OREGON HEALTH & SCIENCE UNIVERSITY
SCHOOL OF MEDICINE – GRADUATE STUDIES

CAN THE USE OF ELECTRONIC TOOLS IMPROVE THE CODING OF PATIENT PROBLEM LISTS TO THE SNOMED CT TERMINOLOGY?

By

Jeffery Emch

A CAPSTONE PROJECT

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirement of the degree of

Master of Biomedical Informatics

December 2010

OREGON HEALTH & SCIENCE UNIVERSITY
SCHOOL OF MEDICINE – GRADUATE STUDIES

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's capstone project of

Jeffery Emch

Has been approved

Advisor: Judith R. Logan, MD, MS

Table of Contents

List of Tables

List of Figures

Acknowledgements

This project and paper would not have been possible without the unwavering support and guidance of Dr. Judith R. Logan, my project mentor and capstone advisor. Besides providing wide ranging technical advice and editorial guidance, Dr. Logan also provided a problem list dataset from the CORI GI endoscopic reporting system. My deepest thanks to Dr. Logan.

Thanks also to Dr. Francis Lau and Dennis Lee of the University of Victoria for graciously providing the problem list dataset associated with their original research and to Harry Solomon of G.E. Healthcare who also provided a problem list dataset.

Abstract

This paper describes a capstone project that investigated automated methods to assist the mapping of problem list terms to the SNOMED CT terminology. Building on the previous work of Dr. Francis Lau, a database schema and associated matching algorithms were developed and tested. A methodology of recommended matching steps was defined and tested against three different problem list datasets. Associated results and metrics were recorded and areas for improvement and future research were identified. The tools and methodology developed during the project were found to be effective in the mapping of diverse problem list datasets to the SNOMED CT terminology.

Introduction

New U.S. government regulations include a set of “carrots and sticks” that incentivize eligible professionals and hospitals to implement electronic health records. To be eligible for the incentives, certain criteria deemed “meaningful use” must be met. One criterion is the use of either the ICD-9-CM or SNOMED CT medical terminology for the coding of a patient’s current diagnoses, or problem lists, within the electronic health record (EHR).

Automated mapping approaches can be very valuable in the establishment of local vocabularies for use in EHRs. These local vocabularies can be used to populate drop-down menus or pick-lists in the EHRs and their presence can aid in the implementation and utilization of EHRs. The goal of this project is to develop and test automated approaches to improve the mapping of practice problem lists onto the prescribed terminologies for use in the establishment of local vocabularies.

This project investigates various matching techniques to develop a recommended set of steps (a methodology) that can be employed in problem list term matching. A database model and associated match algorithms were developed and tested. The methodology and algorithms were evaluated against three distinct problem list datasets and the results evaluated. A tool to facilitate the recommended mapping methodology was then architected, including a database schema, a set of interface use cases and mock-ups of key interfaces. Finally, results and recommendations were synthesized and reported along with recommendations for future research.

Background

Problem lists contain key information in the medical record. Patients can present with a wide range of problems, particularly in general practice settings. The information recorded as problems can be quite diverse, ranging from patient-expressed symptoms, through personal or family history data, to clinician diagnoses. Specific diagnoses may be represented in multiple ways; for example, the terms “myocardial infarction” and “heart attack” can be used for the same condition. Where problem list data is entered as free text, misspellings can occur and acronyms may be interspersed with descriptive text. Even when an EHR is in use, practices with multiple clinicians will likely have similar problems represented in different ways. All of these issues call for a set of tools to help standardize the problem list terms. Legislation is also emerging to require problem list coding to specific medical terminologies.

The United States Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 includes a set of “carrots and sticks” that will incentivize eligible professionals and eligible hospitals to implement EHRs. The Office of the National Coordinator for Health IT (ONC) in late December 2009 released its proposed criteria for “meaningful use” that will guide the methods of incentives and the requirements for EHR implementations. A key document[1] establishes that one criterion is the use of either the International Classification of Diseases v9 with Clinical Modifications (ICD-9-CM)[2] or the Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT)[3] terminologies for the coding of a patient’s current diagnoses, or problem lists within the EHR.

New EHR adopters, and likely existing EHR users, will need assistance in both standardizing their problem lists and in mapping the standardized lists onto the prescribed terminologies. Automated mapping approaches should reduce the burden of manually determining an appropriate SNOMED CT concept for every problem list term and aid in the establishment of standardized local vocabularies for use in an EHR.

The local vocabularies can be used to populate drop-down menus or pick-lists in the EHR[4]. The presence of these previously validated selections can help minimize redundancy and duplication of problem list entries. The mapping of the problem list terms to standard terminologies significantly improves the ability to utilize the data in downstream analysis and quality measurements.

SNOMED CT

SNOMED CT is described as "a controlled medical terminology with comprehensive coverage of diseases, clinical findings, etiologies, and procedures and outcomes used by physicians, veterinarians, and others"[11]. SNOMED CT was first published as the Systematized Nomenclature of Pathology (SNOP) in 1965 by the College of American Pathologists (CAP) as a tool for organizing data from pathology reports. There have been multiple editions published since then with the most recent released in January 2010. The Systematized Nomenclature of Medicine (SNOMED) first appeared in 1974 and has evolved to its current incarnation through ongoing enhancements and incorporations of other terminologies.

The original SNOP contained four axes[12]. An "axis" is a characteristic that helps define a disease. The SNOP axes were:

- *Topography:*

  The part of the body affected by the disease.

- *Morphology*

  A structural change in tissue.

- *Etiology*

  The cause of the disease or injury

- *Function*

  Physiological or chemical disorders and alterations resulting from a disease or injury.

These axes reflect the original background in pathology. The first release of SNOMED incorporated two additional axes of “Disease”, which organized the original four axes, and “Procedures”. SNOMED Version II was released in 1979 and incorporated a seventh axis called “Occupation” that was derived from the World Health Organization (WHO) International Labour Office. The “Morphology” axis was also expanded to incorporate the International Classification of Diseases – Oncology (ICD-O). By this release the terminology contained over 44,000 records.

SNOMED 3.0 was released in 1993 and was also called SNOMED International. In this version, the “Etiology” axis was divided into 4 axes and a “General Linkage Modifier” axis was added to help link information together across the other axes. This growth of the axes across the releases is shown in Table 1[12].

SNOP	**Original SNOMED**	**SNOMED II**	**SNOMED 3.0**
Topography	Topography	Topography	1. Topography
Morphology	Morphology	Morphology	2. Morphology
Etiology	Etiology	Etiology	3. Living Organisms 4. Chemicals 5. Physical Agents 6. Social Context
Function	Function	Function	7. Function
	Disease	Disease	8. Disease
	Procedures	Procedures	9. Procedures
		Occupation	10. Occupation
			11. General Linkage Modifiers

Table 1: SNOMED Axes growth across early releases

During the 1990's new versions were released with growth in the total number of terms and incorporation of other terminologies. Version 3.4 in 1997 incorporated mappings to the Logical Observation Identifiers, Names and Codes (LOINC) and mappings to ICD-9-CM. In 1998 Version 3.5 was released which added many terms to the disease axis. Total terms had now grown to over 155,000.

Despite its various axes and richness of terms, SNOMED at that time had not been widely adopted. One problem was that its breadth allowed multiple ways to express a single concept. Work began in the late 1990's to address specific issues. CAP began working with a team of physicians and nurses from Kaiser Permanente, an integrated managed care organization in the U.S. The goals were to develop ways to link data together such that the system could recognize equivalent representations of the same concept. The release that resulted also incorporated content from the Digital Imaging and Communications in Medicine (DICOM) community. This led to the release in 2000 of SNOMED Reference Terminology (RT).

Concurrent with the SNOMED changes, work was being performed in the United Kingdom (UK) on developing computer systems for use in general medical practice. Part of this effort entailed designing a set of coding schemes. Dr. James Read, a general practitioner (GP) in England, developed a set of codes for use by GPs that came to be known as the Read Codes. The codes were position-dependent and formed a strict hierarchy. They were designed primarily for use by GPs in their surgery, and not for epidemiology or international comparisons[6]. The U.K. National Health Service (NHS) purchased the Read Codes in 1990. The NHS then formed a Clinical Terms project in 1992 to expand the usage of the Read Codes. This led to the development and release of Clinical Terms Version 3 (CTV3) in 1995 and its widespread use in the U.K.

A merging of SNOMED RT and CTV3 efforts resulted in the release of SNOMED CT in 2002. The National Library of Medicine (NLM), on behalf of the U.S. Department of Health and Human Services, negotiated an agreement with CAP for a perpetual license for SNOMED CT and ongoing updates. Ownership of the SNOMED CT intellectual property passed from CAP to the International Health Terminology Standards Development Organization (IHTSDO) in 2007. The NLM is the U.S. member of IHTSDO and distributes SNOMED CT within the U.S. at no cost.

In 1997, James Cimino published a landmark paper, *Desiderata for Controlled Medical Vocabularies in the Twenty-First Century*[13]. Cimino's paper identified twelve desiderata that he considered essential to a properly designed medical vocabulary. The word *desiderata* is the plural form of desideratum, which can mean "something that is wished for, or considered desirable" [14]. In this case, it can be considered a "wish list" of goals or desirable traits for modern medical vocabularies. SNOMED CT was designed and implemented to align with Cimino's goals. The twelve desiderata, with some sub-points from a 1997 Cimino presentation[14] are shown in Table 2.

Desiderata	**Sub-points**
1. Concept Completeness	• Must seek to provide breadth and depth • Atoms versus molecules
2. Concept Orientation	• Concepts, not terms • One meaning (non-vague) • No more than one meaning (non-ambiguous)
3. Concept Permanence	• Old concepts can't be deleted
4. Identifiers without embedded semantics	• Don't use the name • Don't use a code that will run out of room
5. Polyhierarchy	• Need for tree walking
6. Formal definitions	• Support understanding and maintenance • Structured and controlled (not narrative) • Represented through relationships within the vocabulary
7. Reject NEC (not elsewhere classified)	• Can't have a formal definition
8. Multiple granularities	• Different levels for different purposes
9. Multiple consistent views	• Multiple views for multiple purposes
10. Representing Context	• Needed: a grammar to show usage
11. Graceful evolution	• Will always need to fix mistakes • Medical knowledge will grow
12. Recognize redundancy	• Synonyms are good • Redundant concepts are bad

Table 2. Desiderata and sub-points[14]

SNOMED CT development tried to embody these goals in its structure. It is built with *concepts* with one or more *descriptions* organized and linked by *relationships*. Because they are so central these phrases are defined in an IHTSDO SNOMED CT overview document[15]. The richness of this approach has allowed SNOMED CT to grow with medical knowledge. As of 2008, it contained over 283,000 active concept codes, 732,000 active descriptions and 923,000 relationships.

Due to the specificity and granularity of SNOMED CT, it sometimes is necessary to use multiple concepts to fully express a clinical event. This is called “post-coordination”. An example is the problem list concept “Severe Asthma”. The problem list term would be associated with the SNOMED CT concept of “Asthma”. The concept “severe” would be added, or post-coordinated, to represent the sub-type condition of severe asthma. This manner of sub-typing is very powerful and allows very refined nuances of expression. Because of the detail and richness of SNOMED CT it was chosen as the target vocabulary for this project.

Previous Work

Although significant research[10,11] has been done on the SNOMED CT and ICD- 9-CM terminologies, research on automated problem list coding approaches is somewhat sparse. Key previous works includes the 2004 capstone project[8] of Dr. Greg Fraser, a DMICE professor and alumni, and a 2008 paper[9] by Dr. Francis Lau, et al.

Dr. Fraser developed a consolidated problem list dictionary of over 35,000 records derived from SNOMED CT and ICD-9-CM. A random sample of 1,422 problem entries was extracted from several paper-based OHSU ambulatory care clinics. The problem entries were then mapped to the proposed dictionary using various database queries and lookups using the CLUE browser. Successful mappings were found for 85.3% of the problem list entries and 81% of the discrete problems.

Dr. Lau’s paper explored methods for the encoding of problem lists using SNOMED CT and a series of matching algorithms. The problem list data was derived from a commercial EHR system used by a general practice and contains several thousand records. A methodology of mapping techniques was developed and iteratively applied to the problem list data by Lau and associates. Successfully matched problem list terms were eliminated with each iteration and unmatched terms reviewed for spelling, acronym expansion/elimination or other changes. Spellings were corrected, acronyms expanded and other corrections made prior to the next match iteration. This iterative process resulted in matching success of greater than 90%. Dr. Lau’s approach formed the foundation for this project.

The goal of this project was to further develop and test automated approaches to improve the mapping of problem lists onto the SNOMED CT terminology. The project began by repeating the prior work by Dr. Lau. This included investigating the various matching techniques to develop a recommended set of steps (a methodology) to employ in problem list term matching. A database model and associated matching algorithms were then developed and tested. The methodology and algorithms were tested against the original Lau dataset and a two other problem list datasets. The results were compared and evaluated. A tool to facilitate the recommended mapping methodology was designed including a database schema, a set of interface use cases and mock-ups of key interfaces.

Methods

Problem List Datasets

Dr. Lau graciously provided his original problem list dataset. The problem list data was extracted from a commercial EMR used by a general practice. That practice has four general practitioners working in a township of 100,000 population located east of Vancouver, British Columbia. This dataset included 7,833 terms in which there were 1,822 unique terms.

Harry Solomon, a staff member of GE Healthcare and an OHSU guest lecturer, provided the second problem list dataset. The data represents a broad set of diagnosis and findings primarily to be used in the evaluation and testing of mapping techniques. This dataset has 2,257 terms with 2,121 unique terms.

Dr. Judith R. Logan, of Oregon Health & Science University, provided the third set of problem list data. This data was extracted from the Clinical Outcomes and Research Initiative (CORI) National Endoscopic Database. Data on gastrointestinal (GI) endoscopic procedures is collected during clinical care using software developed by CORI. The CORI software is used at multiple sites across the US for documentation of these procedures. The dataset included over 1.5 million terms but only 268 unique terms. Note that only the discretely collected diagnoses were used and not the diagnosis from free text.

All datasets were provided as Microsoft Excel worksheets. A column with a unique local identifier was added to each spreadsheet to facilitate tracking. Due to Excel's pattern of surrounding commas with quotes on export, macros were built to export the data with a | separator. The export files were then loaded into an Oracle 10g database using Oracle's SQL*Loader. After loading, the data was changed to all lower case characters to remove case comparison issues and all terms were processed to remove multiple consecutive spaces.

SNOMED CT Data

The NLM maintains a terminology system named the Universal Medical Language System (UMLS) "that integrates and distributes key terminology"[5]. The system defines interrelationships between text strings and many terminologies including SNOMED CT and ICD-9-CM. For this project a SNOMED CT "subset" was extracted from the May 2010, 2010AA UMLS distribution and loaded into the project database. The SNOMED CT strings, words, concepts, and other data could then be referenced from within the database.

Automated Matching Methods

Multiple database schemas to support the matching methods were evaluated using an Agile methodology. The final schema is shown in Appendix 1**.** The matching algorithms were modeled on Lau's descriptions. All matching was done using lowercase strings. Since problem lists may contain repeated terms, matching was performed on the set of unique terms derived from each problem list. Two algorithms were used to find matching SNOMED CT concept strings, an "exact-match" algorithm and a "match-all" algorithm. These were implemented in Oracle stored procedures. A third algorithm, "partial-match", was investigated. This algorithm would find a match if any of the words in a problem list terms was present in a SNOMED CT string. This approach yielded so many matches as to be unusable. The algorithms are summarized in Table 3 and detailed at the database level in Appendix 2.

Match Method	Description
Exact match terms (Exact-match)	All words in the problem list term are present in the SNOMED string and the words are in the same order.
Match all terms (Match-all)	All words in the problem list term are present but not necessarily in the same order. Additional words in the SNOMED string are allowed.

Table 3. Matching methods

Term Manipulations

Multiple matching passes were done against the datasets in order to identify the reasons that some terms might not match using either the exact-match or match-all algorithm. This led to a set of term manipulations that were found to improve matching results. When terms were modified, the original term was always retained in the database for tracking purposes. The manipulation approaches are highlighted in Table 4 and detailed below.

Manipulation	Description
Normalization	Terms are processed using the NLM “LuiNorm” tools.
Cleaning	Extraneous characters are removed. This primarily addresses punctuation and multiple white spaces.
Correction	Misspellings are corrected.
Expansion	Acronyms and abbreviations are expanded.
Replacement/Deletion	Redundant words or acronyms are replaced with different strings or deleted entirely.
Post-Coordination	Words or strings relating to laterality, chronicity or other modifiers are removed from the terms to facilitate matching. The terms are then annotated as needing post-coordination.

Table 4. Term manipulation methods

Normalization. Normalization is a series of steps used to transform a text string into a “standard form”. The NLM provides a set of tools, called the Specialist Toolset, to process text strings. The toolset includes a pair of programs, “norm” and “luiNorm” that, when presented an input string, transforms it and outputs a “normalized” string. Within the UMLS, there are tables that hold similarly processed SNOMED CT strings for comparison. The first six steps of the two programs, “norm” and “luiNorm” are identical. The “norm” tool can potentially output multiple normalized strings for a given input. The “luiNorm” tool adds an additional “canonization” step that selects a single output if there are multiple normalization outputs. This project used “luiNorm” to facilitate the automation of the process and remove the need to manually review and select an output from the “norm” tool for each problem list term. An example is the term “Hodgkin’s disease, NOS” that, when normalized by either tool, resulted in the string “disease hodgkin.” The steps of the transformation are shown in Table 5.

Step #	Transformation	Result
1	Remove genitive	Hodgkin’s diseases ⇒ Hodgkin diseases, NOS
2	Strip punctuation	Hodgkin diseases, NOS ⇒ Hodgkin diseases NOS
3	Strip stop words	Hodgkin diseases NOS ⇒ Hodgkin diseases
4	Lowercase	Hodgkin diseases ⇒ hodgkin diseases
5	Uninflect	hodgkin diseases ⇒ hodgkin disease
6	Sort words	hodgkin disease ⇒ disease hodgkin

Table 5. LuiNorm and norm normalization steps[16]

Cleaning. Cleaning is the removal of extraneous characters, primarily punctuation marks. Periods, parenthesis, dashes, and slashes are common candidates. The unmatched data was manually reviewed and a cleaning algorithm configured to replace the candidates with a white space. Multiple contiguous white spaces were also removed.

Inspection of bad words. Terms were parsed into individual words and the words checked for validity by comparison with a UMLS dictionary. Inspection of “bad” words can highlight outright misspellings. The misspellings were replaced with the corrected word. The bad words may also be acronyms. The acronyms may be expanded or deleted depending on their context in the term. For instance, for the term “GERD”, the acronym is expanded to ”gastroesophageal reflux disease”. However, if the complete term is “GERD gastroesophageal reflux” the string “GERD” may simply be eliminated.

Post-coordination. Post-coordination focuses on the elimination of words with SNOMED CT post-coordination implications. This includes words involving laterality (e.g. “left” or “right”), chronicity (e.g. “chronic” or “recurrent”), and other strings such as “query” or “not otherwise specified” (NOS). These are removed and the associated terms are marked for post coordination. After matching the remaining terms to SNOMED CT concepts, the matched terms are then post-coordinated with attributes. An example is the problem list term “acute pneumonia” for which there is no matching SNOMED CT concept. If the term has the word “acute” stripped, the resultant term “pneumonia” maps to SNOMED CT concept id C0085762. The term is then post-coordinated by adding as a second concept the SNOMED CT concept id C0205178 which is “Acute (qualifier value).”

All datasets were processed using the same algorithms and methodology. When manipulations were done to the problem list term the original term was always preserved in the database in order to track the process. Corrections, expansions, replacement/deletions and post-coordination were done individually by dataset and the change terms stored. All measurements were performed identically across the datasets.

Tools Design and Prototyping

This bulk of this project was done in the Oracle database environment. Algorithms were coded and tested using PL/SQL, a proprietary programming language with embedded SQL. The match-all algorithm was designed using Oracle Text, a text processing extension to Oracle. Although Oracle based, some of the functionality was also tested in the MySQL environment.

Since a goal of the project was to define a tool set, the project was iterated in an AGILE development style. Various database schemas were developed, tested and modified, as were associated queries and algorithms. Multi-window query tools such as Toad were used to simulate interface windows. The iterative nature of the process helped develop a set of requirements and use cases that can be used to define and implement true interfaces.

Results

Proposed Process Methodology

The main process of the proposed methodology has three phases. In the first phase, the unique set of terms is extracted from the problem list and serves as the input for all further matching efforts. The *exact-match algorithm* is applied. The remaining unmatched terms are iteratively modified in various ways as detailed below and the *exact-match algorithm* is reapplied after each manipulation. After these manipulations, the remaining unmatched terms are then normalized and the *exact-match algorithm* is applied to the normalized strings. Statistics are gathered at all matching steps. All matches are reviewed for clinical appropriateness and any erroneous matches are removed. The remaining unmatched terms serve as input into Phase 2.

During Phase 2 the terms remain unchanged. Now the *match-all algorithm* is applied to the unmatched terms. The *match-all algorithm* is also applied to the normalized string of the remaining unmatched terms. The matches from Phase 2 are then reviewed for clinical appropriateness and any erroneous matches are removed.

During Phase 3 the remaining unmatched terms are reviewed for possible post-coordination. This process was described above and results in the modification of some terms. The *exact-match algorithm* is applied to the modified terms. The remaining unmatched terms are then normalized and the *exact-match algorithm* is applied to the normalized strings. The *match-all algorithm* is next applied to the remaining unmatched terms and finally the *match-all algorithm* is applied to the normalized strings of the unmatched terms. This culminates with a final clinical review. The remaining unmatched terms are then available for manual review and the manual selection of an appropriate SNOMED CT concept. The high-level process phases are shown in Figure 1. Each phase is detailed below with the key to all Figures shown in Figure 2.

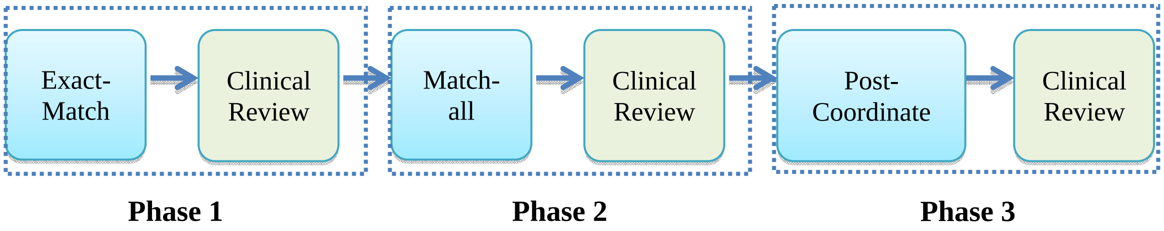


Figure 1. High-level phases of process methodology

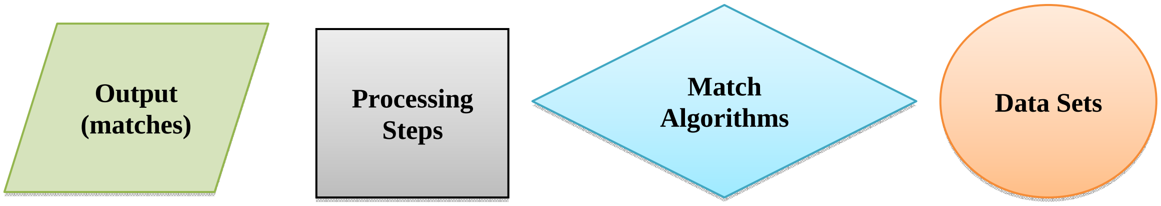


Figure 2. Key to figures in process diagrams

Phase 1: The Exact-Match Algorithm

Phase 1 focuses on the exact-match algorithm and has multiple steps. Each step is described below and illustrated in Figures 3a and 3b.

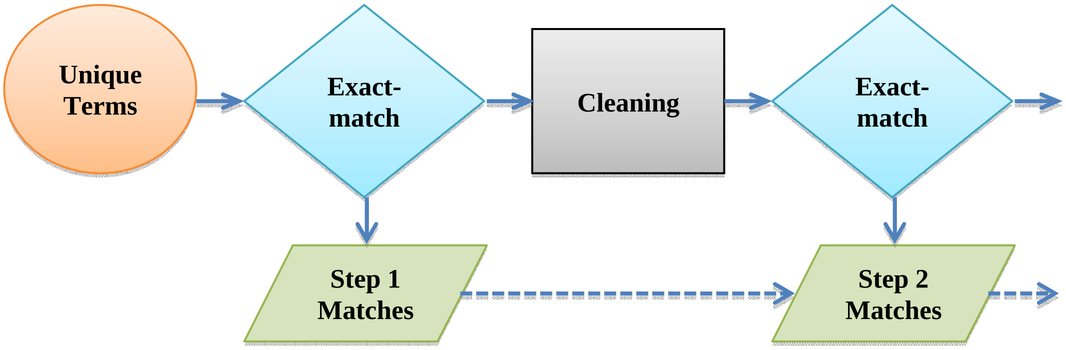


Figure 3a. Phase 1 – Steps 1 & 2

Step 1: Exact match of unique terms. Phase 1 begins with the set of unique terms extracted from the original problem list terms. The exact-match algorithm is applied to the unique terms and the matches recorded. These terms are tentatively marked as matched. The matches are reviewed and any duplicate matches for a unique term are adjudicated. The matches are then marked and the original terms are associated with the matched SNOMED CT concept id (CUI).

Step 2: Cleaning. A set of unique words is parsed from the remaining unmatched unique terms. These words are matched against the UMLS dictionary table and any non-matching words are marked as suspect. The set of unique words, both suspect and matched, are reviewed for their characteristics. The suspect words review can be guided by the presentation of the words in the count of their inclusion in unique terms. It is likely that various punctuation characters, including parentheses, commas, dashes, and slashes impact the word strings. The unique words are annotated by iteration number and preserved for analysis.

A “cleaning” tool is then used to strip the unwanted punctuation characters from the unique terms. Extra spaces between words are reduced to single spaces. The cleaned terms then replace the original terms in further analysis; the original terms are stored in the database to allow match comparisons to both the cleaned term and the original problem list term.

Since the cleaned terms may now be identical to terms that were already successfully matched, a check is done to determine if any cleaned terms match existing matched terms. If so they are annotated with the match CUI.

A new set of unmatched unique terms is now extracted, this time from the cleaned terms. The exact-match algorithm is applied to the unmatched terms and the matches reviewed as in step 1.This prepares the system for step 3.

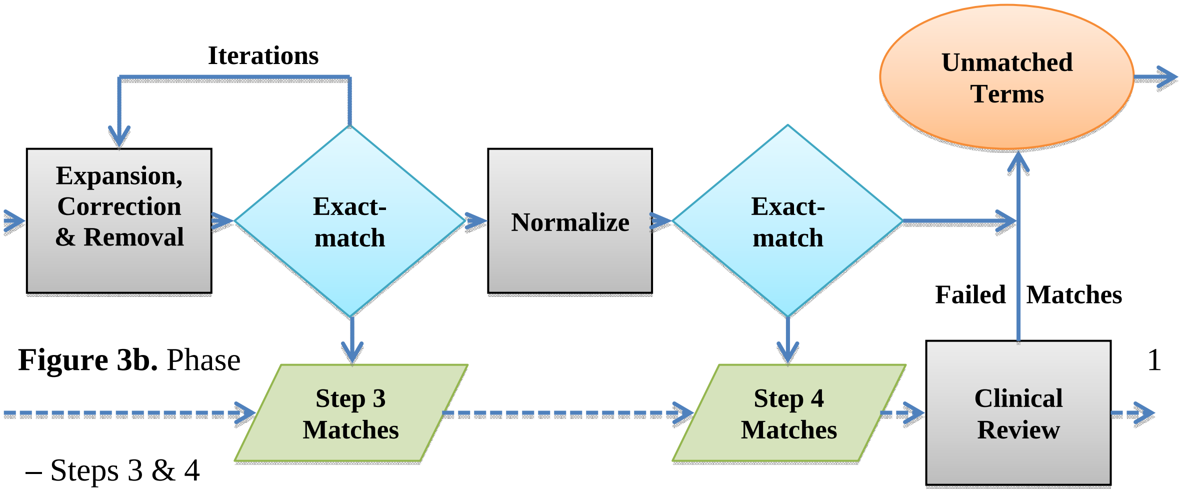


Figure 3b. Phase 1 – Steps 3 & 4

Step 3: Expansion, Correction and Removal. A set of unique words is created as in previous steps. The words are again compared against the UMLS dictionary and non-matches are marked as suspect. With the stripping of punctuation, it is likely that the list of suspect words is significantly reduced. Tracking of words created by each iteration is done to identify any words created as a function of the multiple step iterations. This step focuses on the correction of misspellings and the expansion or elimination of acronyms. The word replacement values are stored in the database. The unique terms are then modified with a word replacement algorithm. The resulting modified terms are stored and checked for pre-existing matches as in previous steps.

A set of unmatched unique terms is again extracted, this time from the set of modified terms. The exact-match algorithm is applied to the unmatched terms and the matches reviewed as in previous steps. This prepares the system for possible iterations of step 3.

The word review process is followed as in previous steps. Particular attention is given to new suspect words created via a previous iteration to evaluate the effect of potential erroneous word replacements or corrections. It is likely that the suspect word list has been reduced but not entirely eliminated. Another set of replacement words can be generated and the step 3 word replacement and terms matching cycle repeated.

Step 4: Normalization. The unique problem list terms unmatched after step 3 are normalized using the luiNorm tool. The resulting strings are stored with the associated unique terms. The exact-match algorithm is applied to the normalized strings and is matched against the UMLS table of normalized SNOMED CT terms. Multiple matches could exist for each normalized term and these need adjudication as in the prior steps. No changes or manipulations to the unique terms are done; the normalized terms are stored separately and associated with the unique term. Matches are reviewed as in previous steps. This prepares the system for the Phase 2.

Clinical Review. All matches are reviewed for clinical appropriateness and any erroneous matches are removed.

Phase 2: The Match-All Algorithm

Phase 2 focuses on the match-all algorithm. The match-all algorithm is applied to the unique problems list terms unmatched after the Phase 1. This type of match requires all words in the problem list term to be present in the matched string. The word order can differ and the matched string may have additional words. No additional term manipulations are done during these steps. The match-all phase is detailed in Figure 4.

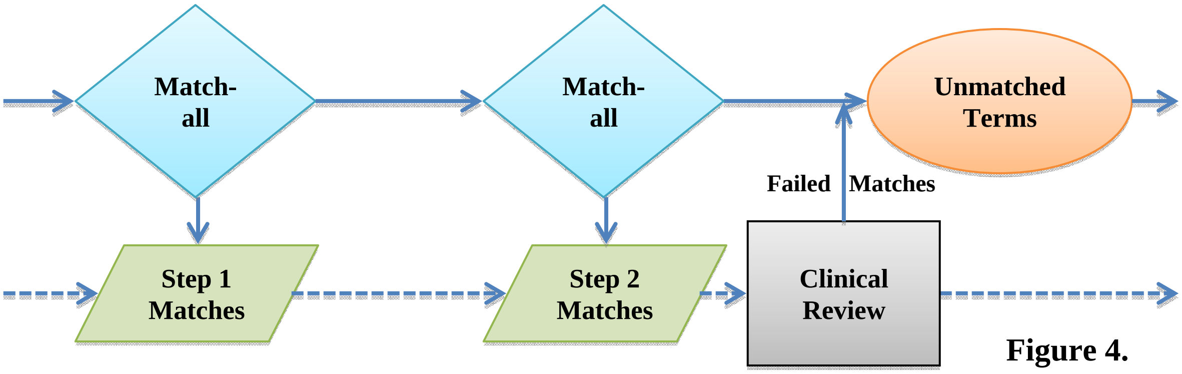


Figure 4. Phase 2 details

Step 1. Match-all on non-normalized terms. The match-all algorithm is applied to unmatched unique terms remaining after Phase 1. The match-all algorithm can generate multiple matches; single word terms can generate hundreds of potential matches. Therefore careful review and adjudication of the proposed matches is necessary. When the match review is complete the matched terms are updated as in the previous phase.

Step 2. Match-all on normalized terms. The match-all algorithm is applied to the normalized strings of the remaining unmatched strings. A review of proposed matches is done as in step 1. When the match review is complete the matched terms are updated as in the step 1. After these steps a clinical review of the new matches is done.

Clinical Review. All matches are reviewed for clinical appropriateness and any erroneous matches are removed**.**

Phase 3: Post-coordination

This phase focuses on the removal of words with post-coordination implications in the remaining unmatched unique strings. The words expressing laterality, chronicity and other words appropriate for post-coordination are removed and the associated terms are marked for later post-coordination. A new set of unmatched unique terms is generated from the stripped terms for additional matching efforts. The two match algorithms are then executed in the steps detailed in Figures 5a and 5b.

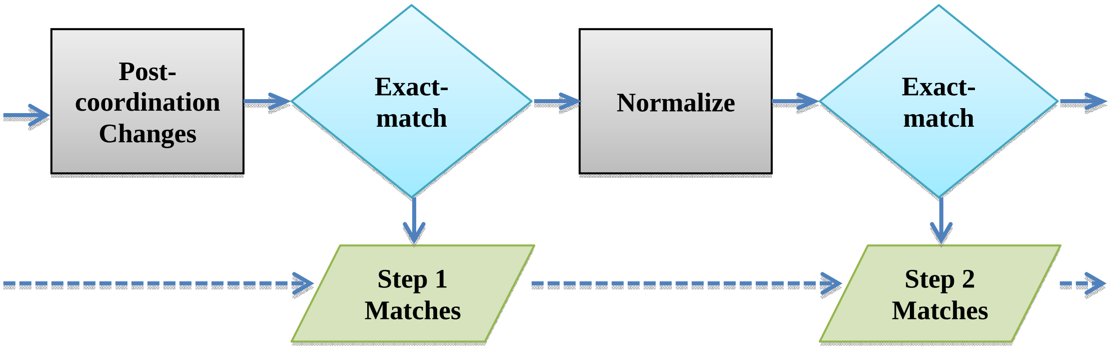


Figure 5a. Phase 3 – Steps 1 & 2

Step 1. Exact-match on non-normalized terms after post-coordination modifications. The exact-match algorithm is applied to unique terms remaining after Phase 2 and removal of words for post-coordination. Matches are reviewed and processed as in previous steps.

The unmatched terms are then normalized as in previous steps.

Step 2. Exact-match on normalized terms after post-coordination modifications. The exact-match algorithm is applied to the normalized strings of the remaining unmatched terms. Matches are reviewed and processed as in previous steps.

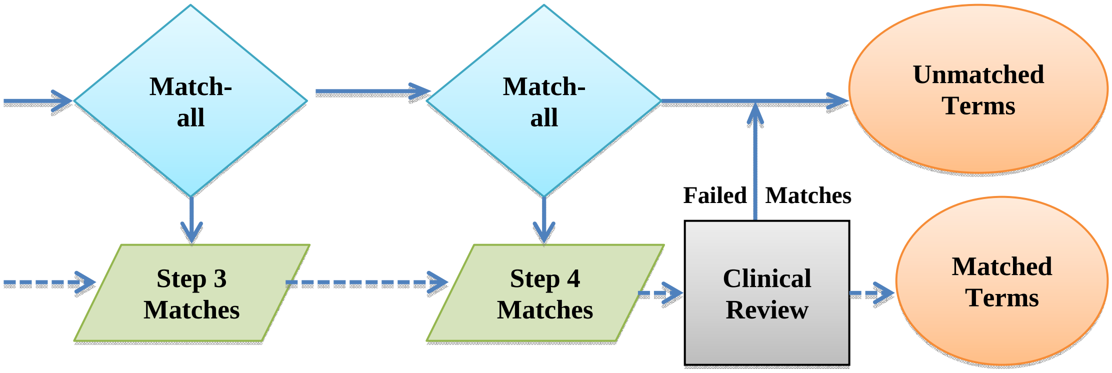


Figure 5b. Phase 3 – Steps 3 & 4

Step 3. Match-all on non-normalized terms after post-coordination modifications. The match-all algorithm is applied to remaining unmatched terms. Matches are reviewed and processed as in previous steps.

Step 4. Match-all on normalized terms after post-coordination modifications. The exact-match algorithm is applied to the normalized strings of the remaining unmatched terms. Matches are reviewed and processed as in previous steps.

Clinical Review. All matches are reviewed for clinical appropriateness and any erroneous matches are removed.

This concludes the automated processing of the problem list terms. The remaining unmatched unique terms are available for manual review and assignment of appropriate SNOMED CT concept IDs.

Dataset 1 (Lau Family Practice problem list)

	Original Terms	Unique Terms	Unmatched Unique Terms	Step Matches	Unique Term Matches (Total)	Original Term Matches (Total)
Phase 1						
Step 1	7833	1822	1822	626	626	4652
Step 2	7833	1817	1209	0	626	4652
Step 3	7833	1792	1184	44	652	5170
Step 4	7833	1792	1140	306	958	5947
					53.5%	**75.9%**
Phase 2						
Step 1	7833	1792	834	201	1159	6550
Step 2	7833	1792	633	20	1179	6616
					65.8%	**84.5%**
Phase 3						
Step 1	7833	1706	527	44	1218	6936
Step 2	7833	1706	483	25	1243	7051
Step 3	7833	1706	458	24	1272	7123
Step 4	7833	1706	434	0	1272	7123
					74.6%	**90.9%**

Table 6. Dataset 1 results. Stepwise and overall matching of problem list terms to SNOMED CT using the developed methodology.

Comparison with Lau Results

Table 6 shows the results of applying the described methodology to the first dataset (the Lau Family Practice problem list.) Overall, 90.9% of terms were matched to SNOMED CT concepts; 74.6% of unique terms were matched.

Initial dataset metrics. Lau provided an initial set of metrics on the problem list terms. These included total and unique counts for terms and words and other term and word metrics. Similar metrics were gathered on the problem list terms after loading into this project's analysis system. Comparisons are shown in Table 7.

	Total Terms	Unique Terms	Total Words	Unique Words	Median Word Length	Most Common Word
Lau (ref)	7,833	1,713	16,455	1,764	8	Hypertension, 585 times
Current Methodology	7,833	1,822	16,936	1,811	7	Hypertension, 585 times
Difference	0	109	471	47	1	None

Table 7. Comparison of key dataset metrics. Data metrics as published by Lau compared with the metrics on the same problem list using the current methodology.

Matching metrics. Lau's process involved two cycles. The first used the mapping algorithms against the initially generated set of unique terms. After the first cycle, multiple iterations of term modification and matching were done. Partial-match algorithms were also used. Details and counts of these interim iterations were not available. After all modifications were completed, a final cycle of matching was performed. Comparisons of Lau's matching results to this project's results are shown in Table 8.

	Initial Unique Terms	Initial Cycle Exact-match	Initial Cycle Match-all	Initial Cycle Exact-Match & Match-all	Final Cycle Exact-Match & Match-all	Final Unique Terms
Lau (ref)	1,713	52.8%	9.8%	62.3%	91.6%	1,409
Current Methodology	1,822	53.4%	12.4%	65.8%	74.6%	1,706

Table 8. Comparison of key matching metrics. Data matching as published by Lau compared with the metrics on the same problem list using the current methodology.

Dataset 2 (Solomon problem list)

	Original Terms	Unique Terms	Unmatched Unique Terms	Step Matches	Unique Term Matches (Total)	Original Term Matches (Total)
Phase 1						
Step 1	2257	2121	2121	68	68	71
Step 2	2257	2060	1992	8	76	79
Step 3	2257	2040	1964	156	232	254
Step 4	2257	2040	1808	228	460	503
					22.5%	**22.3%**
Phase 2						
Step 1	2257	2040	1580	238	698	768
Step 2	2257	2040	1342	13	711	783
					34.9%	**34.7%**
Phase 3						
Step 1	2257	1903	1192	253	964	1122
Step 2	2257	1903	968	133	1097	1323
Step 3	2257	1903	806	168	1265	1537
Step 4	2257	1903	638	10	1275	1553
					67.0%	**68.8%**

Table 9. Dataset 2 results. Stepwise and overall matching of problem list terms to SNOMED CT using the developed methodology.

Dataset 3 (CORI GI Endoscopy problem list)

	Original Terms	Unique Terms	Unmatched Unique Terms	Step Matches	Unique Term Matches (Total)	Original Term Matches (Total)
Phase 1						
Pass 1	1,5142,40	268	268	67	67	235,118
Pass 2	1,5142,40	267	200	3	70	235,485
Pass 3	1,5142,40	265	195	14	84	249,631
Pass 4	1,5142,40	265	181	82	166	990,485
					62.6%	**65.4%**
Phase 2						
Step 1	1,5142,40	265	99	34	200	1,346,259
Step 2	1,5142,40	265	65	2	202	1,348,063
					76.2%	**89.0%**
Phase 3						
Step 1	1,5142,40	265	63	3	205	1,352,711
Step 2	1,5142,40	265	60	0	205	1,352,711
Step 3	1,5142,40	265	60	0	205	1,352,711
Step 4	1,5142,40	265	60	0	205	1,352,711
					77.4%	**89.3%**

Table 10. Dataset 3 results. Stepwise and overall matching of problem list terms to SNOMED CT using the developed methodology.

Comparisons Across Datasets

Table 9 shows the results of application of the current methodology to the second dataset, contributed by H. Solomon. Of 2,257 terms, 68.8% overall and 67.0% of unique terms were matched to SNOMED CT concepts.

Table 10 shows the results of application of the current methodology to the third dataset, the CORI GI Endoscopy problem list. Of 1,514,240 terms, 89.3% overall and 77.4% of unique terms where mapped to SNOMED CT concepts.

Problem list dataset metrics. After loading into the database and prior to processing, metrics were gathered on the three problem list datasets. A computed metric, “Terms to Unique Terms Ratio”, was created to express the “uniqueness” of each of the original terms. It is the quotient of the total terms divided by the unique terms. The metrics are shown in Table 11.

Dataset	Total Terms	Unique Terms	Terms to Unique Terms Ratio	Total Words	Unique Words
1	7,833	1,822	4.29	16,455	1,764
2	2,257	2,121	1.06	10,652	1,710
3	1,514,240	265	5,714.11	586	115

Table 11. Comparison of dataset metrics.

1 = Lau Family Practice problem list; 2 = Solomon problem list; 3 = CORI GI Endoscopy problem list.

Problem list matching metrics. The matching percentages and matched and unmatched unique term quantities were gathered at each project step. The metrics at the end of each phase in the methodology are shown in Tables 12a and 12b.

	Start	Phase 1 Matching		Phase 2 Matching		Phase 3 Matching	
Dataset	**Begin Unique Terms**	**Terms Matched**	**Ending Unique Terms**	**Total Matched Terms**	**Ending Unique Terms**	**Total Matched Terms**	**Ending Unique Terms**
1	1,822	958	1,792	1179	1,792	1272	1,706
2	2,121	460	2,040	711	2,040	1,275	1,903
3	267	166	265	202	265	205	265

Table 12a. Matched terms and total unique terms at different phases of the process

Dataset	Phase 1 Matching		Phase 2 Matching		Phase 3 Matching	
	Unique Terms	**Total Terms**	**Unique Terms**	**Total Terms**	**Unique Terms**	**Total Terms**
1	53.5%	75.9%	65.8%	84.5%	74.6%	90.9%
2	22.5%	22.3%	34.9%	34.7%	67.0%	68.8%
3	62.6%	65.4%	76.2%	89.0%	77.4%	89.3%

Table 12b. Percentage of total terms and unique terms matched at each phase of the process.

1 = Lau Family Practice problem list; 2 = Solomon problem list; 3 = CORI GI Endoscopy problem list.

Key Interfaces

During the many iterations of the data, notes were kept on functionality and interfaces that could facilitate the proposed process. From these, use cases were created for documentation of functionality that is important in the user interface of an application designed to help users in the steps of this matching process. Two steps in the process are particularly well suited to improvement through quality interfaces. The use cases for those steps are the “View and Select Matches” and the “View Words and Propose Changes” use cases and are presented below. Paper prototypes of proposed user interfaces were created for these use cases. Additional use cases are presented in Appendix 3.

A prototype for implementing the "View and Select Matches" use case is shown in Figure 6. Figure 7 is a prototype for implementing the "View Words and Propose Changes" use case.

Use Case: **View and Select Matches**

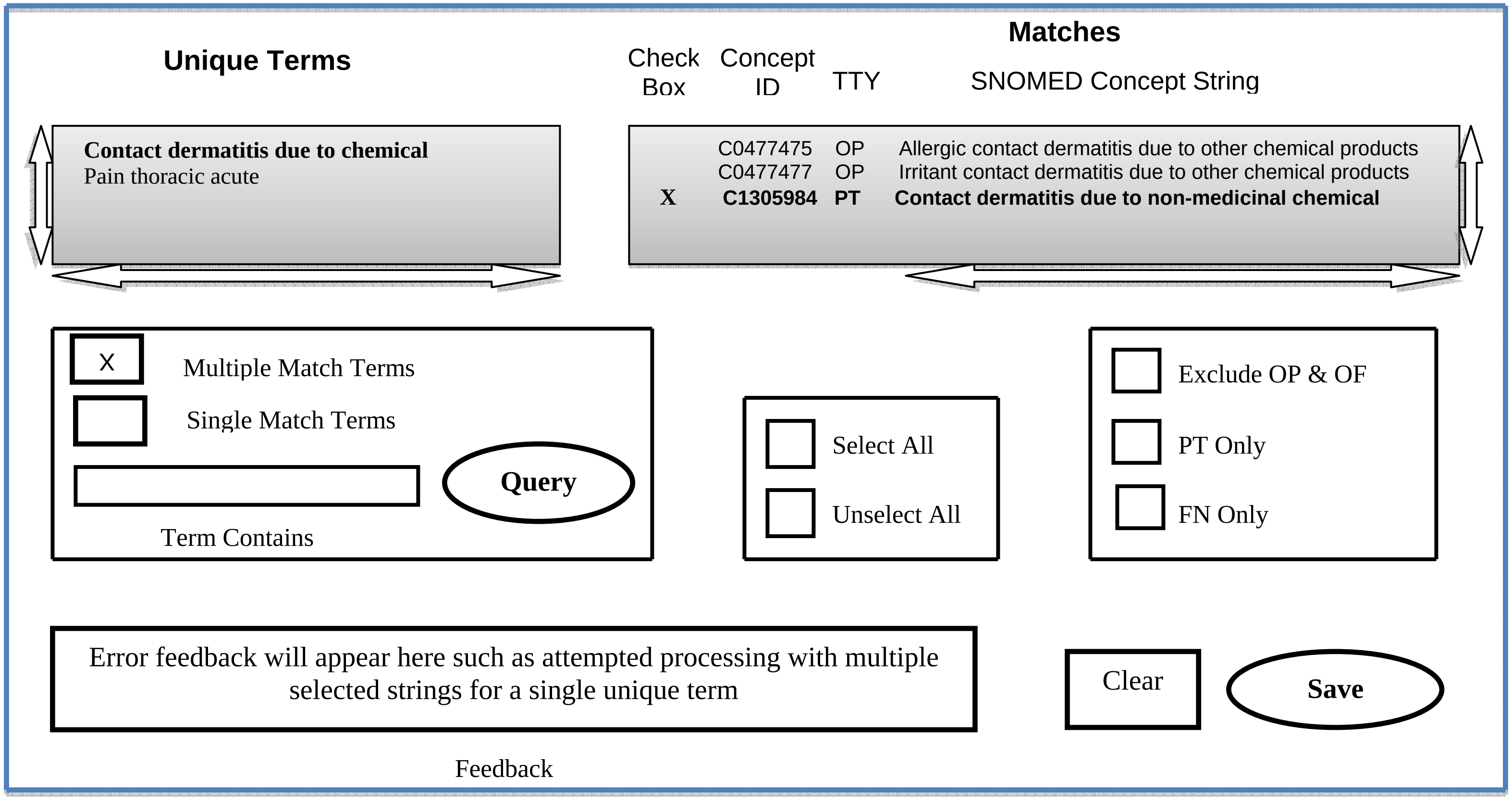
Description: User reviews matches associated with the unique terms that have been proposed by the algorithms. User selects the appropriate match and/or excludes inappropriate matches. Only one match is allowed per unique term. If a match is confirmed, on saving the system updates the match status for the term to “matched” and associates the matched CUI with the original term. If no matches are deemed appropriate for a unique term, that term’s status is reset to “unmatched”.

1. User queries the unique problem list terms choosing one of the three following sets of terms:
    a. Terms with multiple matches (default)

    b. Terms with single matches
    c. Terms that contain a specified string.

2. System displays the unique problem list terms with their suggested SNOMED CT matches in a side-by-side manner with a check box available for each SNOMED CT term record. Data displayed includes:
    a. SNOMED CT Concept string (STR)
    b. SMOMED CT Concept ID (CUI)
    c. SNOMED CT Term Type (TTY)
3. User checks the appropriate SNOMED CT term. User can also “unselect all” previously selected matches.
4. User selects “Save”
    a. The system verifies that there is only one match record selected. If more than one record is selected the system warns the user and returns to the previous step.
    b. If no match records are selected the system alerts the user that all matches will be discarded and requires a positive response to continue.
5. System processes matches where there is one match selected per unique term.
    a. System marks match as selected.
    b. All original terms associated with matched unique term are updated with the selected match’s CUI.
    c. Any non-selected or excluded match records are marked appropriately.
6. Feedback is provided including the total number of unique terms and original terms matched.

Process Step: View and Select Matches



Unique Terms

Contact dermatitis due to chemical
Pain thoracic acute

Matches

Check Box	Concept ID	TTY	SNOMED Concept String
	C0477475	OP	Allergic contact dermatitis due to other chemical products
	C0477477	OP	Irritant contact dermatitis due to other chemical products
X	**C1305984**	**PT**	**Contact dermatitis due to non-medicinal chemical**

☒ Multiple Match Terms
□ Single Match Terms
Term Contains

□ Select All
□ Unselect All

□ Exclude OP & OF
□ PT Only
□ FN Only

Error feedback will appear here such as attempted processing with multiple selected strings for a single unique term

Feedback

Figure 6. Prototype of interface to implement the “View and Select Matches” use case

This example is the match results from a Match-all algorithm match. The user has selected multiple match terms and two terms are displayed in the unique terms window. The window can be scrolled to see the complete set. The user has currently highlighted the first unique term, which populated the Matches window with three matching SNOMED CT records. The user has selected the third matching term as the correct match by clicking the check box.

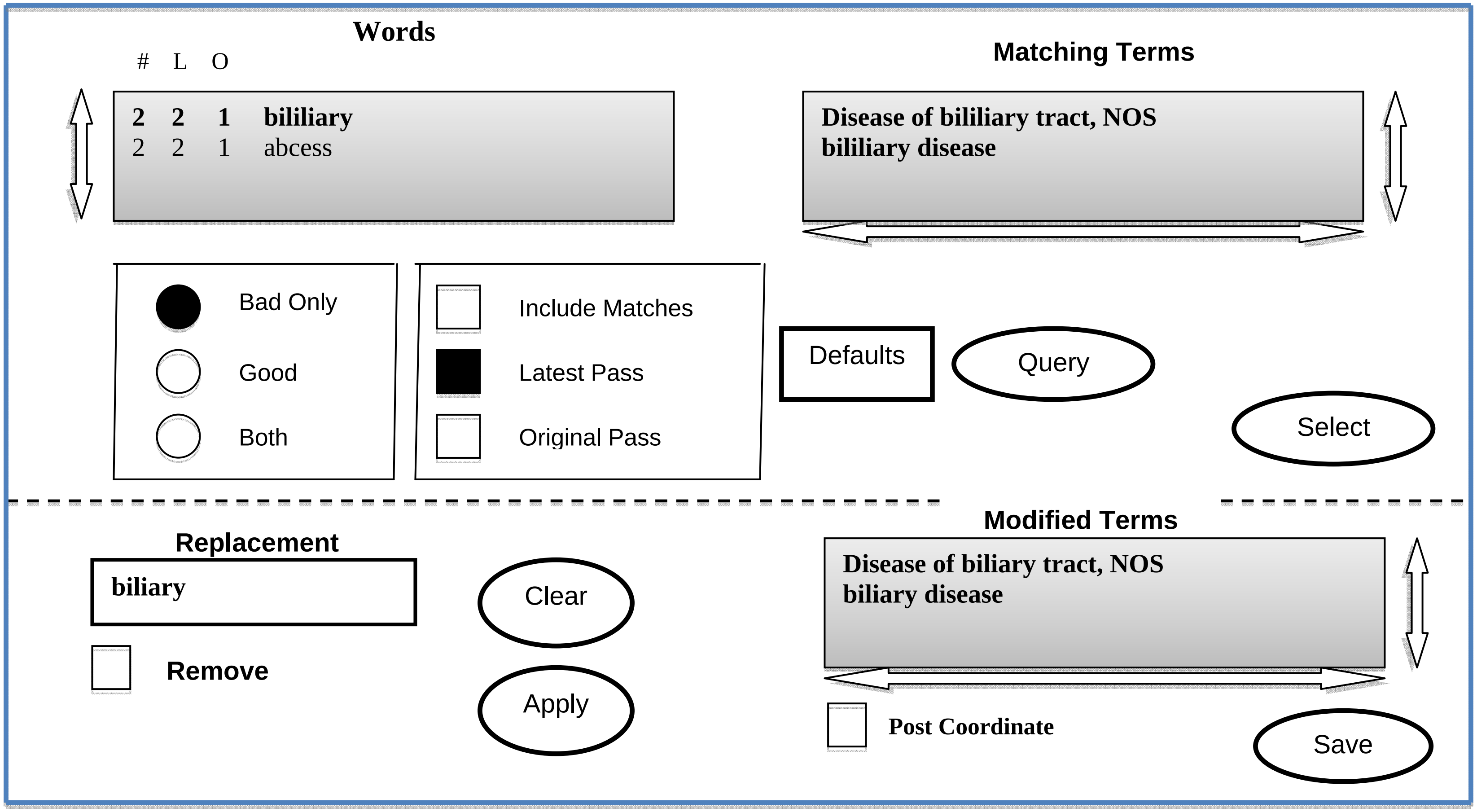
Use Case: **View Words and Propose Changes**

Description: During each step in Phase 1, the user will review the unique words derived from the remaining unmatched unique terms. These words are checked against a UMLS dictionary and, if not found, are marked as “bad”. The user can select a word and view all of the terms that contain the word. If the selected word needs to be replaced or removed, the user can specify the replacement or removal and see the effect on the associated terms. If the replacement or removal is found to be appropriate, the user can save the change. The prototype interface that implements this use case is found in Figure 7.

1. User selects the subset of words to review: “bad”, “good” or all words may be displayed. The system defaults to “bad”.
2. User can request that the already matched terms be included in the display. The default is to not include matched terms. The already matched terms are shown for informative purposes; word replacements or removal do not affect already matched terms.
3. User can select a combination of the original or latest step values to analyze words created by each Phase 1, step 3 iteration. The system defaults to latest iteration and displays the latest iteration number as a default.
4. The system displays the appropriate words along with a count of the number of unique terms in which the word appears. The list is sorted by term count descending and includes:
    a. Term count of terms containing bad word
    b. Latest iteration # and Original iteration # of bad word detection
    c. Word

5. User selects a word by and all of the terms containing the word are displayed.
6. If appropriate, the user indicates a replacement word or removes the word. The proposed replacement or removal in terms containing the word is displayed.
    a. User accepts or clears the suggested modification or removal.
    b. If accepted and the change requires the new term to be marked for post-coordination the user indicates this.
    c. The system stores the word replacement/removal data and updates a table of the unmatched terms to reflect the changed terms. This is to support the possible editing of a term with two or more bad words.
7. User iterates through the process by selecting a new word for processing.

Process Step: View Word, Analyze & Save Changes



Words

#	L	O	
2	**2**	**1**	**bililiary**
2	2	1	abcess

Matching Terms

Disease of bililiary tract, NOS
bililiary disease

● Bad Only
○ Good
○ Both

□ Include Matches
■ Latest Pass
□ Original Pass

Replacement

biliary

□ **Remove**

Modified Terms

Disease of biliary tract, NOS
biliary disease

□ **Post Coordinate**

Figure 7. Prototype of interface to implement the "View Words and Propose Changes" use case

In this example the user has used the default query attributes. Two words are shown in the Words window. The user selected "bililiary" and the system returned on the right two unmatched terms containing the word. The user clicked the select button which populated the lower part of the screen for replacement of the term, with effect on the Modified Terms shown.

Discussion

Comparison with Lau Results

Lau's work provided two sets of measurements, the first resulting from the initial loading and matching of the problem list dataset and the second from the final matching cycle after completion of all manual modifications to the problem list terms. Interim measurements were not available. The first step in this project was to attempt to reproduce the work performed by Lau and associates. As presented in Table 7 above, however, the initial metrics did not match, which may have had an effect on the final percent of terms with matches to SNOMEC CT. Analysis of the processes used by both groups led to the discovery of a key difference in the handling of word and term parsing. The Lau project used MySQL for the database infrastructure. The MySQL text-processing algorithms by default ignore words of 3 characters or less in length in the parsing of words from strings and in the generation of unique terms. The Oracle text-processing algorithms used by the current project default to a word length of 1. The parameter in MySQL is configurable, however. When a MySQL instance was created and with the parameter set to 1 rather than 3 characters, we found a similar number of unique terms in the first dataset.

In addition, Lau's process involved two cycles. The first used the mapping algorithms against the initially generated set of unique terms. After the first cycle, multiple iterations of term modification and matching were performed. Partial-match algorithms were also used. Details and counts of these interim iterations are not available. After all modifications were completed, a final cycle of matching was performed.

As this project’s methodology included term modifications in its exact-match phase, the initial cycle metrics were not exactly comparable however the percentages were acceptably close. Since Lau’s final matching cycle was preceded by multiple manual term modification iterations, it was expected that this project’s final cycle results would be less than Lau’s. The effect of Lau’s manual modification of problem list terms is illustrated in the final cycle’s unique terms count. The Lau cycle concluded with 1,409 unique terms while this project’s automated approaches resulted in a final unique term count of 1,706. It is expected that manual term modification of this project’s unmatched terms could have driven the matching results closer. While we were unable to precisely reproduce the Lau results, we felt this project’s database, matching algorithms and methodology were robust enough to proceed with additional datasets.

Multiple Problem List Comparisons

Initial dataset metrics: Some notable differences exist in the three problem sets. The magnitude of the total terms, unique terms and word metrics are similar in datasets 1 and 2. Dataset 3 is quite different in all metrics. This is perhaps best illustrated in the ratio metric. This metric expresses the amount of problem term “reuse”. On average, each unique term in dataset 1 is used 4.29 times in the complete problem list set. The ratio for dataset 2 is 1.06. This could be expected from a dataset designed for testing or validation purposes. It should have a broad span and little reuse of terms. Dataset 3, from a real world GI endoscopic reporting application, is a more focused problem list. It has far fewer unique terms and a much larger ratio metric (5,714). The total and unique word counts are also much smaller. This application has a specific medical focus and all terms are drawn from a controlled vocabulary, which limits variation.

Dataset matching metrics. Overall, the matching percentages of datasets 1 and 3 were encouraging. Both had similar final unique term matching over 70% and total term matches near 90%. Phase two (match-all) was more effective for dataset 3 while phase three (post-coordination) was more effective for dataset 1. However the matching percentages were lower at all points for dataset 2.

Analysis found that the composition of dataset 2 differed from the other datasets in one significant characteristic that affected the matching rates. Almost half of dataset 2's initial set of terms (1,045 of 2,121) contained one of the abbreviations of “hx” (for “history”) or “fhx” (for “family history”). These abbreviations did not appear in either dataset 1 or 3 and the strings “history” or “family history” occurred only twelve times in dataset 1 and not at all in dataset 3. During Phase 1 for dataset 2, the strings “hx” and “fhx” were expanded to “history” and “family history”, respectively. The expansions enabled 35.2% of the associated terms to be matched. In Phase 3, the strings “history” and “family history” were stripped and the terms marked for post-coordination. This enabled an additional 37.9% of the terms to be matched and accounts for the larger match percentage increase in Phase 3 for dataset 3 relative to datasets 1 and 2.

Dataset 3 was further analyzed to assess a complete, real-world mapping process. The availability of the owner of the data allowed a more thorough review of the project's mappings and the unmatched terms. The 60 unmatched terms remaining after Phase 3 were manually reviewed for assignment to SNOMED CT concepts. Four of the unmatched terms were found to contain multiple concepts and needed adjudication. Matching SNOMED CT concepts were found for all of the remaining fifty-six unmatched terms. A complete local vocabulary is being generated for the CORI application with potential mappings to additional terminologies.

Project Methodology Assessment

Multiple process paths were tested during the development of the recommended methodology. While the project's goal was to provide a set of tools that would automate the mapping of problem list terms, it was expected that some amount of manual mapping would be required following the application of the automated tools. The final methodology was selected because it provided a straightforward automated process while delivering acceptable matching results. Other process flows provided higher matching results but required more manual intervention. One key issue in the methodology was the process of term normalization.

Normalization changes. An issue with normalization was discovered late in the project. The luiNorm tool, because it makes one selection from potentially multiple transformations (a process called canonization), sometimes displays some idiosyncrasies, at least from this project's point of view. Uninflection, the fifth step in normalization as shown in Table 5 above, can lead to multiple forms for a given word. An example is the word "left," which can have an adjective form (opposite of right) or verbal form (to depart). It can also be the "s-stripped" form of the noun "leaves" (parts of plants). The issue is that luiNorm maps all of the base forms to the string "leaf" as shown in Table 13.

Base Form	Uninflected Forms	Canonical Form
leaf	leaf	leaf
leaves	leave, leaf	leaf
left	leave, left	leaf

Table 13. Canonization of the word "left"

Therefore, any unmatched problem list terms containing the word “left” will contain the word “leaf” in its normalized string. The SNOMED CT normalized strings are not similarly transformed; they use the “norm” tool and are manually reviewed or canonized by the NLM project group. The normalized strings of SNOMED CT terms containing the word “left” contain the word “left” and none contain the word “leaf”. A similar situation occurs for the word “colon” which luiNorm canonizes to “cola”. Any problem list terms containing these strings would not get matched in their normalized form with either matching algorithm. Review of the matching data confirmed this.

Using luiNorm involves a tradeoff of potential inaccuracies in canonization versus the significant process automation advantage of a single normalization output. As discussed in the Methods section, the “norm” tool could replace luiNorm but its use would require the manual review and selection of each normalized transformation. It is recommended that luiNorm continue to be used, but that a list of undesirable word transformations be developed to apply to the normalized strings of the unmatched problem list terms prior to matching. This compromise allows the benefits of automation while mitigating the potential inaccuracy in the luiNorm canonization step.

Future Work

Several future projects are suggested by the current work. The processing of more datasets, from diverse sources and medical foci, could potentially identify areas for improvements. This research could include matching against additional target terminologies such as ICD10-CM. Of particular interest is the potential use of the Meta-Map tools. These tools could potentially replace all of the project’s matching algorithms and provide automation to the post-coordination process.

Additional datasets. While the three datasets were diverse, exercise of the system with additional datasets would be desirable. The luiNorm canonization issue was highlighted by the use of dataset 3 which, being GI-related, helped identify the incorrect normalization of the string "colon" and triggered the investigation that identified the canonization issues.

Development of Key Interfaces. Most of the methodology was instantiated by the manual execution of PL/SQL scripts and packages, operating system level commands, and various queries. This approach is both time-consuming and has the potential for process errors. Development of interfaces would allow both more accurate process execution and the potential for multiple users to access the system simultaneously.

Extension to Other Terminologies such as ICD10-CM. The project utilized a UMLS "subset" of data including only the SNOMED CT terminology. A different subset could be generated using another, or multiple terminologies. Some minor modifications to the matching algorithms, particularly involving the choice of attribute selection values, would be necessary.

The most recent releases of the UMLS include predefined relationships from SNOMED CT to other terminologies such as ICD10-CM. This project's toolset could be used to map a dataset to both SNOMED CT and ICD10-CM. It would be interesting to then compare the project's SNOMED CT to ICD10-CM relationships to the UMLS predefined relationships.

Potential Use of NLM Meta-Map Tool: In addition to the UMLS and Specialist Toolset, the NLM has released a set of tools under the MetaMap[7] moniker that can be configured or extended for mapping purposes in conjunction with the UMLS. These tools are part of the NLM’s natural language processing (NLP) efforts. The tools are focused on “providing access from biomedical text to the concepts in the unified medical language system” [17]. Though primarily focused on NLP processing of sentences from biomedical texts, an option within MetaMap is targeted at the processing and mapping of terms. MetaMap can be configured to utilize specific vocabularies and can return sets of concepts that when combined provide the highest weighted mapping.

During the project some investigation of these tools was done using sets of unmatched terms and the NLM’s web-based MetaMap interface. When specifying SNOMED CT as the target vocabulary, the output essentially provided post-coordinated sets of concepts. The results were promising and further research is warranted to determine if the MetaMap toolset could completely replace the set of matching algorithms and post-coordination activities. The Meta-Map toolset is a key research area for the NLM and leveraging their work could be a strategic advantage in long term mapping efforts.

Conclusion

Overall, the project achieved its goal of developing and testing a set of automated tools and a process methodology that can aid the mapping of problem lists onto a target terminology, specifically SNOMED-CT. The project's matching performance against the reference Lau dataset was judged sufficient to continue the project and while not achieving 100% mapping of problem list terms, the number of problem list terms requiring manual review was significantly reduced across all datasets. A manual review and assignment of unmatched terms was done for dataset 3 to assess a complete real world process and the results are being used to develop a local application vocabulary.

Using an Agile methodology, a database and associated matching algorithms were iterated and finalized. Interface use cases were developed and paper prototypes were created for key interfaces. The project also identified how a step in the NLM normalization toolset could adversely affect terminology mappings. Finally, several paths for future research were identified.

The tools and methodology developed during the project were found to be effective in the mapping of diverse problem list datasets to the SNOMED CT terminology and show promise for extension to other target terminologies.

References

1. U.S. Health and Human Services. Standards & certification interim final rule: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology. [Internet]. (2010) [cited 2010 Jan. 6] Available from: http://edocket.access.gpo.gov/2010/E9-31216.htm

2. Centers for Disease Control and Prevention, National Center for Health Statistics, International classification of diseases, Tenth revision, Clinical Modification (ICD-10-CM), [Internet]. (2010) [cited 2010 May 30] Available from: http://www.cdc.gov/nchs/icd/icd10cm.htm

3. International Health Terminology Standards Development Organization, SNOMED CT, [Internet]. (2010) [cited 2010 Feb. 8] Available from: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

4. AHIMA, Best Practices for Problem Lists in an EHR - Appendix A, [Internet], [cited 2010 Jan. 22] Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_036240.pdf

5. U.S. National Library of Medicine, National Institute of Health, Unified Medical Language System, [Internet]. (2010) [cited 2010 Feb. 8] Available from: http://www.nlm.nih.gov/research/umls/

6. Benson T, Principles of health interoperability: HL7 and SNOMED, London: Springer-Verlag; 2010 (pp 173 – 216) [cited 2010 June 1]

7. Aronson A, MetaMap: Mapping text to the UMLS Metathesaurus, U.S. National Library of Medicine, National Institute of Health, (2006), [Internet] [cited 2010, Mar. 3] Available from: http://skr.nlm.nih.gov/papers/references/metamap06.pdf

8. Fraser G, Capstone: An efficient approach to the development and validation of a SNOMED clinical terms subset for use as a problem list dictionary, Department of Medical Informatics & Clinical Epidemiology, O.H.S.U., (2004), [Internet]. [cited 2010 Jan 2] Available from: http://www.ohsu.edu/dmice/people/students/theses/2004/upload/Fraser-Capstone.pdf

9. Lau F, Simkus R, Lee D, A methodology for encoding problem lists with SNOMED CT in general practice, Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)

10. Ruch P, Gobeill J, Lovis C, Geissbuhler A, Automatic medical encoding with SNOMED categories, BMC Medical Informatics and Decision Making, (2008), 8(Suppl 1):56, doc:10.1186/1472-6947-8-SI-S6 Available from: http://www.biomedicalcentral.com/1472-6947/8/S1/S6

11. Elkin PL, et al., Evaluation of the content coverage of SNOMED CT: Ability of SNOMED clinical terms to represent clinical problems, Mayo Clin Proc. 2006 Jun;81(6):741-748 Available from: http://www.ncbi.nlm.nih.gov/pubmed/16770974

12. Foley M, Chapter 7 In: K. Giannangelo editor, Healthcare code sets, clinical terminologies, and classification systems Chicago: AHIMA; 2010 (pp105-126) [cited 2010 June 1]

13. Cimino JJ, Desiderata for controlled medical vocabularies in the twenty-first century. Meth Inform Med. 1998;37:394-403 [cited 2010 June 3] available from: http://www.sahs.uth.tmc.edu/evbernstam/HI5300/Articles%20(Reading%20Materials)/cimino_desiderata-for-controlled-medical.pdf

14. Cimino JJ, 1997-Imia-Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.ppt, [cited 2010 June 3] available from: http://people.dbmi.columbia.edu/cimino/Presentations/1997-Imia-Desiderata%20for%20Controlled%20Medical%20Vocabularies%20in%20the%20Twenty-First%20Century.ppt

15. IHTSDO, SNOMED Clinical Terms Overview, [cited 2010 June 3] available from: http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Recourses/Introducing_SNOMED_CT/SNOMED_CT_Overview_IHTSDO_Taping_Aug08.pdf

16. National Library of Medicine. Web Tools Tutorial [Internet]. (2004) [cited 2010 Oct 31]. Available from: http://umlslex.nlm.nih.gov:8100/WebLvg/html/luiNormTutorial.html

17. Aronson AR, Lang FM, An overview of MetaMap: historical perspective and recent advance. JAMIA, 17(3):229, 2010

Appendices

Appendix 1: Database Schema

The entity relationship diagram for the database is shown in Figure 8**.** The diagram represents the custom designed tables; the standard UMLS tables are not shown.

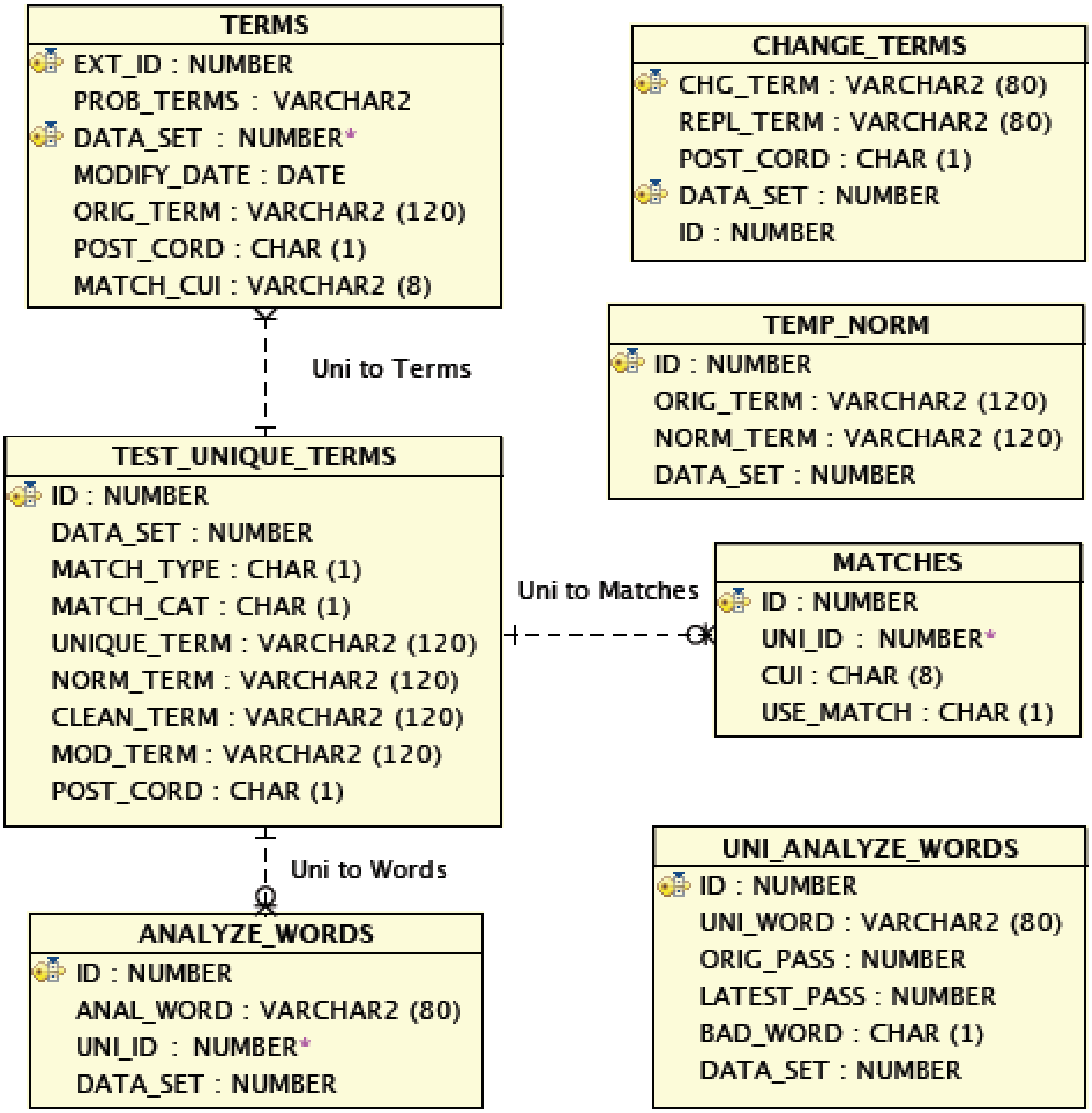


Figure 8. Entity relationship diagram for database

Appendix 2: Match Algorithms

SNOMED CT Terms: A SNOMED CT, English only subset was derived from the United Medical Language System (UMLS). The subset was loaded into Oracle for matching purposes. Three tables, MRCONSO, MRXNS_ENG and MRXW_ENG were used for various mapping purposes that are explained below. The table and columns utilized by the matching algorithms are summarized in Table 14.

Table Name	Table Description	Key Columns
MRCONSO	Table holding SNOMED CT strings, concept IDs and other attributes	STR - String associated with concept CUI - Concept ID TTY – Concept type ISPREF – Preferred term flag
MRXNS_ENG	Table holding normalized SNOMED CT strings	NSTR – Normalized string associated with concept CUI - Concept ID
MRXW_ENG	Table holding valid English words	WD – Word string

Table 14. Key UMLS tables used in matching algorithms

Exact Match: This method attempts to match a unique problem list term to the STR column of the MRCONSO table where all words are the same and are in the same sequence. Both the problem list term and STR column are lowercased for the match. No additional words in the STR column are allowed.

Exact Match – Normalized Strings: The unmatched problem list terms are extracted and normalized using the NLM luiNorm program. The luiNorm additionally selects a single phrase based on a weighting method called "canonization"**.** The normalize terms are then reloaded into the database and associated with their corresponding unique term. The normalized string is matched against the NSTR column of the MRXNS_ENG table. Similar to exact match, all words must be the same and in the same order.

Match-All: The match-all method uses attempts to match the STR column of MRCOSNO where all words in the problem list term are present but not necessarily in the same order. Additional words can be present in the STR column. This match uses the functionality of Oracle Text to accomplish the matching.

Match-All – Normalized Strings: Using the same normalization approach as above, a match-all method is used against the NSTR column of the MRXNS_ENG table. Additional words can exist in the NSTR column. Multiple matches can exist for a PL term; these must also be reviewed to select the most appropriate match.

Word Matching: Words parsed from the problem list terms are matched against the WD column of the MRXW_ENG table. All matching is done using lower case. Words not found are marked as "bad" for review.

Appendix 3: Additional Interface Use Cases

Title: Define Mapping Project

Description: Setup of project for mapping effort. Project information such as data source or owner may be recorded. Key process is the assignment of a "dataset identifier" to a project. This will allow multiple datasets to be active in the system simultaneously.

1. User requests a new project. A unique numeric identifier is provided by the system.
2. User enters required data including data source, project name and other data.
3. User saves or cancels request.

Title: Load Project Terms

Description: User loads a set of project terms into the system.

1. User requests a load of terms and provides a project id.
2. System checks for the existence of terms data with associated id. Loading is not allowed if existing data exists.
3. System prompts for file name or allows browsing to selected file.
4. Appropriate format is verified and if correct data is loaded.
5. Feedback is provided as to success of load and total number of records loaded.

Title: Derive Unique Terms

Description: User requests a derivation of unique terms from an existing set of unmatched terms.

1. User supplies a project id and requests a derivation run.
2. System checks for the existence of unmatched, unique terms for the project and deletes those found.
3. System checks for the existence of unmatched terms and if found derives the unique set and stores it in the database.
4. System provides feedback including the total number of unique terms derived and the total number of associated project terms.

Title: Find Exact Matches

Description: Using a project set of unique unmatched terms, the system identifies potential SN CUI's by matching the term to a SN string. The strings must match in words, punctuation and order. Capitalization differences are suppressed.

1. User supplies a project id and requests exact match processing.

2. System applies an exact match algorithm to all terms not already identified as matched.
3. All matches are stored with associations to the unique term.
4. Matched unique terms are marked with an identifier indicating a potential exact match exists.
5. Feedback is provided as to success of matching effort including the total number of unique terms associated with matches, the total number of matches and the total number of duplicate matches (unique terms with more than 1 potential exact match).