


A Methodology for Encoding Problem Lists with SNOMED CT in General Practice



Researchers

Francis Lau, Ray Simkus, Dennis Lee
University of Victoria, BC, Canada

Funding support by Canadian Institutes for Health Research Strategic Training Initiatives

Overview

- Purpose, Setting and Rationale
- Methods – Algorithms, Normalization, Process, Encoding, Navigation
- Results – PL Matching Summary, Encoded Terms, Navigating Hierarchy
- Discussion – Methodology, Next Steps

Purpose, Rationale and Setting

□ Purpose

- Apply reverse mapping methodology to encode problem list (PL) terms in SNOMED CT for small Canadian general practice settings

□ Rationale

- Small-medium size vendors, GP offices
- No SCT sites; demonstrations/guidelines needed

□ Setting

- 4 general practitioners, 8 yrs with EMR system
- PL terms from 1 GP extracted for study

Methods

- Term Matching Algorithms
- Normalization – UMLS/Porter, Expanded
- Term Matching Process
- Encoding PL Terms
- PL Hierarchy Navigation

Methods: Term Matching Algorithms

Algorithm	Explanation
Exact Match	Exact string match where all words are same and in same sequence for both source and target terms, including punctuation
Match All	String match where all words are same but not necessary in same order; additional words allowed in target term
Partial Match	String match where one or more words in source term is found in target term
Semantic Match	For inactive concepts found, use historical relationships of WAS-A, SAME-AS, MAY-BE-A, REPLACE-BY to find current concepts
Unmatchable	Assigned when no match is found

Methods: Normalization UMLS

Steps 1 to 6	Example
Remove genitive	Hodgkin <i>'s</i> diseases, NOS – Hodgkin diseases, NOS
Remove stop words	Hodgkin diseases, <i>NOS</i> – Hodgkin diseases,
Convert to lowercase	<i>H</i> odgkin diseases, – hodgkin diseases,
Strip punctuation	hodgkin diseases, – hodgkin diseases
Uninflect phrase	hodgkin disease <i>s</i> – hodgkin disease
Sort words	<i>hodgkin disease</i> –disease hodgkin

Methods: Normalization Expanded

Step 2	Explanation
Stop words	Frequent short words that do not affect the phrase: and, by, for, in, of, on, the, to, with, no, and (nos)
Exclude words	Words that may change meaning of the word but if ignored help to locate a term otherwise missed: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while, without
SNOMED Prefixes	[X] – concepts with ICD-10 codes not in ICD-9 [D] – concepts in ICD-9 XVI and ICD-10 SVII [M] – morphology of neoplasm concepts in ICD-O [SO] – concepts in OPCS-4 chapter Z in CTV3 [Q] – temporary qualifying terms from CTV3 [V] – concepts in ICD-9 and ICD-10 on factors influencing health status and contacts with health services (V and Z codes)

Methods: Term Matching Process

- ❑ Cycle thru algorithms to find best candidate SCT terms
 - Exact match, Match-all-only, Match-all, Partial, Semantic
- ❑ For each algorithm, cycle thru source terms
 - Original term, normalized UMLS terms, stemmed terms
- ❑ In each cycle, review candidate concepts
 - Determine if there is a match and what type of match
 - If inactive concepts found, do semantic match to find current concepts thru historical relationships
- ❑ If no matching concepts found, label as unmatchable
- ❑ Tally frequency statistics on matches, count only first-matches

Results

- Matching Summary
- Encoded PL Terms
- Navigating PL Hierarchy

Results – Matching Summary

- Total patients 2,894
- Total PL entries 7,833
- Total words 16,455
- Unique words 1,764
- Longest word “Hypercholesterolemia” 20 chars
- Median length 8 chars
- Most common Hypertension 585 times

Results: Matching Summary, Top 10

- Top 10 PL terms by frequency
 - Hypertension
 - Hypercholesterolemia
 - diabetes mellitus
 - Hypothyroid
 - Asthma
 - Atrial fibrillation
 - Gastroesophageal reflux
 - Depression
 - Congestive heart failure
 - Chronic kidney disease

Results – Summary, Match Types

Match Type	Initial Cycle Frequency (%)	2 nd Cycle Frequency (%)
Exact-match	905 (52.8%)	1,296 (88.2%)
Match-all	167 (9.8%)	52 (3.5%)
Partial-match	633 (37.0%)	120 (8.2%)
Semantic-match*	49 (2.9%)	20 (1.4%)
Unmatched	8 (0.4%)	2 (0.1%)
Post-coordination	Not done	In-progress
Total	1,713	1,468

* Not included in total count

Results – Summary, Top PL Matches

Code	Description	Exact			All (w exclude words)			Partial
		OT	UMLS	PS	OT	UMLS	PS	
173	Asthma	6	5	0	0	0	0	0
190	Atrial Fibrillation	3	2	0	0	0	0	40
367	Chronic Kidney Disease	0	0	0	0	0	0	40
432	Congestive Heart Failure	2	0	0	0	0	0	40
490	Depression	3	4	6	0	0	0	0
501	Diabetes Mellitus	5	1	0	0	0	0	40
673	Gastroesophageal reflux disease	2	0	0	0	0	0	40
788	Hypercholesterolemia	3	4	0	0	0	0	0
804	Hypertension	3	2	0	0	0	0	0
826	Hypothyroid	1	0	8	0	0	0	0

No	Match	Lex/Sem	DescriptionId	Term	DescriptionStatus	DescriptionType	ConceptId	FullySpecifiedName	ConceptStatus

(104) 173 Asthma									
1	Exact	Lexical	301485011	Asthma	Current (0)	Preferred (1)	195967001	Asthma (disorder)	Current (0)
2	Exact	Lexical	535919016	Asthma	Concept inactive	Fully Specified Na	155574008	Asthma	Duplicate (2)
3	Exact	Lexical	242335016	Asthma	Concept inactive	Preferred (1)	155574008	Asthma	Duplicate (2)
4	Exact	Lexical	2547707010	Asthma	Concept inactive	Preferred (1)	187687003	Asthma (disorder)	Duplicate (2)
5	Exact	Lexical	35840014	Asthma	Concept inactive	Preferred (1)	21341004	Asthma (disorder) [Ambiguous]	Ambiguous (4)
6	Exact	Semantic	350147016	Childhood asthma	Current (0)	Preferred (1)	233678006	Childhood asthma (disorder)	Current (0)
7	Exact	Semantic	2157699014	Asthmatic bronchitis	Current (0)	Preferred (1)	405944004	Asthmatic bronchitis (disorder)	Current (0)
8	Exact	Semantic	1483199016	Allergic asthma	Current (0)	Preferred (1)	389145006	Allergic asthma (disorder)	Current (0)
9	Exact	Semantic	2157501011	Allergic bronchitis	Current (0)	Preferred (1)	405720007	Allergic bronchitis (disorder)	Current (0)
10	Exact	Lexical	415456016	Asthma	Concept inactive	Synonym (2)	278517007	Asthmatic bronchitis (disorder)	Ambiguous (4)
11	Exact	Lexical	35841013	Asthma, NOS	Retired without a	Synonym (2)	21341004	Asthma (disorder) [Ambiguous]	Ambiguous (4)
12	Exact	Lexical	301523015	Asthma NOS	Concept inactive	Synonym (2)	195983001	(Asthma: [exercise induced] or [all	Ambiguous (4)
13	Exact	Semantic	52467014	Exercise-induced asthma	Current (0)	Preferred (1)	31387002	Exercise-induced asthma (disorder)	Current (0)
14	Exact	Lexical	396251010	Asthma NOS	Concept inactive	Synonym (2)	266398009	Asthma: [NOS] or [attack]	Ambiguous (4)
15	Exact	Semantic	396119015	Asthma attack	Current (0)	Preferred (1)	266364000	Asthma attack (disorder)	Current (0)
16	Exact	Lexical	242351012	Asthma NOS	Concept inactive	Synonym (2)	155579003	Asthma: [NOS] or [attack]	Ambiguous (4)
17	Exact	Lexical	396120014	Asthma NOS	Limited (6)	Preferred (1)	266365004	Asthma NOS (disorder)	Limited (6)

(117) 190 Atrial Fibrillation									
1	Exact	Lexical	82343012	Atrial fibrillation	Current (0)	Preferred (1)	49436004	Atrial fibrillation (disorder)	Current (0)
2	Exact	Lexical	241913016	Atrial fibrillation	Concept inactive	Synonym (2)	155364009	(Atrial fibrillation) or (atrial flutter	Ambiguous (4)
3	Exact	Semantic	9988012	Atrial flutter	Current (0)	Preferred (1)	5370000	Atrial flutter (disorder)	Current (0)
4	Exact	Lexical	395943011	Atrial fibrillation	Concept inactive	Synonym (2)	266306001	(Atrial fibrillation) or (atrial flutter	Ambiguous (4)
5	Exact	Lexical	241912014	Fibrillation - atrial	Concept inactive	Synonym (2)	155364009	(Atrial fibrillation) or (atrial flutter	Ambiguous (4)
6	Exact	Lexical	395940014	Fibrillation - atrial	Concept inactive	Synonym (2)	266306001	(Atrial fibrillation) or (atrial flutter	Ambiguous (4)

(276) 432 Congestive Heart Failure									
1	Exact	Lexical	70653017	Congestive heart failure	Current (0)	Preferred (1)	42343007	Congestive heart failure (disorder)	Current (0)
2	Exact	Lexical	300176012	Congestive heart failure	Concept inactive	Synonym (2)	195108009	Heart failure: [right] or [congestive	Ambiguous (4)
3	Exact	Semantic	490972013	Right ventricular failure	Current (0)	Preferred (1)	367363000	Right ventricular failure (disorder)	Current (0)

Results: Encoded PL Terms, Revision

- ❑ 2 cycles of encoding PL terms
- ❑ 3 missing SCT terms, i.e. vasculopath, pyocystitis, hypotestosteronemia
- ❑ 5 misspelled, idiosyncratic and ambiguous PL terms
- ❑ 313 PL parts became modifiers - left, right, lower, midline, chronic, recurring, active, query, multiple
- ❑ Modifiers – time course, number, location, severity
- ❑ Preference – PL term of “chronic kidney disease” over SCT term of “chronic renal failure”
- ❑ 89 post-coordination of 2 SCT concepts

Original PL Term	Type of Match	Identifier	Id Type	SNOMED Term	Descn Type	Descn Status
Atrial Fibrillation	Exact	49436004	C	Atrial fibrillation (disorder)	F	0
		82343012	D	Atrial fibrillation	P	0
Hypertension	Exact	38341003	C	Hypertensive disorder, systemic arterial (disorder)	F	0
		1215744012	D	Hypertensive disorder	P	0
		64176011	D	Hypertension	S	0
Gastroesophageal Reflux - GERD	All	235595009	C	Gastroesophageal reflux disease (disorder)	F	0
		2535970019	D	GERD – Gastro-esophageal reflux disease	S	0
Cirrhosis	Semantic	155809006	C	Cirrhosis	U	4
		19943007	C	Cirrhosis of liver (disorder)	F	0
		33568015	D	Cirrhosis of liver	P	0
Atrial Fibrillation - Chronic	Post, Exact	82343012	D	Atrial fibrillation	P	0
		288524001	C	Courses (qualifier value)	F	0
		428182017	D	Courses	P	0
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	Chronic	P	0
Chronic Kidney Disease - CKD	Post	90708001	C	Kidney disease (disorder)	F	0
		150315015	D	Kidney disease	P	0
		263502005	C	Clinical course (attribute)	F	0
		391753013	D	Clinical course	P	0
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	Chronic	P	0
Headache Migraine	Post, Exact	37796009	C	Migraine (disorder)	F	0
		63055014	D	Migraine	P	0
		246090004	C	Associated finding (attribute)	F	0
		367802015	D	Associated finding	P	0
		25064002	C	Headache (finding)	F	0
		41990019	D	Headache	P	0

Results: Encoded PL Terms, Links

Linking PL to SCT terms

Rec	PL-Id	PL-Term	ConceptId	DescriptionId	Match	AttributeId	SequenceId
1	160	Atrial Fibrillation	49436004	83243013	Exact	0	0
2	789	Hypertension	38341003	64176011	Exact	0	0
3	685	Gastroesophageal Reflux - GERD	235595009	2535970019	All	0	0
4	32666	Chronic Kidney Disease CKD	90708001	150315015	Post	0	0
5	32666	Chronic Kidney Disease CKD	90734009	150360019	Post	263502005	1
6	431	Cirrhosis	19943007	33568015	Semantic	0	0
7	1044	Headache Migraine	37796009	63055014	Post, Exact	0	0
8	1044	Headache Migraine	25064002	41990019	Post, Exact	246090004	1

Table 5. Examples of the index table linking the original PL terms to matched SNOMED terms.

Legend: SequenceId indicates the relative ordering of the post-coordinated records

Two sets of post-coordinated terms shown above

Results: Navigating PL Hierarchy

- Can a PL hierarchy improve retrieval?
- Creating the PL hierarchy
 - Select 32 PL terms for cardiovascular disorders
 - Tally frequency counts from EMR
 - Start with Disorder of cardiovascular system
 - Prune tree, include concepts with counts
 - Leave immediate supertype intact
- More refined/selective concept retrieval?

49601007 Disorder of cardiovascular system (disorder) - 1

128487001 Acute disease of cardiovascular system (disorder)

127337006 Acute heart disease (disorder)

57054005 Acute myocardial infarction (disorder)

70211005 Acute myocardial infarction of anterolateral wall (disorder) - 1

73795002 Acute myocardial infarction of inferior wall (disorder) - 5

307140009 Acute non-Q wave infarction (disorder) - 5

12026006 Paroxysmal tachycardia (disorder) - 1

9904008 Congenital anomaly of cardiovascular system (disorder)

363028003 Congenital anomaly of cardiovascular structure of trunk (disorder)

13213009 Congenital heart disease (disorder) - 1

10818008 Congenital malposition of heart (disorder)

27637000 Dextrocardia (disorder) - 1

27550009 Disorder of blood vessel (disorder)

359557001 Disorder of artery (disorder)

72092001 Arteriosclerotic vascular disease (disorder)

53741008 Coronary arteriosclerosis (disorder) - 9

414024009 Disorder of coronary artery (disorder)

53741008 Coronary arteriosclerosis (disorder) - 9

55855009 Disorder of pericardium (disorder)

3238004 Pericarditis (disorder) - 2

15555002 Acute pericarditis (disorder) - 1

56265001 Heart disease (disorder) - 1

127337006 Acute heart disease (disorder)

57054005 Acute myocardial infarction (disorder)

70211005 Acute myocardial infarction of anterolateral wall (disorder) - 1

73795002 Acute myocardial infarction of inferior wall (disorder) - 5

307140009 Acute non-Q wave infarction (disorder) - 5

12026006 Paroxysmal tachycardia (disorder) - 1

Discussion

- Encoding Methodology
- Next steps

Discussion: Methodology

- 1) Extract PL entries and tabulate frequency of terms present
- 2) Catalogue all unique words across PL terms
- 3) Examine unique PL terms to identify and revise acronyms, abbreviations, spelling variants and errors
- 4) Match PL terms to SCT using “reverse mapping” method
- 5) Create detailed/summary outputs to show matches
- 6) Review matched SCT for accuracy, remove successful matches from further cycles
- 7) Repeat steps 3-6 until no further matches found
- 8) Post-coordinate remaining PL terms with qualifier, refinement and combined concepts
- 9) Create pruned PL hierarchy tree with counts+supertypes
- 10) Create index table with unique identifiers for PL+SCT terms

Discussion: Next Steps

- Tools/Resources
 - What tools and resources most suited for small GP settings?
 - Bundled with HL7 v3 messages?
- EMR Systems
 - Histories, investigations, drugs, billing, referrals, etc.
 - How to work with EMR vendors?
- Hierarchy/Subset
 - Use of existing SCT hierarchy, primary care subset?
- Value
 - How to demonstrate value of SCT in small GP settings?
 - Interoperability, accuracy, patient safety?