Medical Terminologies: Why, What and How?

Kin Wah Fung MD, MS, MA

U.S. National Library of Medicine National Institutes of Health Department of Health and Human Services



Preparing for Interoperability – Terminology Standard for eHR, 2011

Outline

Why do we need medical terminologies

- Problems with free text
- Breaking the semantic barrier
- Coded vs. uncoded information
- What should we use
 - Available choices, and why so many
 - Terminologies vs. classifications, and what to use
- How to harness the power of terminologies
 - Desiderata for medical terminologies
 - Principles and best practices



Medical data representation

- 'The representation of medical data is the foundation on which the entire field of Medical Informatics is based, yet it is often given little attention and taken for granted' (*James Cimino*)
- Medical knowledge representation is more than just using words to represent medical ideas (we already have that for hundreds of years)

The formal manipulation of medical data by computer systems places demands on the representation of these data that have not been required of paper-based information – so we cannot simply adopt the same techniques for data capture in paper records



What are medical terminologies?

- Terminology: A finite, enumerated set of terms used to convey information unambiguously (a controlled vocabulary)
- Main use of medical terminologies
 - To encode (index/annotate) medical information e.g. Electronic Health Records (EHR), medical literature
- Why do we need them?
 - Computers cannot reliably 'understand' (respond to/act upon) natural language (free text)
 - Controlled vocabularies help to reduce complexity by restraining content (what can be said) and syntax (how it is said) analogous to reserved words in programming languages e.g. GOTO, DO...WHILE, FOR...NEXT
 - Electronically-stored information is much more useful if it is encoded



Problems with free text (natural language)

- Medical utterances are difficult for computers:
 - Same meaning, different words (redundancy) e.g. oral, per oral, by mouth, P.O, orally
 - Same word, different meanings (ambiguity) e.g. gunshot in the atrium, COLD, MI, DM
 - Meaning-modifying context e.g. family history of breast cancer
 - Probabilistic expressions and negation e.g. possible pneumonia, myocardial infarction excluded
 - Typos and lexical variants e.g. anemia vs. anaemia
- Natural Language Processing (NLP) a lot of progress, but far from perfect
- Encoded information is still a more reliable way to ensure that the computer 'thinks' and 'acts' correctly





The Electronic Health Record (EHR)

- The transformation of paper records to EHR is more than just a change in storage medium
- "...future patient records must be more than a way to store patient data – they must also support the clinical decision process and help improve the quality of patient care.... Achieving these objectives requires that [EHRs] be more than automated paper records."

The Computer-Based Patient Record: An Essential Technology for Health Care, IOM 1997



From paper records to EHR



From paper records to EHR



EHR works better with encoded data

Functions of the EHR	Free-text data	Encoded data
Electronic record keeping	Limited	\checkmark
Clinical decision support (alerts and reminders, access to on-line information sources)	X	\checkmark
Administrative functions e.g. billing, public health reporting	X	\checkmark
Quality assurance e.g. outcome measurements, clinical guidelines	Limited	\checkmark
Clinical research	Limited	\checkmark



Semantic interoperability

- Terminologies are indispensable for semantic interoperability - the 'Holy Grail' of data sharing
- 'The ability to import utterances from another computer without prior negotiation, and have your decision support, data queries and business rules continue to work reliably against these utterances'*
- ♦ 3 levels of interoperability:
 - Syntactic HL7 message structure, XML schema
 - Narrative HL7 Clinical Document Architecture, HTML
 - Semantic HL7 Reference Information Model, Terminologies, Medical Logic Modules...

* Dolin and Alschuler 'Approaching semantic interoperability in Health Level Seven JAMIA 2011 Jan



Outline

Why do we need medical terminologies

- Problems with free text
- Breaking the semantic barrier
- Coded vs. uncoded information
- What should we use
 - Available choices, and why so many
 - Terminologies vs. classifications, and what to use
- How to harness the power of terminologies
 - Desiderata for medical terminologies
 - Principles and best practices



Terminologies are great.....BUT

• There are so many of them:

- Diagnosis/signs and symptoms ICD9CM, ICD10, ICD10CM, ICD10AM, ICD-O, ICPC, ICF, SNOMED CT, Read Codes, MedDRA, CTCAE, WHOART, MEDCIN, DSM
- Procedures CPT, CDT, HCPCS, OCPS, SNOMED CT, ICD9CM, ICD10-PCS
- Nursing NANDA, NIC, NOC, OMS, HHC
- Diagnostic tests LOINC, UltraSTAR
- Drugs VANDF, NDC, RXNORM, NDDF
- Medical devices UMDNS, GMDN, SPN
- Genomics GO, HUGO, NCBI Taxonomy
- ...



Why are there so many?

- Whenever there is a new need for information encoding, there is a temptation to create a new terminology
 - The terminology landscape is obscure and fragmented difficult to find out what is out there and who uses what
 - No existing terminology covers everything 'I can't find the terms I need'
 - Existing terminologies are often big and complicated it is often easier to create your own (maintenance is another story)
 - The 'update penalty' the burden to keep up-to-date
 - Geographical/historical/political reasons
- The result multiple terminologies with varying degrees of overlap. Problems:
 - Wastage of resources 'reinventing the wheel'
 - Data interoperability becomes much more difficult





Role of NLM in health data standards

- The U.S. National Library of Medicine is the central coordinating body for clinical terminology standards within the Dept of Health and Human Services
- NLM supports the development, enhancement and distribution of clinically specific terminologies to facilitate the exchange of clinical data and improve retrieval of health information
 - clinical terminologies supported, licensed or developed by NLM: SNOMED CT, LOINC, RxNorm
 - uniform distribution mechanism for HIPAA code sets and clinical terminology standards through the UMLS
 - coordinates efforts to develop mappings between HIPAA code sets and standard clinical terminologies



Terminology of terminologies

Other artifacts with 'terminology-like properties'

- classification
- vocabulary
- dictionary
- code set
- thesaurus
- nosology
- taxonomy
- lexicon
- Ontology...



Medical classifications

- Unlike medical terminologies which are relatively recent creations, medical classifications have been in use for hundred of years
 - ICD codes can be traced back to the London Bills of Mortality, 1662 (about 60 categories)
 - Primary purpose is statistical
 - Main user and audience is human, not computers

 Since ICD classifications also have 'terminology-like' characteristics e.g. codes, terms and hierarchies, they are used like terminologies in computer systems



Terminology vs. classification

	Terminology	Classification
Example	SNOMED CT	ICD-9-CM
Usage	General - clinical documentation	Specific - mortality statistics, population health, reimbursement
Goal	Capture of meaning without loss	Abstraction of data to facilitate aggregation and comparison
Structural features:		
Hierarchy	Polyhierarchy	Strict hierarchy (no double counting)
Coding level	Any level	Only at leaf level (needs 'unspecified'/'NOS')
Overlap between concepts	Possible	Pairwise disjoint ('pigeon hole' approach – one code only)
Exhaustiveness	Desirable but not essential	Jointly exhaustive (a code for everything, needs 'Other'/'NEC')





VAI

Poly vs. Strict hierarchy

- Hepatorenal syndrome a life-threatening medical condition that consists of rapid deterioration in kidney function in individuals with cirrhosis or fulminant liver failure
- SNOMED CT

ICD-9-CM



NOS and NEC terms

◆ PNEUMONIA AND INFLUENZA (480-488)

- 480 Viral pneumonia
 - 480.0 Pneumonia due to adenovirus
 - 480.1 Pneumonia due to respiratory syncytial virus
 - 480.2 Pneumonia due to parainfluenza virus
 - 480.3 Pneumonia due to SARS-associated coronavirus
 - <u>480.8 Pneumonia due to other virus not elsewhere classified</u>
 - <u>480.9 Viral pneumonia, unspecified</u>

Not Otherwise Specified (NOS)

Not Elsewhere Classified (NEC)



Semantic drift of NEC terms

ICD-9-CM in 2003

480 Viral pneumonia
480.0 Pneumonia due to adenovirus
480.1 Pneumonia due to respiratory syncytial virus
480.2 Pneumonia due to parainfluenza virus
480.8 Pneumonia due to other virus not elsewhere classified
480.9 Viral pneumonia, unspecified

ICD-9-CM in 2010

- 480 Viral pneumonia
 - 480.0 Pneumonia due to adenovirus
 - 480.1 Pneumonia due to respiratory syncytial virus
 - 480.2 Pneumonia due to parainfluenza virus
 - 480.3 Pneumonia due to SARSassociated coronavirus
 - 480.8 Pneumonia due to other virus not elsewhere classified
 - 480.9 Viral pneumonia, unspecified





Semantic drift of NEC terms

ICD-9-CM in 2003

480 Viral pneumonia
480.0 Pneumonia due to adenovirus
480.1 Pneumonia due to respiratory syncytial virus
480.2 Pneumonia due to parainfluenza virus
480.8 Pneumonia due to other virus not elsewhere classified
480.9 Viral pneumonia, unspecified

ICD-9-CM in 2010

- 480 Viral pneumonia
 - 480.0 Pneumonia due to adenovirus
 - 480.1 Pneumonia due to respiratory syncytial virus
 - 480.2 Pneumonia due to parainfluenza virus
 - 480.3 Pneumonia due to SARSassociated coronavirus
 - 480.8 Pneumonia due to other
 virus not elsewhere classified
 480.9 Viral pneumonia,
 - unspecified





Use of ICD in the EHR

 ICD or ICD-based vocabularies are still quite commonly used in EHRs

- Directly generates codes for statistical reporting and administrative needs (e.g. reimbursement)
- Free to use (some restrictions for ICD10CM)
- Availability of trained coding professionals



For the EHR

 SNOMED CT is inherently more suitable than ICD for clinical documentation in the EHR

- Content coverage
- Clinical orientation
- Flexible data entry and retrieval
- Extensibility



Content coverage

- SNOMED CT has been repeatedly shown to provide better coverage (i.e. more expressive) than ICD
- Number of concepts:
 - SNOMED CT (Clinical finding): 100,000
 - ICD-9-CM: 14,000
 - ICD-10-CM: 68,000
- ICD's focus is statistical rare diseases get lumped together in grouper categories e.g. NEC codes (Not elsewhere classified) which are like 'information black holes'
- SNOMED CT is clinically-based document what is important for patient care



	ICD-9-CM	ICD-10-CM	SNOMED CT
Asperger's disorder	299.8 Other specified pervasive developmental disorders	F84.5 Asperger's disorder	23560001 Asperger's disorder
Apert syndrome	755.55 Acrocephalosyndact yly	Q87.0 Congenital malformation syndromes predominantly affecting facial appearance	205258009 Apert syndrome
Metabolic acidosis	276.2 Acidosis	E87.2 Acidosis	59455009 Metabolic acidosis
Respiratory acidosis	276.2 Acidosis	E87.2 Acidosis	12326000 Respiratory acidosis
Lactic acidosis	276.2 Acidosis	E87.2 Acidosis	91273001 Lactic acidosis





Clinical orientation

- SNOMED CT is designed for direct use by healthcare providers during the process of care
- ICD is for use by coding professionals after the episode of care
- ICD's limitations in clinical documentation
 - Some terms are not 'clinical user-friendly'
 - Some contents have little clinical relevance
 - Presumed knowledge of coding rules and conventions can lead to confusion or miscoding



'Unfriendly' ICD terms

Awkward names:

• E878.2 Surgical operation with anastomosis, bypass, or graft, with natural or artificial tissues used as implant causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation (ICD-9-CM)

Clinically irrelevant details:

 V30.2xxD Person on outside of three-wheeled motor vehicle injured in collision with pedestrian or animal in nontraffic accident, subsequent encounter (ICD-10-CM)



ICD coding rules and conventions

- Patient admitted with gastrointestional bleeding and found to be anemic. Should it be coded as
 - 280 Iron deficiency anemias secondary to blood loss (chronic), or
 - 285.1 Acute posthemorrhagic anemia
 - There is no general code to cover both cases of acute and chronic blood loss
- In ICD convention, words in parenthesis e.g. (chronic) are known as 'non-essential modifiers', so 280 is the correct code if clinical course is uncertain. But without this knowledge, coding can be inconsistent
- SNOMED CT: 413532003 Anemia due to blood loss



Flexible data entry and retrieval

- ICD dictates level of granularity of coding
 - NOS (Not otherwise specified) codes cases with insufficient information for more specific codes
 - NEC (Not elsewhere classified) cases with more specific information but not covered by existing codes
- SNOMED CT allows coding at any level of specificity as appropriate for the clinical situation (no need for NOS, NEC codes)
- Flexible data retrieval
 - SNOMED CT has multiple hierarchy (single hierarchy for ICD)
 - SNOMED CT concepts are defined logically by their attributes (only textual rules and definitions in ICD)



Clinical guideline for hypertension

◆ ICD-9-CM

- All codes in the range HYPERTENSIVE DISEASE (401-405)
- Will be missing
 - 410.9 Myocardial infarction with hypertension
 - 642 Hypertension complicating pregnancy, childbirth, and the puerperium

SNOMED CT

• All descendants of 38341003 Hypertensive disorder



Data retrieval using attributes

 Find all diseases caused by occlusion of artery affecting any artery except mesenteric or renal arteries

SNOMED CT

- Find all descendants of 2929001 Occlusion of artery
- Exclude those with Finding site = 'Structure of mesenteric artery' or 'Structure of renal artery'
- ◆ ICD-9-CM
 - 440 Atherosclerosis and descendants except 440.1 Of renal artery
 - 433 Occlusion and stenosis of precerebral arteries and descendants
 - 437.0 Cerebral atherosclerosis
 - 414.0 Coronary atherosclerosis
 - 416.0 Idiopathic pulmonary arteriosclerosis
 - 443.9 Peripheral vascular disease, unspecified...
- Update burden



Extensibility

- No single terminology will ever be complete, there is always a need for extension
- ◆ ICD no standard way for extension
- SNOMED CT well-defined rules to extend coverage by combining existing concepts (post-coordination) e.g.
 - Left kidney stone can be represented by adding the concept '7771000 Left' as a laterality attribute to '95570007 Kidney stone'

Advantages:

- Can compute equivalence of new concepts to existing concepts
- The new concept will be recognized as a sub-class of existing concepts



Outline

Why do we need medical terminologies

- Problems with free text
- Breaking the semantic barrier
- Coded vs. uncoded information
- What should we use
 - Available choices, and why so many
 - Terminologies vs. classifications, and what to use
- How to harness the power of terminologies
 - Desiderata for medical terminologies
 - Principles and best practices



1. Content

- Must seek to provide breadth and depth
- Structure must not limit size
- Formal editorial policy and methodology
- 2. Concept orientation
 - Concepts, not terms
 - Non-vagueness: at least one meaning per concept
 - Non-ambiguity: at most one meaning per concept
 - Non-redundancy: one concept per meaning
 - Explicit synonyms

* Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century Methods of Information in Medicine 1998



- 3. Concept permanence
 - Old concepts should be explicitly retired (not deleted) e.g. Non-A, non-B hepatitis
 - Don't re-use concept identifiers
 - Names can be changed as long as meaning doesn't change (retronyms) e.g.
 - pacemaker can be renamed implantable pacemaker, and percutaneous pacemaker added as a new concept
 - It is however not correct to rename Non-A, non-B hepatitis as Hepatitis C



4. Non-semantic identifier

- Don't use the name
- Don't use a mnemonic
- Don't use a code that will run out of room
- Don't use a hierarchical code
- Meaningless integer (+/- check digit)



- 5. Polyhierarchy
 - Needed for understanding meaning
 - Needed for tree walking
 - Needed for inferencing
- 6. Formal definitions
 - Support understanding and maintenance
 - Structured and controlled (not narrative)
 - Represented through relationships within the vocabulary
 - Definitional versus assertional knowledge



Formal Definitions

Serum Potassium Test: "a test that measures potassium ion concentration in the serum"

Serum Potassium Test:

is_a: Test has_specimen: Serum Specimen measures_substance: Potassium Ion



Formal Definitions









- 7. Reject 'Not Elsewhere Classified'
 - Can never have a formal definition
 - Terminology changes induce semantic drift
 - There are valid alternatives e.g. store general term and additional qualifier
 - Old way:
 - "Pneumonia" coded as "Pneumonia, NOS"
 - "Informatician's Pneumonia" coded as "Pneumonia, NEC"
 - New way:
 - "Pneumonia" coded as "Pneumonia"
 - "Informatician's Pneumonia coded as "Pneumonia" +
 "Informatician



- 8. Multiple granularities
 - Different levels for different purposes
 - Uncertainty in medical language is inevitable, but strive to represent that uncertainty with precision
- 9. Multiple consistent views
 - No inconsistency between views



























10. Representing context

- Needed: a grammar to show usage
- "What is sensible to say"



Representing Context

Controlled Medical Terminology



LIBRARY OF MEDICINE

Representing Context

Controlled Medical Terminology



LIBRARY OF MEDICINE

11. Graceful evolution

- Reasons for change: user needs, mistakes, new knowledge
- Bad changes e.g. major name changes, code re-use, changed codes
- Good changes e.g. simple addition, refinement, minor name changes, disambiguation, obsolescence



12. Recognize redundancy

- Synonyms are good
- Redundant concepts are bad
- Redundant expressions are inevitable



Recognize Redundancy





Left Lower Lobe Pneumonia is-a: Pneumonia has-site: Left Lower Lobe participates-in: Finding



- Having a high-quality, up-to-date system of terminologies is very important
 - Major impact on the quality of the data
 - Determines what the data can be used for



Terminologies should be deployed pro-actively

- Terminology services as part of the infra-structure of the information systems
- Not as passive reference information or post-hoc coding efforts



New York Presbyterian Hospital Clinical Information Systems Architecture



Think ahead (and often) about maintenance

- Terminologies are not an one-time effort. In fact, they are never finished! Maintenance is the hard part.
- Establish editorial policy different levels of granularity, subspecialty requirements, how to hide/show content where necessary
- Update cycle depends on practical needs, reference standard update schedule etc.
- Manage user requests built-in new term submission mechanism, glean free-text entries for missing concepts, turn-around time must be reasonable



- Manage terminologies intelligently
 - Manual review alone is time-consuming and errorprone
 - Knowledge-based terminology management will improve quality, facilitate maintenance and enhance utility formal definitions, attributes of concepts, auto-classification and subsumption, computation of equivalence
 - Build appropriate tools for browsing and editing move beyond Excel spreadsheets!



Think internationally

- International standards are adopted for good reasons
 - Clinical data SNOMED CT
 - Laboratory tests and results LOINC
- Can benefit from others' efforts
 - Tooling IHTSDO Workbench of SNOMED CT
 - Useful subsets CORE Problem List Subset, Primary Care Subset of SNOMED CT
 - Maps from SNOMED CT to ICD9CM, ICD10, ICD10CM, ICPC2
- In future, there may be a need to share data across national borders



Reap the benefits

- Showcase the power of encoded data e.g. clinical decision support, intelligent data query and reporting, support clinical research
- Provide incentive to healthcare providers to code adequately, consistently and accurately
- One important way of quality assurance of a terminology is to examine the encoded information and its fitness for the intended purpose





- Terminologies are an important building block of the EHR, many advance functions (e.g. clinical decision support) need them to work
- SNOMED CT has a lot of desirable features as a clinical terminology
- Requirements of good terminologies and principles of management are quite wellestablished
- The investment in terminologies will pay off



