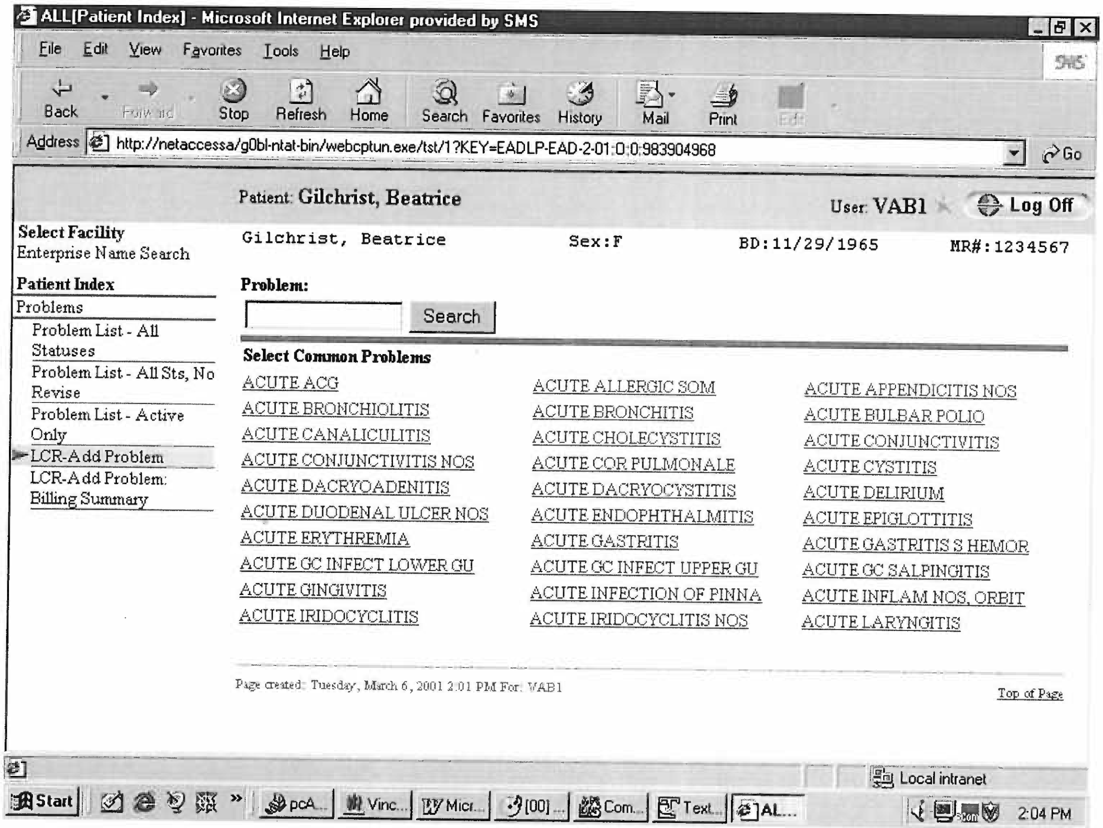
It has been recognized for some time that an electronic problem list requires a standard coding scheme.[4] It is considered to be one of the essential informational attributes of an effective problem list within an electronic health record.[5] A standardized and coded problem list is necessary if the full benefits of an electronic health record are to be achieved,[4] including interoperability, clinical decision support, data mining, and clinical research. The ICD-9-CM codes have been used for many years to capture diagnostic data for billing purposes, but Payne et al. [6] found that for 45% of problem list entries, clinicians were dissatisfied with ICD-9 representation, and concluded that ICD-9 was an inadequate vocabulary for an automated problem list. During the 1990's a number of studies compared available clinical vocabularies. Although none of the vocabularies were found to be ideal, SNOMED was consistently found to do the most complete job of capturing clinical concepts.[7-10] With the latest iteration of SNOMED, SNOMED Clinical Terms (SNOMED CT), the healthcare industry has recognized that it meets nearly all the qualities that characterize an ideal controlled medical terminology (CMT).[11] In November of 2003, the National Committee on Vital and Health Statistics (NCVHS), an advisory panel of the Department of Health and Human Services, recommended that SNOMED CT become "the general core terminology to support the patient medical record information (PMRI)".[12] Moreover, effective January of 2004, SNOMED CT has been licensed by the federal government in order to be freely utilized in the United States.[13] Based upon these circumstances, we chose SNOMED CT to be the basis for the problem list dictionary.

User feedback during informal focus group sessions at OHSU had indicated that availability of ICD-9-CM codes in the problem list would have a significant positive impact on usability. In order to successfully create a problem list dictionary, then, it was necessary to identify a subset of SNOMED CT that: (a) contained descriptions of a comprehensive list of disorders, problems, and procedures, (b) included cross-mapping to relevant ICD-9-CM codes when available, and (c) fit the technical requirements of the EHR software. The subset was then tested for completeness against known common diagnostic codes as well as a sample of existing problem lists.

Materials

LCR

The LCR uses a Common Vocabulary Engine (CVE) to provide a "single, standard source of information about the medical vocabulary used by the care providers".[14] For the purposes of the problem list, it stores and controls the attributes of problem definitions and allows either an industry standard or customized organizational coded terminology to be loaded. Furthermore, it can support mapping to other industry standard coding systems, for example ICD-9-CM.[14]


ALL[Patient Index] - Microsoft Internet Explorer provided by SMS
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address
Go
Log Off
Search
Local intranet
Start
pcA... Vinc... Micr... [00] ... Com... Text... AL...
2:04 PM

Address: http://netaccessa/g0bl-ntat-bin/webcptun.exe/tst/1?KEY=EADLP-EAD-2-01:0:0:983904968

Patient: Gilchrist, Beatrice   User: VAB1

Select Facility
Enterprise Name Search

Patient Index
Problems
- Problem List - All Statuses
- Problem List - All Sts, No Revise
- Problem List - Active Only
- LCR-Add Problem
- LCR-Add Problem: Billing Summary

```
Gilchrist, Beatrice          Sex:F          BD:11/29/1965          MR#:1234567
```

Problem:

Select Common Problems

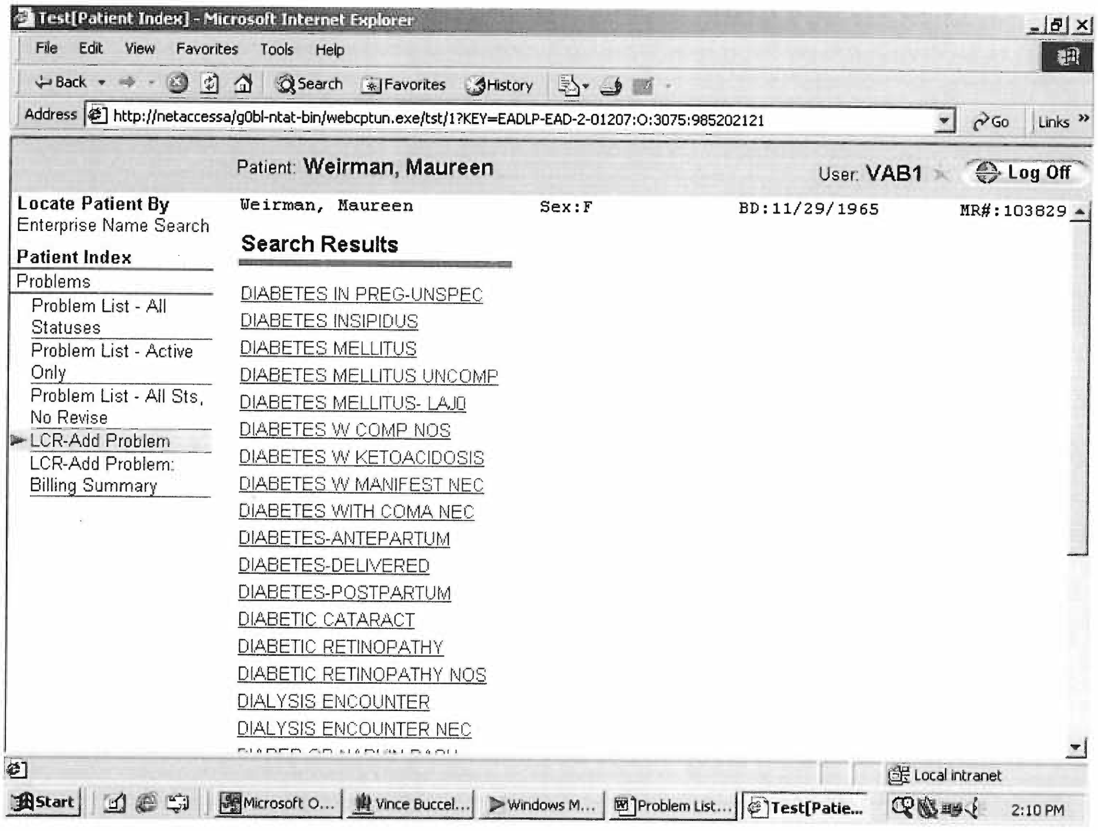
ACUTE ACG
ACUTE ALLERGIC SOM
ACUTE APPENDICITIS NOS
ACUTE BRONCHIOLITIS
ACUTE BRONCHITIS
ACUTE BULBAR POLIO
ACUTE CANALICULITIS
ACUTE CHOLECYSTITIS
ACUTE CONJUNCTIVITIS
ACUTE CONJUNCTIVITIS NOS
ACUTE COR PULMONALE
ACUTE CYSTITIS
ACUTE DACRYOADENITIS
ACUTE DACRYOCYSTITIS
ACUTE DELIRIUM
ACUTE DUODENAL ULCER NOS
ACUTE ENDOPHTHALMITIS
ACUTE EPIGLOTTITIS
ACUTE ERYTHREMIA
ACUTE GASTRITIS
ACUTE GASTRITIS S HEMOR
ACUTE GC INFECT LOWER GU
ACUTE GC INFECT UPPER GU
ACUTE GC SALPINGITIS
ACUTE GINGIVITIS
ACUTE INFECTION OF PINNA
ACUTE INFLAM NOS, ORBIT
ACUTE IRIDOCYCLITIS
ACUTE IRIDOCYCLITIS NOS
ACUTE LARYNGITIS

Page created: Tuesday, March 6, 2001 2:01 PM For: VAB1

Top of Page

Figure 1. Sample LCR Search Page

For each available problem, once loaded, the CVE stores a unique identifier, a textual description, and the coding system used. When the clinician wishes to add a new problem to a patient's problem list, he selects a new problem either by selecting a common problem from a pre-defined list, or by performing a problem search by entering a search string. Figure 1 illustrates a sample "Add Problem" page.[15] The search algorithm performs word completion (i.e. left stem matching) and returns an alphabetical list of search results. It does not, however, rank the results according to relevancy, nor does it capture synonyms or abbreviations. The user can then choose a problem from the search results, or, alternatively, save a problem as free text. Figure 2 illustrates a sample search results page.[15]



Address: http://netaccessa/g0bl-ntat-bin/webcptun.exe/tst/1?KEY=EADLP-EAD-2-01207:O:3075:985202121

Patient: Weirman, Maureen User: VAB1

Locate Patient By
Enterprise Name Search

Patient Index
Problems
Problem List - All Statuses
Problem List - Active Only
Problem List - All Sts, No Revise
LCR-Add Problem
LCR-Add Problem: Billing Summary

Weirman, Maureen Sex:F BD:11/29/1965 MR#:103829

Search Results

DIABETES IN PREG-UNSPEC
DIABETES INSIPIDUS
DIABETES MELLITUS
DIABETES MELLITUS UNCOMP
DIABETES MELLITUS- LAJO
DIABETES W COMP NOS
DIABETES W KETOACIDOSIS
DIABETES W MANIFEST NEC
DIABETES WITH COMA NEC
DIABETES-ANTEPARTUM
DIABETES-DELIVERED
DIABETES-POSTPARTUM
DIABETIC CATARACT
DIABETIC RETINOPATHY
DIABETIC RETINOPATHY NOS
DIALYSIS ENCOUNTER
DIALYSIS ENCOUNTER NEC

Figure 2. Sample Search Results Page

For the purposes of this project, a very important limitation of the CVE is that the problem descriptions are limited to 32 characters. However, the medical staff has indicated that they wish for mapped ICD-9-CM codes to appear along with the search results. This is only achievable if the ICD-9-CM code appears within the description itself; otherwise, only 1 problem per ICD-9-CM code can appear in the vocabulary. Because ICD-9-CM codes can be up to 6 characters (xxx.xx), problem descriptions in the CVE database are effectively limited to 26 characters. Despite this limitation, the linkage to ICD-9-CM codes is felt to be important as it has been shown that "outpatient medical diagnosis coding can be simplified and accuracy improved by using a computerized dictionary of practice-specific diagnoses and synonyms linked to appropriate ICD-9-CM codes."[16]

Table 1. Limitations of LCR

Description field limited to no more than 32 characters
Only one problem per ICD-9-CM code using built-in fields
Search algorithm limitations: ● Left-stem matching only ● List of matches sorted alphabetically, not by relevancy ● No matching of synonyms ● No matching of abbreviations

SNOMED CT

SNOMED CT was first released in early 2002. It is the result of the merging of the UK National Health Service's Clinical Terms Version 3 (CTV3, formerly known as the Read Codes) with SNOMED RT. It has a comprehensive lexicon of clinical terminology arranged in a hierarchical format, that includes, as upper level concepts, not only findings, diseases, and procedures, but also many others as illustrated in Appendix A. Clearly, only some of these are appropriate to consider for use in a problem list dictionary. Table 2 illustrates some pertinent statistics about the release of SNOMED CT used, which was the most recent version available at the beginning of the project.

Table 2. SNOMED CT Statistics

SNOMED CT, July 31, 2003 Release	
Total Descriptions	939,705
Total Unique Concepts	352,661
Preferred Terms	384,830
Active Preferred Terms	295,466
Mean Length of Preferred Term Descriptions	30.9 characters
Active Preferred Term Descriptions > 26 Characters	150,618 (51.0%)
Concepts in "Disease" category	73,171
Concepts in "Finding" category	40,106

An important feature of SNOMED CT is that the coded concept ID's themselves do not carry any meaning. This is by design, as "it is neither appropriate nor useful for code IDs to carry any meaning, classification, or hierarchy."[10] This is because the "classification of the term or the element that is being represented in the code may change over time, rendering any hierarchy in the code itself irrelevant or obsolete."[10]

SNOMED CT is an ontology, containing not only concepts but qualifiers and relationships as well. However, the CVE is capable of storing only a coded ID and description, so it cannot utilize SNOMED CT's complete functionality. Since similar concepts cannot be aggregated together by their concept ID codes, as noted above, the concepts ID's alone do not assist in subset creation. However, by having a coded problem list, problems could be extracted at a later time and the advantages of the hierarchies could be achieved.

The completeness of SNOMED CT has been well documented. McClay and Campbell [17] found that SNOMED CT contained a lexical match or synonym for 93% of text entries of the reason for a visit to an emergency room. Wasserman and Wang established that "a majority of diagnosis/problem list terms (88.4%) were found in SNOMED CT", using only the concepts within the "disease" hierarchy.[18] It would appear that SNOMED CT is the best available CMT on which to base a problem list dictionary, but the entirety of SNOMED CT it is very large and contains a great number of concepts that are not pertinent to a problem list.

Descriptions

SNOMED CT includes a number of descriptions for each concept, including a preferred term, one or more synonyms, and a "fully specified name", which is the preferred term and a tag containing the top-level hierarchical concept to which it belongs.

Cross mapping

SNOMED CT includes a table providing cross mapping to ICD-9-CM codes, in which most ICD-9-CM codes map to multiple SNOMED CT concepts, but most

SNOMED concepts map only to a single ICD-9-CM code. A study of the validity of the provided cross map was not part of this project.

CLUE Browser

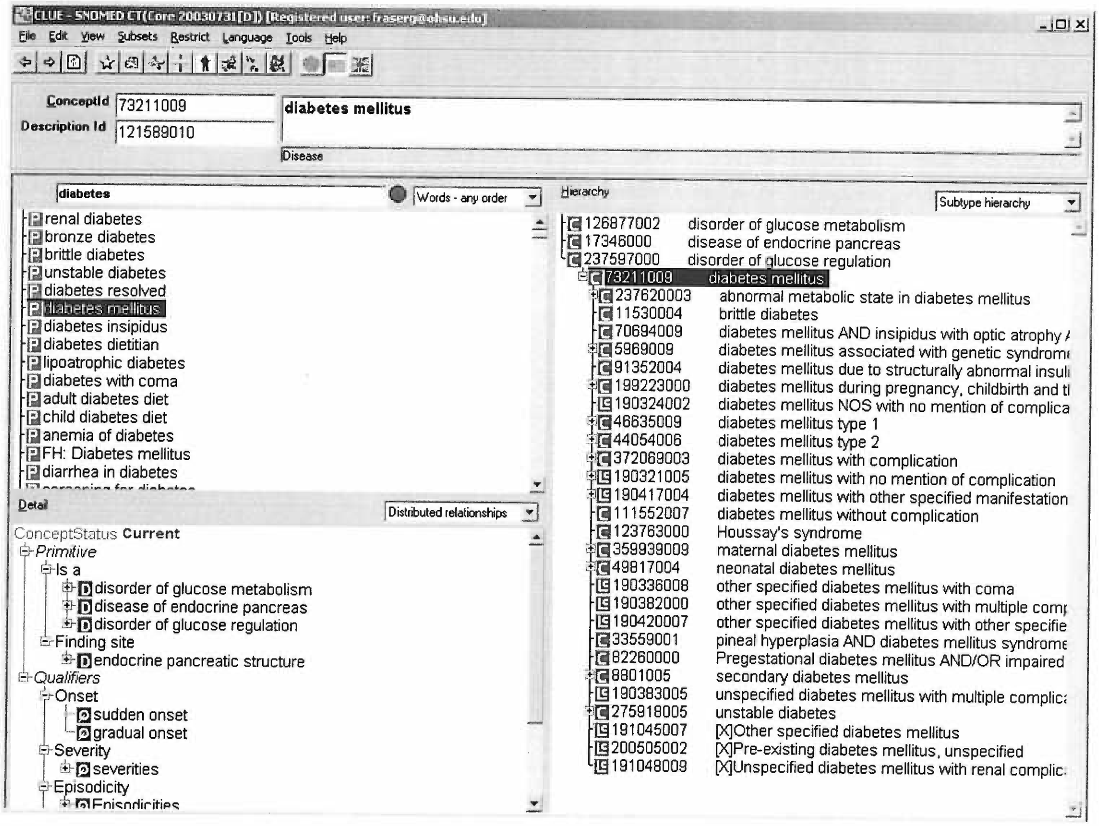
The CLUE Browser is a tool provided with SNOMED CT that allows for lexical searching of SNOMED CT, viewing of the details associated with each concept, and viewing of the hierarchy to which each concept belongs. Concept-specific details can be dragged-and-dropped into a Windows-based word processor, one at a time. Neither SNOMED CT itself nor the CLUE Browser allow for selection of a portion of a given hierarchy into a database or table. Without this functionality, there are no available tools designed to assist in the creation of a customized SNOMED CT subset. A typical CLUE browser screen is illustrated in Figure 3.

UK GP Subset

In the United Kingdom, a subset of SNOMED CT was created to reflect those concepts considered relevant to general practice in that country. Because the OHSU problem list dictionary was to be used in primary care clinics, the author was provided with this subset in the hope that it would provide a useful starting point to identifying concepts for the OHSU problem list dictionary.

Read 2 Codes

Also originating in the UK, a later iteration of the Read 2 Codes (CTV3) was merged with SNOMED RT to form SNOMED CT. The Read 2 Codes, although not updated since 1999, were primary care based but also had two advantages relevant to this project: Each concept had an associated 30-character (or less) description, and the codes themselves were grouped according to meaning. Although this is not normally desirable, it created a means by which selected portions of its hierarchy could be selected.


CLUE - SNOMED CT(Core 20030731[D]) [Registered user: fraserg@ohsu.edu]
File
Edit
View
Subsets
Restrict
Language
Tools
Help

```
ConceptId       73211009     diabetes mellitus
Description Id  121589010
                             Disease

diabetes                        Words - any order
├ renal diabetes
├ bronze diabetes
├ brittle diabetes
├ unstable diabetes
├ diabetes resolved
├ diabetes mellitus
├ diabetes insipidus
├ diabetes dietitian
├ lipoatrophic diabetes
├ diabetes with coma
├ adult diabetes diet
├ child diabetes diet
├ anemia of diabetes
├ FH: Diabetes mellitus
├ diarrhea in diabetes

Detail                          Distributed relationships
ConceptStatus Current
⊟ Primitive
   ⊟ Is a
      ⊞ disorder of glucose metabolism
      ⊞ disease of endocrine pancreas
      ⊞ disorder of glucose regulation
   ⊟ Finding site
      ⊞ endocrine pancreatic structure
⊟ Qualifiers
   ⊟ Onset
      ├ sudden onset
      └ gradual onset
   ⊟ Severity
      ⊞ severities
   ⊟ Episodicity
      ⊞ Episodicities

Hierarchy                                        Subtype hierarchy
├ 126877002     disorder of glucose metabolism
├ 17346000      disease of endocrine pancreas
└ 237597000     disorder of glucose regulation
   ⊟ 73211009      diabetes mellitus
      ⊞ 237620003     abnormal metabolic state in diabetes mellitus
      ├ 11530004      brittle diabetes
      ├ 70694009      diabetes mellitus AND insipidus with optic atrophy A
      ⊞ 5969009       diabetes mellitus associated with genetic syndrome
      ├ 91352004      diabetes mellitus due to structurally abnormal insuli
      ⊞ 199223000     diabetes mellitus during pregnancy, childbirth and tl
      ├ 190324002     diabetes mellitus NOS with no mention of complica
      ⊞ 46635009      diabetes mellitus type 1
      ⊞ 44054006      diabetes mellitus type 2
      ⊞ 372069003     diabetes mellitus with complication
      ⊞ 190321005     diabetes mellitus with no mention of complication
      ⊞ 190417004     diabetes mellitus with other specified manifestation
      ├ 111552007     diabetes mellitus without complication
      ├ 123763000     Houssay's syndrome
      ⊞ 359939009     maternal diabetes mellitus
      ⊞ 49817004      neonatal diabetes mellitus
      ├ 190336008     other specified diabetes mellitus with coma
      ├ 190382000     other specified diabetes mellitus with multiple comp
      ├ 190420007     other specified diabetes mellitus with other specifie
      ├ 33559001      pineal hyperplasia AND diabetes mellitus syndrome
      ├ 82260000      Pregestational diabetes mellitus AND/OR impaired
      ⊞ 8801005       secondary diabetes mellitus
      ├ 190383005     unspecified diabetes mellitus with multiple complic
      ⊞ 275918005     unstable diabetes
      ├ 191045007     [X]Other specified diabetes mellitus
      ├ 200505002     [X]Pre-existing diabetes mellitus, unspecified
      └ 191048009     [X]Unspecified diabetes mellitus with renal complic
```

Figure 3. Screen Capture of CLUE Browser

ICD-9 Codes for Family Practice

The American Academy of Family Physicians (AAFP) has published a list of about 1500 ICD-9-CM codes that are commonly used in family practice.[19] This was obtained in ASCII text format and then converted into a database table for the purpose of querying the eventual problem list dictionary for inclusiveness.

Office and Outpatient Department Visits

Data from the 2001 National Ambulatory Medical Care Survey and the 2001 National Hospital Ambulatory Medical Care Survey were obtained from the Center for

Disease Control (CDC) National Center for Health Statistics (NCHS).[20, 21] These data represent the number and distribution of primary diagnoses by ICD-9-CM code for office and outpatient department visits, respectively.

Existing Problem Lists

A random sample of existing paper-based problem lists was obtained from several OHSU ambulatory care clinics, representing family practice, internal medicine, and pediatrics. All personally identifiable information had been removed. In all, 126 separate paper-based problem lists were obtained, from which 1422 problem entries were retrieved.

Methods

Identification of Subset

Given the comprehensiveness of SNOMED CT and the means by which the LCR CVE would perform searches of and display results from the problem list dictionary, the goal of the project was to identify a usable subset that was as complete as possible. An additional factor is that once a concept ID and description are added to the CVE, they cannot be deleted. Time resources were insufficient to allow a multi-year effort to manually generate a customized problem list lexicon for OHSU, so means to automate the process were sought.

First, the tables containing the aforementioned SNOMED CT concepts, descriptions, and cross maps were imported into Microsoft Access [Microsoft Corporation, Redmond, WA]. Likewise, the UK GP Subset, Read 2 codes, and AAFP common ICD-9-CM codes were loaded into Access. From there, a lengthy series of database queries were undertaken to generate a series of potential subsets. The majority of these candidates were rejected because obvious common concepts were missing, until the following procedure netted what was judged to be the best result.

Disorders

After it became apparent there were significant gaps in concept coverage with any of the three starting points on its own, a subset of the SNOMED-CT diseases hierarchy was obtained by combining terms from the UK GP subset, Read 2 Codes, and AAFP common ICD-9-CM codes, and limiting the results to unique concept ID's representing current active concepts belonging to the upper level "Disease" hierarchy. From the available descriptions for each concept, only the preferred term was selected.

Procedures

Clinicians frequently enter past surgical procedures into the problem list, often using "history of…" or "status post…" descriptors. SNOMED does not provide a significant number of pre-coordinated terms that specifically qualify a past history of a procedure. However, JCAHO specifically requires that "the medical record contains a Summary List of not only known significant conditions but also procedures."[3] It was therefore felt to be important that surgical procedures be available in the problem list dictionary, even without qualifiers. A means was identified to perform a database query on the Read 2 Codes for a subset of the "Procedure" hierarchy consisting solely of surgical procedures, all of which were available with 30-character descriptions.

Symptom Codes

Another frequent occurrence in problem lists is that of nonspecific descriptions, representing symptom complexes or laboratory findings that remain under investigation or which have not yet been diagnosed. It turns out that an inactive group of such concepts was present within SNOMED-CT. These were lexically matched to active concepts and mapped to ICD-9-CM codes when possible. These would be familiar to clinicians as ICD-9 780 and 790 codes.

The resulting subsets of disorders, surgical procedures, and nonspecific codes were appended together to form a subset consisting of 35,942 concepts, of which 10,537 were surgical procedures. The preferred term was utilized in every instance in which it was 30 characters or less; otherwise, the Read 2 30-character description was utilized when available. In order to limit the size of the problem list dictionary, it was decided not to include SNOMED-defined synonyms for each concept.

Spellchecking and correction

Because a great many of the concepts and 30-character descriptions originated in the UK, it became apparent that variances in spelling of medical terminology between the US and the UK would be an issue. For example, "anemia" appeared as "anaemia" and "esophagus" appeared as "oesophagus". Additionally, "fracture" was abbreviated to "#". The latter problem was corrected using a find and replace function, while a US/UK spellchecking script was used to convert the former. The script required a leading space in order to recognize the target word, so most, but not all, instances of British spelling were corrected.

Validation

It is known that the created subset contains at least one matching concept for all 1470 ICD-9-CM codes on the AAFP list of common diagnoses, as its codes were used to help generate the subset. Additionally, at least one matching concept for all of the ICD-9-CM codes that appeared in the table of most common primary diagnoses in the NCHS surveys were confirmed to be present. "Face validation" of the represented concepts revealed few obvious deficiencies, and although this suggested that the proposed subset would perform well when queried by clinicians for a specific problem to add to the problem list, further validation was certainly necessary.

Using the existing problem lists obtained from the outpatient clinics, a listing of problem list entries was made for each of family practice, internal medicine, and pediatrics. A lexical search was carried out by way of a simple text search on the proposed subset. In the absence of a match, a search of SNOMED CT was performed using the CLUE browser. If an exact lexical match was found, it was recorded. If no exact lexical match was present, partial lexical matching was attempted, using what was felt to be the likeliest keyword(s) from the problem list entry and examining matched concepts for meaning. It should be noted that the CLUE Browser does search for synonymy, so synonyms of preferred terms were accepted. When a matching SNOMED CT concept was found, the subset was searched again for the matched concept ID. Problem list entries were then determined to be “present in the subset”, “present in SNOMED CT but not in the subset”, or “not in SNOMED CT”. The concept ID of terms found to be “missing” from the subset were recorded.

Results

Problem List Dictionary

The end result of the project is a problem list dictionary consisting of 35,942 unique concepts. As mentioned, 10,537 of these are surgical procedures, none of which, of course, require an ICD-9-CM code, so a maximum of 32 characters is available for their descriptions. Fortunately, all of these have a 30-character description available because they were derived from the Read 2 Codes.

There are 25,405 diagnosis-related concepts, of which 16,491 have an available description, either a preferred term or Read Code, of 26 characters or less. Nearly all have been mapped to a corresponding ICD-9-CM code. Note that if synonyms had been included, this would have added 26,381 additional descriptions.

Validation

Unlike any of the UKGP Subset, the Read 2 Codes, or the AAFP long list of ICD-9-CM codes on its own, visual inspection of the combined subset did not reveal any glaring deficiencies.

Table 3 illustrates the results of testing the entire problem list dictionary with the existing problem lists obtained from the OHSU outpatient clinics:

Table 3. Lexical Matching of Problem List Entries

Clinic Type	Lists	Entries	Match in Dictionary	% Matched	Terms not in SNOMED CT
Family Practice	44	408	353	86.5%	5
Internal Medicine	65	853	720	84.4	30
Pediatrics	17	161	140	86.9	3
TOTAL	126	1422	1213	85.3	38

When these results are adjusted to eliminate duplicate entries, as shown in Table 4, the success rate in finding a match is slightly different.

Table 4. Matching of Non-Duplicate Problems

Clinic Type	Discrete Problems	Match in Dictionary	% Matching	Terms not in SNOMED CT
Family Practice	304	256	84.2	5
Internal Medicine	541	437	80.8	15
Pediatrics	122	103	84.4	3
COMBINED	857	694	81.0	23

A total of 126 problem lists were examined, containing 1422 entries, of which 857 were non-duplicates. Of the 1422 entries, 1213 (85.3%) were found to have a lexical match within the proposed problem list dictionary. One would expect this to correlate with the likelihood that a match can be found for the next search string entered by a clinician. Of the 857 discrete problems, 694 (81.0%) were matched in the problem list dictionary. This value is lower because common problems are more likely to be duplicated.

Family Practice and Pediatric Problem list entries were matched at a slightly higher percentage than those from Internal Medicine, which may be because Internal Medicine patients were more complex and had a higher incidence of relatively infrequently-occurring problems.

A significant result is that 8914 of the 35,942 concepts in the problem list dictionary do not have an available description that can be implemented automatically, i.e. all available descriptions for them are greater than 26 characters.

Discussion

Based upon what we know about commonly-used diagnostic codes, the proposed problem list dictionary should represent the most common problems encountered in outpatient primary care. This impression is confirmed by its performance in comparing it with existing problem lists. One would estimate that its coverage of new problem list entries would be in excess of 80%. It is the impression of the author, who identified alternative SNOMED CT concepts for most of the terms missing a lexical match, that many of the unmatched terms were quite general. Many times there were several more-specific terms within the subset. One wonders if a clinician would have chosen this more descriptive and specific term if it had been presented to him on a pick list.

Because the CLUE browser was sometimes able make lexical match with a synonym of the preferred term, it was determined that some of the concept ID's initially thought to be missing, were, in fact present. The author did not track how frequently this occurred, but the presence of the appropriate concept ID was counted as a match within the subset, even if the problem list entry did not match the preferred term exactly.

It has proven to be quite valuable for the person working on the subset and validation to have a clinical background, and to have the direct assistance of the Scientific Director of SNOMED International. These may be important factors in the success of the project as it continues.

Overall, the number of available concepts seems quite large, and it would not be surprising if many of the terms are never used. This is unfortunate, as it will be impossible to delete unwanted terms after they are loaded in, although they can be labeled as inactive. The reality, however, is that a considerable effort and time commitment would be necessary to distill the proposed problem list dictionary to a smaller size while still preserving its comprehensiveness.

Others have taken a variety of approaches to generating a problem list dictionary or lexicon, many of which were generated manually as a local vocabulary. Yarnall et. al. [16] started with a dictionary of only 717 diagnoses which had been generated from billing data. Similarly, Zelingher et. al. [22] generated a dictionary of 846 terms. After years of development, Campbell [5] started with a problem list lexicon of over 11,000 terms. Brown et. al. [23] developed a "Canonical Clinical Problem Statement System" (CCPSS) consisting of 15,534 terms, but it took 2000 person-hours over 2 years. Conversely, Wang et. al. [24] derived a "canonical problem list dictionary" of merely 465 terms, but now contains over 840 concepts and 4500 synonyms.

It is only recently that evaluation of SNOMED CT as the problem list dictionary in and of itself has begun. Wasserman and Wang [18] described creating a base lexicon of all concepts and synonyms under the top-level hierarchical concept of "disease". Their dictionary consisted of over 80,000 terms. It would appear that few, if any, have tried to incorporate procedures into their local problem list, as has been done with this project.

The commonest approach to "problem statement vocabulary analysis" is to compare "clinically generated problem statements with existing controlled vocabularies".[23] Results have been variable. Zelingher et. al.'s dictionary matched about 66%.[22] Brown et. al. report that their CCPSS "fully captured" clinical intent 84% of the time.[23] The results in this project appear to compare favorably. Using a proactive rather than retroactive approach, Wasserman and Wang found coverage of

88.4%, which improved to 98.5% after adding additional synonyms.[18] Wang et. al. reported an overall successful coding rate of 69.1%.[24]

The OHSU LCR Problem List implementation is not complete. The proposed problem list dictionary requires further refinement. A priority will be to identify needed concepts among those that do not currently have a description short enough to be used by the CVE. It will probably be wiser to modify the preferred term manually rather than to truncate it, which will involve considerable manpower. Additionally, residual British spelling should be identified and corrected manually. Where concept ID's mapped to more than one ICD-9-CM code, it will have to be decided which one most closely represents the preferred term.

Validation should be repeated, preferably by someone with a clinical background, and perhaps with an even larger selection of existing problem lists. It will be necessary to decide whether to add those terms found to be "missing" from the subset.

Once the problem list dictionary has been loaded into the CVE, it is clear that maintenance and continuous improvement of it will be required on an ongoing basis. Campbell et al. found that they added 800 new terms to their lexicon in the first six months, after which 15 new terms were being added per week, based upon new requests.[5] It will have to be decided whether an individual or team will be responsible for responding to user feedback and making decisions about what additions to make. Another aspect of maintenance is that a new version of SNOMED is released approximately every six months, so changes in the vocabulary may affect the problem list and will have to be evaluated and accounted for.

One important decision to be made in the near future is whether to include SNOMED CT's built-in synonyms for the selected concepts. Wasserman and Wang, in particular, emphasized that their clinicians utilized a synonym equally as frequently as they used the preferred term, and were highly predisposed to search using abbreviations.[18] The inclusion of synonym functionality, then, may improve the CVE search function. It would be simple to accomplish, but would have to be weighed against the cost of an even larger dictionary in terms of the number of choices being presented to the clinician. Also, it is likely that a significant percentage of the synonyms would have descriptions greater than 26 characters in length.

This restriction on the length of the concept descriptions has been an important limitation throughout this project, and is not easily overcome. It was exacerbated by the need to take up valuable description space to include ICD-9-CM codes. Table 1 demonstrates why it is difficult to limit SNOMED CT descriptions to such a small number of characters. The mean length of a preferred term description in SNOMED CT is 30.9 characters, and a majority (51%) of them are greater than 26 characters. This will continue to be a challenge as new concepts are added to the problem list dictionary.

It should be noted that Wang et. al. found that users frequently used free text even though an appropriate coded term was available.[24] They concluded that "user interface failure" was the commonest reason for this occurrence. It will be interesting to see how LCR performs in this regard.

Another important limitation is the lack of built-in subset editing functionality either in SNOMED CT or in the CLUE browser. If this function is not added at some point in the future, then some consideration might be given to simply including all of the concepts and descriptions within a top-level hierarchical concept, such as “disease”, “finding”, and/or “procedure”, in order to provide some degree of standardization between facilities. Future research will likely help to elucidate the best approach.

CONCLUSION

In order to avoid a multi-year effort that others have employed to generate a customized problem list dictionary, a variety of tools were used to create a subset of SNOMED CT that could function as a problem list dictionary within the confines of the OHSU ambulatory electronic health record, Siemens Lifetime Clinical Record (LCR). The problem list dictionary will be accessible to the clinician users by either text string or ICD-9-CM search.

Initial testing suggests that when added to LCR's Common Vocabulary Engine, the subset would be expected to perform well, although some additional work is required before the entirety of the subset can be loaded in. Afterward, a process of user testing, feedback, improvement, and ongoing maintenance will be needed.

It is hoped that production of the problem list dictionary will assist OHSU in its ongoing project of fully implementing Summary List functionality. Ultimately, the electronic health record will contain problem list entries codified using the best available controlled medical vocabulary.

REFERENCES

1. Weed, L.L., *Medical records that guide and teach.* N Engl J Med, 1968. **278**(11): p. 593-600.
2. Institute of Medicine (U.S.). Committee on Improving the Patient Record., et al., *The computer-based patient record : an essential technology for health care*. Rev. ed. 1997, Washington, D.C.: National Academy Press. xx, 234.
3. Joint Commission on Accreditation of Healthcare Organizations., *2003 JCAHO Hospital Accreditation Standards*. 2003, Oakbrook Terrace, IL: Joint Commission Resources.
4. Kuperman, G. and D. Bates, *Standardized coding of the medical problem list.* J Am Med Inform Assoc, 1994. **1**(5): p. 414-415.
5. Campbell, J.R., *Strategies for problem list implementation in a complex clinical enterprise.* Proc AMIA Symp, 1998: p. 285-9.
6. Payne, T.H., G.R. Murphy, and A.A. Salazar, *How well does ICD9 represent phrases used in the medical record problem list? (Abstract).* Proc Annu Symp Comput Appl Med Care, 1992: p. 654-7.
7. Campbell, J.R. and T.H. Payne, *A comparison of four schemes for codification of problem lists (Abstract).* Proc Annu Symp Comput Appl Med Care, 1994: p. 201-5.
8. Chute, C., et al., *The content coverage of clinical classifications. For The Computer- Based Patient Record Institute's Work Group on Codes & Structures.* J Am Med Inform Assoc, 1996. **3**(3): p. 224-233.
9. Campbell, J.R., et al., *Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity.* J Am Med Inform Assoc, 1997. **4**(3): p. 238-250.
10. Chan, W., C. Centiu, and J.A. Morris, *Uniform data standards for capturing patient medical record information at the point of care.* J Healthc Inf Manag, 1999. **13**(3): p. 85-95.
11. Rose, J.S., et al., *Common medical terminology comes of age, Part Two: Current code and terminology sets--strengths and weaknesses.* J Healthc Inf Manag, 2001. **15**(3): p. 319-30.
12. College of American Pathologists. (2003). Federal advisory panel backs use of SNOMED. Retrieved April 3, 2004, from http://www.snomed.com/about/Nov2003.html.
13. Morrissey, J., *A defining moment. Database to help establish universally accepted clinical terminology, easing development of a standardized medical record.* Mod Healthc, 2003. **33**(30): p. 30, 32.
14. Guappone, K., *Implementation of the Siemens LCR web problem list at OHSU hospitals and clinics: A white paper (Unpublished).* 2003, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University School of Medicine.
15. Buccellato, V., *LCR 24 Problem list functional detail and net access set-up.* 2001, Siemens.
16. Yarnall, K.S., et al., *Computer-prompted diagnostic codes.* J Fam Pract, 1995. **40**(3): p. 257-62.

17. McClay, J.C. and J. Campbell, *Improved coding of the primary reason for visit to the emergency department using SNOMED (Abstract).* Proc AMIA Symp, 2002: p. 499-503.
18. Wasserman, H. and J. Wang, *An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List.* Proc AMIA Symp, 2003: p. 699-703.
19. American Academy of Family Physicians. (2003). ICD-9 Codes for Family Practice 2003-2004: The *FPM* Long List. Retrieved January 20, 2004, from http://www.aafp.org/x20096.xml.
20. Centers for Disease Control and Prevention National Center for Health Statistics. (2003). National Ambulatory Medical Care Survey, 2001. Retrieved January 20, 2004, from http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm.
21. Centers for Disease Control and Prevention National Center for Health Statistics. (2003). National Hospital Ambulatory Medical Care Survey, 2001. Retrieved January 20, 2004, from http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm.
22. Zelingher, J., et al., *Categorization of free-text problem lists: an effective method of capturing clinical data (Abstract).* Proc Annu Symp Comput Appl Med Care, 1995: p. 416-20.
23. Brown, S.H., et al., *Empirical derivation of an electronic clinically useful problem statement system.* Ann Intern Med, 1999. **131**(2): p. 117-26.
24. Wang, S.J., et al., *Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool.* Int J Med Inf, 2003. **72**(1-3): p. 17-28.

APPENDIX A: SNOMED CT UPPER LEVEL HIERARCHICAL CONCEPTS

Type	Top Level Hierarchy	Brief Description
Major Hierarchies	Finding	Concepts that result from an assessment or judgment
	Disease or Disorder	Concepts in which there is an explicit or implicit pathological process
	Procedure	Concepts that represent the purposeful activities performed in the provision of health care
Supporting Hierarchies	Observable entity	Concepts that represent potential findings or "things you can ask questions about" e.g. visual acuity
	Body structure	Concepts include normal anatomical concepts and abnormal body structures
	Organism	Represent organisms' etiological significance in diseases
	Substance	Contains biologic substance, body substance, diagnostic substance, dietary substance, toxic substance, industrial substance, etc.
	Pharmaceutical or Biological Product	Represents virtual (clinical) drugs and actual (branded) drugs
	Specimen	Concepts representing entities that are obtained for examination or analysis
	Physical Object	Concepts include natural and man-made objects
	Physical Force	Includes motion, friction, gravity, electricity, magnetism, sound, radiation, thermal forces, humidity, air pressure, etc.
	Event	Concepts that represent occurrences that result in injury and exclude procedures and interventions
	Environment or Geographical Location	Includes types of environment as well as named locations
	Social Context	Contains social conditions and circumstances of importance to medicine
	Context-dependent Category	Added information which changes the meaning of the concept
	Staging and Scales	Contains the name of staging schemes and assessment scales
	Attribute	Concepts that are used to relate two different SNOMED CT concepts, usually definitional characteristics
	Qualifier Value	Contains values for SNOMED CT Attributes
	Special Concept	Concepts that have been set aside from the logical hierarchy of semantic subtypes e.g. inactive

APPENDIX B – DATABASE TABLES AND SQL QUERIES

Tables supplied by SNOMED

concepts	(concept information with added CTV3 in hexadecimal notation – necessary because MS Access cannot distinguish upper- and lower-case characters)
descriptions	(descriptions, description types)
crossmaps	(ICD-9-CM cross mapping)
crossmaptargets	(also for ICD-9-CM cross mapping)
ukgp	(UK GP subset)
termkeys	(Read 2 Codes with added hexadecimal notation)

Table generated by author

icdlong	(AAFP ICD-9-CM codes)

Note: Some actual table names changed for simplicity

Queries

1. Table “ukgpdisorders” created by:

```
SELECT ukgp.MEMBERID, descriptions.TERM, descriptions.DESCRIPTIONTYPE,
descriptions.CONCEPTID
FROM ukgp, descriptions
WHERE ukgp.MEMBERID = descriptions.CONCEPTID
AND descriptions.DESCRIPTIONTYPE = 3
AND Right([descriptions.TERM], 10) = ‘(disorder)’
```

2. Table “UKGPICD9” created by:

```
SELECT ukgpdisorders.CONCEPTID, crossmaptargets.TARGETCODES
FROM ukgpdisorders, crossmaptargets, crossmaps
WHERE ukgpdisorders.CONCEPTID = crossmaps.MAPCONCEPTID
AND crossmaps.MAPTARGETID = crossmapargets.TARGETID
```

3. Table “AAFPICD9” created by:

```
SELECT icdlong.CODE, crossmaps.MAPCONCEPTID
FROM icdlong, crossmaptargets, crossmaps
WHERE icdlong.CODE = crossmaptargets.TARGETID
```

```
AND crossmaptargets.TARGETID = crossmaps.MAPTARGETID
```

4. Table “RCDDisorders” created by:

```
SELECT concepts.CONCEPTID, descriptions.TERM,
descriptions.DESCRIPTIONTYPE
FROM concepts, descriptions, termkeys
WHERE concepts.CTVHEX = termkeys.rcdhex
AND concepts.CONCEPTID = descriptions.CONCEPTID
AND descriptions.DESCRIPTIONTYPE = 3
AND Right([descriptions.TERM], 10) = ‘(disorder)’
```

5. Table “RCDICD9” created by:

```
SELECT RCDDisorders.CONCEPTID, crossmaptargets.TARGETCODES
FROM RCDDisoders, crossmaptargets, crossmaps
WHERE RCDDisorders.CONCEPTID = crossmaps.MAPCONCEPTID
AND crossmaps.MAPTARGETID = crossmaptargets.TARGETID
```

6. Table “DTerms” created by:

```
SELECT CONCEPTID, FULLYSPECIFIEDNAME
FROM concepts
WHERE Left([FULLYSPECIFIEDNAME], 3) = ‘[D]’
```

7. Table “DCode” created by:

```
SELECT DTerms.TERM, descriptions.TERM, descriptions.DESCRIPTIONTYPE,
descriptions.CONCEPTID
FROM DTerms, descriptions
WHERE descriptions.DESCRIPTIONTYPE = 1
AND DTerms.TERM = descriptions.TERM
```

8. Table “DCodeICD9” created by:

```
SELECT DCode.CONCEPTID, crossmaptargets.TARGETCODES
FROM DCode, crossmaptargets, crossmaps
WHERE DCode.CONCEPTID = crossmaps.MAPCONCEPTID
AND crossmaps.MAPTARGETID = crossmaptargets.TARGETID
```

9. Table “RCDProcedures” created by:

```
SELECT TERM-30, READ_CODE, rcdhex
FROM termkeys
WHERE READ_CODE >= ‘7%’
AND READ_CODE < ‘8%’
```

10. Table “Procedures” created by:

```
SELECT concepts.CONCEPTID, RCDProcedures.TERM_30
FROM concepts, RCDProcedures
WHERE concepts.CTVHEX = RCDProcedures.rcdhex
```

11. Appended UKGPICD9, AAFPICD9, RCDICD9, DCodeICD9, Procedures, with distinct CONCEPTID's only, to create table CID-ICD9

12. Table “CID-TERM30” created by:

```
SELECT CID-ICD9.CONCEPTID, descriptions.TERM, termkeys.TERM_30, CID-ICD9.CODE, descriptions.DESCRIPTIONTYPE
FROM CID-ICD9, descriptions, concepts, termkeys
WHERE CID-ICD9.CONCEPTID = descriptions.CONCEPTID
AND descriptions.CONCEPTID = concepts.CONCEPTID
AND concepts.CTVHEX = termkeys.rcdhex
AND descriptions.DESCRIPTIONTYPE = 1
```

13. Table “active” created by:

```
SELECT concepts.CONCEPTID, CID-TERM30.TERM_30, CID-TERM30.CODE, CID.TERM30.TERM, concepts.CONCEPTSTATUS
FROM concepts, CID-TERM30
WHERE concepts.CONCEPTSTATUS = 0
```

14. “active” updated by:

Spell-check script
Manual addition of field for NAHCS common ICD-9-CM codes
Updated TERM_30 with TERM if TERM<31 characters

Result: Proposed problem list dictionary

15. Selection of concepts with descriptions <= 26 characters:

```
SELECT CONCEPTID, TERM_30, CODE, TERM, CONCEPTSTATUS,
COMMONICD9
FROM active
WHERE Len([TERM_30]) < 27
```